This is the post peer-review accepted manuscript of:

D. Rossi et al., "A −1.8V to 0.9V body bias, 60 GOPS/W 4-core cluster in low-power 28nm UTBB FD-SOI technology," 2015 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), Rohnert Park, CA, 2015, pp. 1-3.

The published version is available online at:

https://ieeexplore.ieee.org/document/7333483

© 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

A -1.8V to 0.9V Body Bias, 60 GOPS/W 4-core Cluster in low-power 28nm UTBB FD-SOI technology

Davide Rossi[1], Antonio Pullini[2], Michael Gautschi[2], Igor Loi[1], Frank Kagan Gurkaynak[2], Philippe Flatresse[3], Luca Benini[1,2]

[1]University Of Bologna, Bologna, Italy; [2]ETHZ, Zurich, Swizerland; [3]STMicroelectronics, Crolles, France
davide.rossi@unibo.it, Tel. +39 051 20 93843, Fax. +39 051 20 93839

Abstract

A 4-core cluster fabricated in low power 28nm UTBB FD-SOI conventional well technology is presented. The SoC architecture enables the processors to operate "on-demand" on a 0.44V (1.8MHz) to 1.2V (475MHz) supply voltage wide range and -1.2V to 0.9V body bias wide range achieving the peak energy efficiency of 60 GOPS/W, (419µW, 6.4MHz) at 0.5V with 0.5V forward body bias. The proposed SoC energy efficiency is 1.4x to 3.7x greater than other low-power processors with comparable performance.

Introduction

Ultra-low power operation and extreme energy efficiency are strong requirements for a number of high-growth Internet of-Things applications requiring near-sensor processing. A promising approach to achieve major energy efficiency improvements is near-threshold computing. However, frequency degradation due to aggressive voltage scaling may not be acceptable for performance-constrained applications. The SoC presented in this work exploits multi-core parallelism with explicitly-managed shared L1 memory to overcome performance degradation at low voltage, while keeping the flexibility typical of instruction processors. Moreover, enabling the cores to operate on-demand over a wide supply voltage and body bias ranges allows to achieve high energy efficiency over a wide spectrum of computational demands.

Conventional well UTBB FD-SOI technology

Past work on UTBB FD-SOI processors focused on the high-performance flavor of the technology (flip well) where aggressive forward body-biasing led to major operating frequency boost [1]. In this work, we focus for the first time on low-power multi-processor design based on low-leakage FD-SOI transistors where NMOS and PMOS are on P-well and N-well, a theoretical reverse body-biasing (RBB) up to -3V and a forward body-biasing (FBB) up to [VDD/2 + 300 mV] can be applied (Fig. 1). As opposed to the flip well flavor of the technology, supporting both RBB and FBB conventional well enables flexible management of leakage power and it is very well suited for low-power applications [1]. Moreover, when applied to conventional well, FBB is useful to achieve maximum energy efficiency and not only as a speed-boosting method (as opposed to the flip-well flavor).

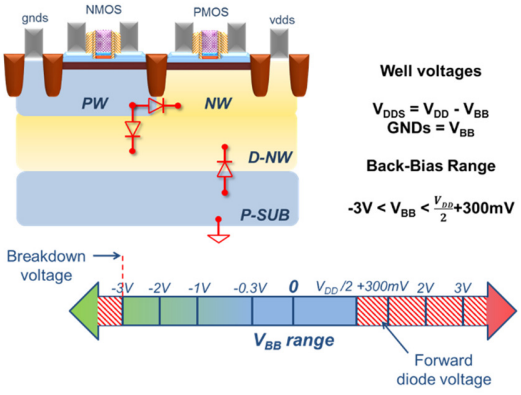

Fig. 1 UTBB-FDSOI conventional well transistors architecture.

