Asynchronous Design Methods for Dark Silicon Chips

Milos Krstic

IHP, Frankfurt (Oder), Germany

krstic@ihp-microelectronics.com

Dedicated to the dear friend and colleague Alex Yakovlev at the occasion of his 60th birthday.

1 Introduction

The scaling of CMOS technologies continues to enable each two years more and more transistors to the ASIC designers and eventually application users. In theory this would mean that the processing architectures could also exponentially develop in the direction of further many-processor systems scaling. Nevertheless there are significant issues which may disable such scenario with the current (20 nm, 14 nm) and future (10 nm, 8 nm and beyond) technology nodes. The power consumption which can be provided to the silicon unfortunately cannot scale further. This leads to the phenomenon named as dark silicon [1]. As a consequence, even if the increased number of available transistors is used to implement additional processor cores, all available cores cannot be powered at the same time, in order not to overload the thermal budget of the chip. Dark silicon is potentially very significant issue, practically disabling further simple scaling of homogenous processor architectures.

There are several consequences of dark silicon paradigm:

- The business as usual" strategy of scaled complex chips is not possible anymore and novel architectures and strategies are required. The future designs shell instead of parallel operation of many homogenous programmable processing cores, rather use large number of dedicated co-processors which will execute the dedicated tasks in a power-optimal manned only when needed and during idle time stay unpowered [2].
- The advanced power reduction methods are becoming completely unavoidable. The use of power gating, dynamic frequency and voltage scaling and even adaptive voltage scaling are the cornerstones of successful dark silicon ASIC implementation.
- The applied methods/architectures need to be utilized in a dynamic adaptive way, ensuring that the performance is available when applications needs it, but that it can be dramatically reduced in case of inactive operation or it can be utilized with additional robustness again when needed.

– The intelligent power management is needed to utilize the rich processing architectures and advanced power control mechanisms in an optimal and reliable way.

Use of standard synchronous methodologies is quite challenging with dark silicon limitation. Continuous clocking of the circuits adds increase to the power budget and creates the need for extensive clock gating which comes with its own overhead. Moreover the advanced power reduction methods, such as adaptive voltage scaling and use of the circuits in Near-Threshold (NT) modes, becomes to be very critical or suboptimal in synchronous systems due to the large on-chip variability. One important alternative is the partial or intensive use of the asynchronous logic in dark silicon systems.

2 Power Optimal Asynchronous Circuits

Asynchronous circuits are for many years proposed as low-power alternative to the standard synchronous approach. The main difference between the synchronous and asynchronous paradigm is that in the synchronous case there is only one global control signal - clock, which is periodically active regardless whether there is a need for processing in particular pipeline stage or not. In case of asynchronous design, the local activity is there only if there is a valid control initiator (token) which activates the local pipelines. With this methodology the dynamic power consumption could be significantly reduced compared to the synchronous counterparts.

The limitations of asynchronous methods, namely complicated design and purely developed test flow, disabled the pervasive use of the methodology for the applications beyond the academic demonstrators and few industry examples.

In the recent years the novel asynchronous design methods have been proposed, including the concept of desynchronization [3]. Desynchronization enables the seamless conversion of arbitrary synchronous circuit into the bundled data asynchronous design. Using such approaches it has been shown that the significant advantages of the asynchronous circuits in ultra-low power domains can be obtained. In particular, according to [4], in near threshold regimes, due to lack of the global timing and avoidance of worst case paradigm, 40% improved power consumption can be achieved. This can be traded also for more performance under the same power budget. Moreover, the novel aggressive fault tolerant voltage scaling approaches on the asynchronous side, such as recently proposed BLADE [5] show important improvements in comparison with state of the art synchronous power saving architectures such as Bubble-RAZOR.

One of the most interesting concepts for power reduction using asynchronous logic comes from the group of Prof. Alex Yakovlev at the University of Newcastle. He introduces the term "energy modulated computing" [6, 7], indicating the ability of asynchronous logic to use the quant of the energy which is currently available, i.e. to self-adjust the performance to the energy level which is currently available. This concept has been utilized in the various valuable architectures focused on ultra-low power usage in for example wireless sensor nodes [8].