SoC Architecture

The SoC consists of a cluster of four cores and 16 kB of L2 memory (Fig. 2). The cores, featuring 1K instruction cache each, are based on a highly power optimized microarchitecture implementing the OpenRISC ISA. GCC and LLVM toolchains are available for the core, OpenMP 3.0 is supported on top of bare-metal parallel runtime. Energy efficiency is boosted by using a carefully tuned pipeline depth to reduce register and clocking overhead, while the data-path is area-optimized to reduce leakage. To avoid memory coherency overhead and increase energy efficiency the cores do not have private data caches, while they share a L1 multi-banked Tightly Coupled Data Memory (TCDM) acting as an explicitly managed data scratchpad memory. The TCDM features 8 word-level interleaved 2kB SRAM banks connected to the processors through a non-blocking interconnect to minimize banking conflict probability. The whole memory space (L1, L2, memory mapped peripherals) is visible to all the cores of the cluster (Global Address Space architecture). L2 cluster memory latency is managed by a DMA featuring 10 cycles programming latency, up to 16 outstanding transactions and 4 physical channels for DMA control (one per core, thereby completely eliminating DMA control port contention). The DMA has a direct connection to the TCDM through 4 dedicated ports on the TCDM interconnect. This eliminates the need for data buffering in the DMA engine, which is very expensive in terms area and power. Starvation of the cores is prevented thanks to the x2 banking factor and fair arbitration.

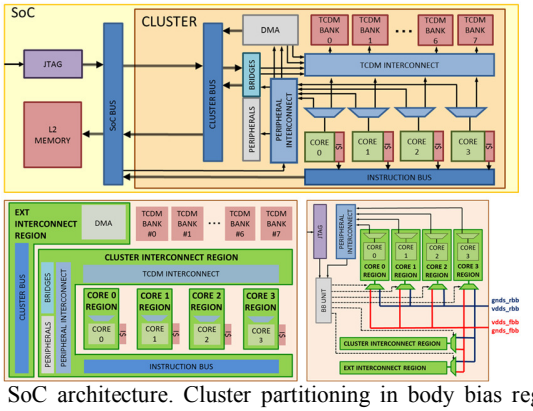

Fig. 2 SoC architecture. Cluster partitioning in body bias regions. Body bias control architecture.

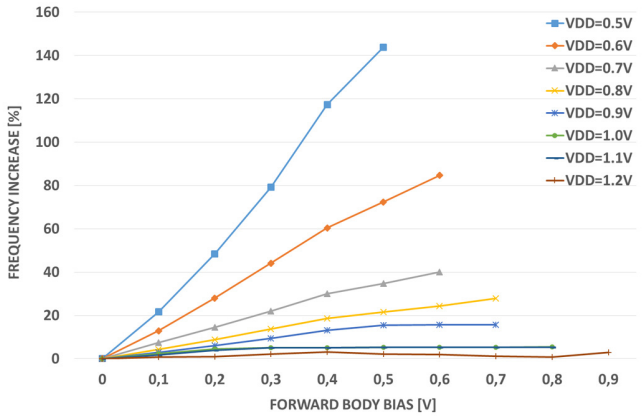

Fig. 3 Impact of forward body bias on maximum operating frequency.

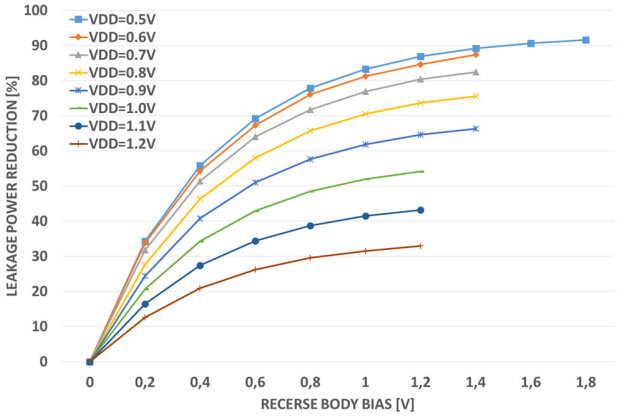

Fig. 4 Impact of reverse body bias on leakage power reduction.

Voltage [V]	Max Frequency [MHz]	Total Power @ Max Frequancy [mW]	Leakage Power @ 25°C [µW]
0.5	3.1	0.23	25.9
0.6	25.5	2.00	31.4
0.7	86.2	8.36	56.5
0.8	168.3	20.41	114.7
0.9	254.9	38.23	241.5
1.0	337.6	61.73	496.4
1.1	360	87.39	760.8
1.2	452	119.72	1695.3

Fig. 5 Maximum operating frequency, total power consumption, and leakage power consumption of the chip.

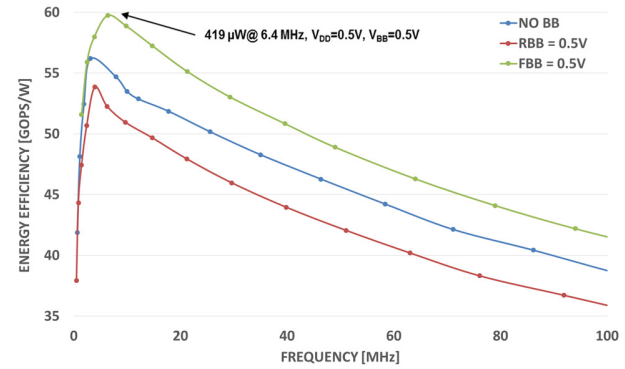

Fig. 6 Energy efficiency in the frequency range [0MHz; 100MHz].

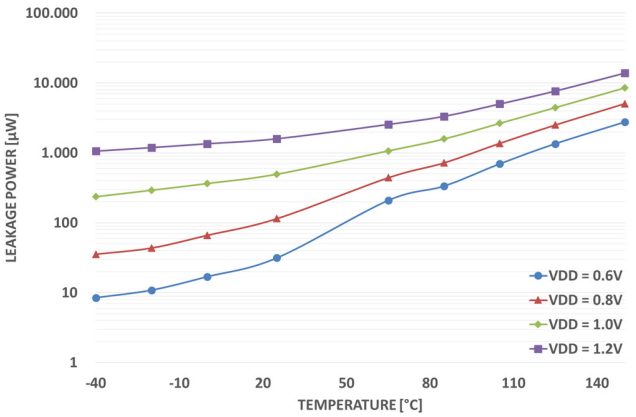

Fig. 7 Leakage current as function of the supply voltage in the temperature range [-40°C; 150°C].

Fine-Grained Body Biasing

To enable the SoC to achieve high energy efficiency for a wide range of workloads and to reduce the overhead of unused cores during the execution of sequential code, the cluster features on-demand shut down of cores by means of fine grained partitioning into 6 regions with separate clock trees and isolated wells (Fig. 2). The memory cuts that implement L2 memory, TCDM and I$ are based on standard SRAMs without support for body bias. 6 digitally controlled body bias multiplexers (BBMUXes) select the polarization of the P-well and the N-well of each region (vdds, gnds) choosing between two couples of global voltages (vdds_rbb, gnds_rbb; vdds_fbb, gnds_fbb) thus enabling per-region body-bias state selection (FBB or RBB). In contrast to DVFS and power gating approaches, this architecture has minimal overhead in term of area (less than 1%) and does not require level shifters and power grid isolation. The measured switching time between the two different body-bias states is lower than 30ns. Fig. 3 and Fig. 4 shows the measured impact of FBB and RBB on the operating frequency and leakage power of the body biased regions. In the near-threshold operating range, where the body bias technique is more effective, RBB provides up to 10x reduction of leakage power, while FBB provides an increase of up to 2.5x of operating frequency. This makes body biasing an excellent knob to modulate the leakage performance trade-off of the cluster regions enabling ultra-fast transitions between the high energy efficient active and the low leakage states.