Based on the prior work, it has been shown that the asynchronous circuit design methods are very effective for low-power applications in the scaled CMOS world especially for the applications with extreme power reduction requirements, such as Internet of Things (IoT). Nevertheless, the power requirements are not important only to mobile and IoT applications. The high-performance computing is also affected and dark silicon issues and causes the need to the radical paradigm change. In this context there is a significant chance for the asynchronous logic design methods to address dark silicon problems.

3 Asynchronous Design for Dark Silicon

In order to address the challenges imposed by the dark silicon issues the asynchronous logic design could be effectively utilized in several ways.

The future architectural concept of dedicated, power-optimal and power-controlled co-processor based design is much more suitable for the use of asynchronous logic then the generic homogenous many-core processing which was a main trend some years ago. The asynchronous co-processors could be effectively power-controlled, they do not need the distribution of the clock source, and could be fully event driven. Also utilization of the modern low-power techniques such as adaptive voltage scaling and power gating can be seamlessly integrated in the design of such co-processors. In general the concept of co-processor is based on irregular activation of the dedicated accelerated hardware processing which again in the event driven manner provides back the processed information. In its nature this function is elastic and does not need exact global synchronization and cycle based processing. Consequently, the use of the asynchronous logic seems to be the natural choice. As an example, the recent study [9] has shown that asynchronous co-processor for Elliptic Curve Cryptography (ECC) can reduce the power consumption in comparison to the synchronous counterpart by 1/3.

Moreover, the use of asynchronous logic is in general beneficial for the operations in the ultra-low power regimes of the operations such as Near Threshold Voltage Computing [4]. This can be additionally utilized in co-processor design in dark silicon chips to increase the performance and/or reduce the power consumption. Since synchronous design style still has significant merits when it comes to the high performance, due to the maturity of the design tools and simpler control protocol, it is plausible to propose also the use of mixed-mode synchronous-asynchronous logic, as illustrated in Fig. 1. In the context of dark silicon the specific co-processors can be designed in such way that their pipelines can be with control signal (Mode in Fig. 1) turn into asynchronous (bundled-data) pipelines. This asynchronous mode of operation could be used in near threshold voltage mode to further reduce the power consumption of the system. In high-performance mode, the regular synchronous pipeline could be activated. The overhead of this technique could be limited to the additional control gates which can be in scaled technologies, with billions of gates available, tolerated in many applications.

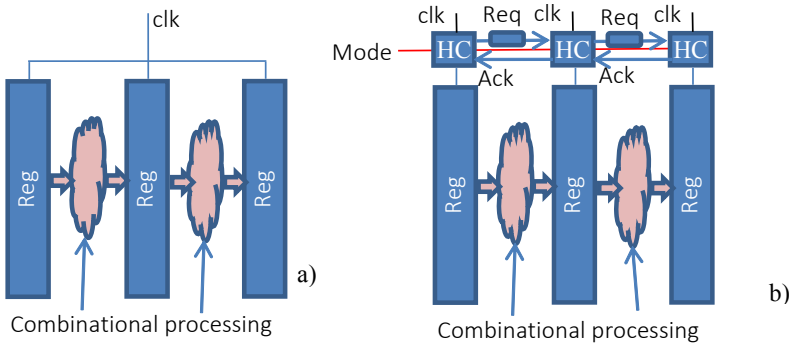

Fig. 1. Standard synchronous pipeline (a), and mixed-mode synchronous-asynchronous pipeline (b)

Finally, the inevitable part of the future dark silicon systems will be interconnects. The concepts of Networks of Chips have be proposed already many years ago, but mainly driven from the academic research. In this context it has been also shown that asynchronous logic can be utilized in much power efficient way compared to the synchronous approach. The power consumption of the asynchronous switches could be reduced by using asynchronous methods from 45% up to 91%, respectively in active and idle operation phases, compared with the clock gated synchronous design [10]. The industry standards (such as AMBA AXI, OCP) and applications push the interconnect development more into point-to-point link direction. However, in this case the use of the asynchronous interconnects is even more natural and simpler than in the case of bus or crossbar interconnect topologies. It is therefore reasonable to expect that in dark silicon chips the role of asynchronous interconnects will be significant. The power management of asynchronous blocks is simpler and more efficient then in case of the synchronous design. In general the free voltage adaptation is always possible and the processing performance will be self-adjusted to the current PVT (process, voltage, temperature) setting, without the need for PLL or similar blocks. Moreover, the blocks may have extremely fast event-based activation which doesn't require active clock being enabled all the time. Finally, there are no additional obstacles in integrating power gating with asynchronous logic.

4 Conclusions

Dark silicon issues impose significant challenges to the design and architectures of the future complex system-on-chip. In this paper it has been emphasized that the asynchronous design methods could be utilized as effective design methods to address the challenges of the dark silicon. Those design methods could be utilized to in general improve the power budget of the system both at the component and at the interconnect level. Moreover, such methods will enable more aggressive voltage scaling techniques and utilization of near-threshold voltages. Finally, the

event based computing is very suitable to the co-processor based architectures which expect to be the main stream solution for the future dark silicon systems.

References

1. H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark Silicon and the End of Multicore Scaling," *SIGARCH Comput. Archit. News*, vol. 39, no. 3, pp. 365–376, Jun. 2011. [Online]. Available: http://doi.acm.org/10.1145/2024723.2000108
2. Q. Zheng, N. Goulding-Hotta, S. Ricketts, S. Swanson, M. B. Taylor, and J. Sampson, "Exploring Energy Scalability in Coprocessor-Dominated Architectures for Dark Silicon," *ACM Trans. Embed. Comput. Syst.*, vol. 13, no. 4s, pp. 130:1–130:24, Apr. 2014. [Online]. Available: http://doi.acm.org/10.1145/2584657
3. J. Cortadella, A. Kondratyev, L. Lavagno, and C. P. Sotiriou, "Desynchronization: Synthesis of Asynchronous Circuits From Synchronous Specifications," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 10, pp. 1904–1921, Oct 2006.
4. J. Cortadella, L. Lavagno, D. Amiri, J. Casanova, C. Macin, F. Martorell, J. A. Moya, L. Necchi, D. Sokolov, and E. Tuncer, "Narrowing the margins with elastic clocks," in *2010 IEEE International Conference on Integrated Circuit Design and Technology*, June 2010, pp. 146–150.
5. D. Hand, M. T. Moreira, H. H. Huang, D. Chen, F. Butzke, Z. Li, M. Gibiluka, M. Breuer, N. L. V. Calazans, and P. A. Beerel, "Blade – A Timing Violation Resilient Asynchronous Template," in *Asynchronous Circuits and Systems (ASYNC), 2015 21st IEEE International Symposium on*, May 2015, pp. 21–28.
6. A. Yakovlev, "Energy-modulated computing," in *2011 Design, Automation Test in Europe*, March 2011, pp. 1–6.
7. F. Xia, A. Mokhov, Y. Zhou, Y. Chen, I. Mitrani, D. Shang, D. Sokolov, and A. Yakovlev, "Towards power-elastic systems through concurrency management," *IET Computers Digital Techniques*, vol. 6, no. 1, pp. 33–42, January 2012.
8. X. Zhang, D. Shang, F. Xia, and A. Yakovlev, "A Novel Power Delivery Method for Asynchronous Loads in Energy Harvesting Systems," *J. Emerg. Technol. Comput. Syst.*, vol. 7, no. 4, pp. 16:1–16:22, Dec. 2011. [Online]. Available: http://doi.acm.org/10.1145/2043643.2043646
9. S. Zeidler, M. Goderbauer, and M. Krstić, "Design of a Low-Power Asynchronous Elliptic Curve Cryptography Coprocessor," in *Proc. of the IEEE International Conference on Electronics, Circuits, and Systems (ICECS'13), Abu Dhabi, UAE*, Dec. 2013, pp. 569–572.
10. A. Ghiribaldi, D. Bertozzi, and S. M. Nowick, "A transition-signaling bundled data NoC switch architecture for cost-effective GALS multicore systems," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2013*, March 2013, pp. 332–337.