Experimental Results

Fig. 5 shows the frequency, total power and leakage power resulting from the silicon measurements of the prototype running a typical parallel workload (matrix multiply). The chip is operational from 1.9MHz at 440 mV, 0.5 FBB to 475 MHz at 1.20V, 0,9 FBB. Fig. 6 shows the energy efficiency of the chip (GOPS/W). The peak energy efficiency of 60 GOPS/W is achieved by the chip at VDD=0.5V, and 0.5V FBB. At this voltage, the measured power is 419 µW at 6.4 MHz. It is interesting to note that moving from a 0.5V RBB to 0.5V FBB condition the energy efficiency increases by 13%, from 53 GOPS/W to 60 GOPS/W, while the best energy point increases by 60%, from 4 MHz to 6.4MHz. The leakage power of the SoC at 25°C ranges between 21µW at 0.44V and 1.7mW at 1.2V. As shown in Fig. 7, at the highest operating voltage of 1.2V and the temperature of 150°C, the leakage power of the chip is bounded to 11 mW. Even in this extreme operating point the leakage power never exceeds 22% of the overall power consumption.

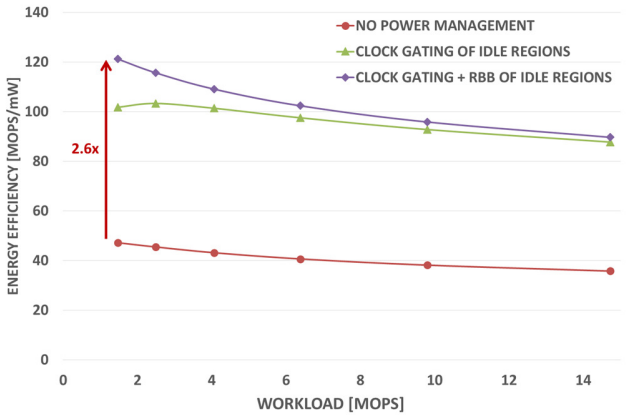

Fig. 8 Impact of power management applied to the idle regions of the cluster when running on a single core. The power considered for calculation of energy efficiency includes the silicon area that can be body biased (i.e. it does not include power consumption of SRAMs).

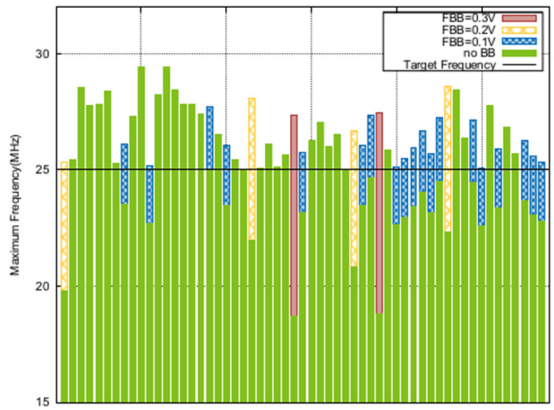

Fig. 9 Maximum measured frequency of 60 samples at Vdd=0.6V. Compensation of die to die frequency variation utilizing forward body bias to reach a frequency larger than the target frequency of 25 MHz.

Figure 8 highlights the effectiveness of the "on-demand" deactivation of the regions of the cluster by means of clock gating and RBB, when executing sequential code on a single core. This technique can be exploited to increase the energy efficiency when parallel execution over multiple cores is not required by the application (i.e., for low workloads), or during the execution of sequential portions of code that cannot be parallelized, due to the Amdahl's law. Eliminating the overhead caused by the idle regions of the cluster it is possible to increase the energy efficiency of the chip by up to 160%.

Although low-voltage operation causes large variation of the maximum operating frequency from die to die, UTBB FD-SOI technology provides an effective knob to fully compensate such variation, namely FBB. The maximum operating frequency measured on 60 chip samples at VDD=0.6V ranges between 19 MHz and 28 MHz, 25 MHz being the average frequency. As shown in Fig. 9 applying FBB ranging from 0.1V to 0.3V to the slow chips allows to bring the maximum frequency of all the 60 dies over the target frequency of 25 MHz. Compensation is achieved with no increase of dynamic power (as opposed to compensation through supply voltage).

A micrograph of the SoC and a comparison with other low-power processors are shown in Fig. 10 and Fig. 11. The proposed SoC energy efficiency (GOPS/W) is 1.4x to 3.7x

	[2]	[3]	[4]	This Work
Technology	CMOS 65nm LP/GP	CMOS 28nm LP	FD-SOI 28nm flip well	FD-SOI 28nm conventional well
Data format	16-bit	32-bit VLIW	32-bit VLIW	32-bit
# of cores	1	1	1	4
I$/D$/L2	64B/2K/16K	16K/32K/256K	4K/4K/n.a.	1Kx4/16K/16K
Voltage range (memories)	0.4V (1.0V)	0.6V - 1.05V	0.4V - 1.3V	0.44V – 1.2V (0.54V – 1.2V)
Max frequency	25 MHz	1.2 GHz	2.6 GHz	475 MHz
Best power density	7.7 µW/MHz	58 µW/MHz	62 µW/MHz	65 µW/MHz
Best performance	12.5 MOPS[1]	3 GOPS[3]	2.6 GOPS	1.8 GOPS
System energy efficiency (MAX)	64.5 GOPS/W[1]	43.1 GOPS/W[3]	16.1 GOPS/W	60 GOPS/W
Core energy efficiency[2] (MAX)	129 GOPS/W[1]	n.a.	n.a.	185 GOPS/W

[1] Normalized to 32-bit operations [2] Does not include SRAMs power [3] An average IPC of 2.5 is assumed

Fig. 10 Comparison with recent ULP and wide-voltage-range processors.

Technology	UTBB FDSOI 28nm
Transistors	Conventional well
Core area	1.5 mm^2
Gates	180K
Memory	72 x 4 Kbit
VDD range	0.44V - 1.2V
BB range	-1.8V - 0.9V
Frequency range	NO BB: 0.74 - 452 MHz FBB: 1.8 – 475 MHz
Power range	NO BB: 0.1 - 119 mW FBB: 0.11 - 127 mW

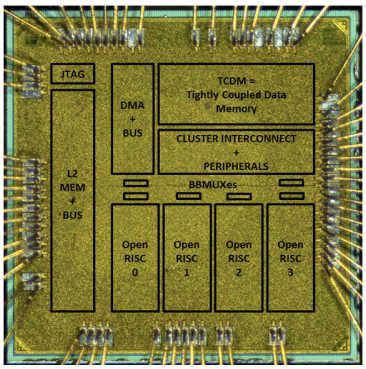

Fig. 11 Die micrograph and main features.

larger than other processors optimized for near-threshold operation with comparable GOPS [3][4]. The chip also outperforms by 144x while achieving a comparable energy efficiency with respect to a leading-edge near-threshold RISC single-issue 16-bit processor optimized for extremely low power applications [2]. Energy efficiency surpasses previously works by more than 43%, when considering only the power consumed in the core silicon area that can be body biased.

Acknowledgments

This work is supported by the European FP7 ERC Advanced project MULTITHERMAN (g.a. 291125) and by the YINS RTD project (no. 20NA21 150939), evaluated by the Swiss NSF and funded by Nano-Tera.ch with Swiss Confederation financing. We thank STMicroelectronics for chip fabrication.

References

[1] D. Jacquet et. al. "2.6GHz ultra-wide voltage range energy efficient dual A9 in 28nm UTBB FD-SOI", VLSI Technology (VLSIT), 2013 Symposium on, pp.C44, C45, 11-13 June 2013.

[2] D. Bol, et al., "A 25MHz 7µW/MHz Ultra-Low-Voltage Microcontroller SoC in 65nm LP/GP CMOS for Low-Carbon Wireless Sensor Nodes", ISSCC 2012, pp. 490-491.

[3] M. Saint-Laurent, et. al., "A 28nm DSP Powered by an On-Chip LDO for High-Performance and Energy-Efficient Mobile Applications", ISSCC Dig. Tech. Papers, 2014, pp. 176 – 179.

[4] R. Wilson, et. al., "A 460MHz at 397mV, 2.6GHz at 1.3V, 32b VLIW DSP, Embedding FMAX Tracking", ISSCC Dig. Tech. Papers, 2014, pp. 452 – 455.