A Shared-Well Dual-Supply-Voltage 64-bit ALU

Yasuhisa Shimazaki, *Member, IEEE*, Radu Zlatanovici, and Borivoje Nikolić

***Abstract*—A shared n-well layout technique is developed for the design of dual-supply-voltage logic blocks. It is demonstrated on a design of a 64-bit arithmetic logic unit (ALU) module in domino logic. The second supply voltage is used to lower the power of noncritical paths in the sparse, radix-4 64-bit carry-lookahead adder and in the loopback bus. A 3 mm^2 test chip in 0.18-μm 1.8-V five-metal with local interconnect CMOS technology that contains six ALUs and test circuitry operates at 1.16 GHz at the nominal supply. For target delay increase of 2.8% energy savings are 25.3% using dual supplies, while for 8.3% increase in delay, 33.3% can be saved.**

***Index Terms*—Adder, ALU, CMOS, dual-supply design, low power.**

I. Introduction

IN ORDER to meet the constant demand for performance enhancement, the cycle time in high-performance digital systems has been steadily reduced. The frequency increase trend is stronger than that provided by scaling of the technology, due to reduction of logic depth between the pipeline stages and slower decrease in supply voltages. As a result, the power dissipation and chip power density have been almost tripling in each new lead processor generation [1]–[3]. Power dissipation and power density are now primary design constraints in high-performance systems because of the thermal design, heat removal and power delivery. On the other end of the spectrum, reducing the energy consumption is essential for battery-operated portable applications.

Several techniques have been proposed for trading off the speed and power savings through transistor sizing, supply voltage scaling, and threshold optimization [4]. A reduction in the supply voltage of a circuit decreases power dissipation, but degrades speed performance. In many signal processing applications, the throughput can be recovered through increased pipelining and parallelism [5], which comes at the expense of increased latency. However, most microprocessor datapaths are latency constrained, while targeting very high clock rates.

In order to avoid performance degradation, the supply voltage can be selectively lowered by using a dual-supply technique [6], whereby a second, lower voltage is supplied to noncritical timing paths. The supply voltages can be assigned either per cell [7] or per block [8]. Reported dual-supply designs were applied to standard cell-based ASICs and achieved power savings from 30% to 70% over the reference single-supply design [7], [8]. Additionally, this second supply voltage can be employed to selectively reduce the power of gates that drive large switched capacitances with small impact on overall system speed. Examples of use of low supply voltage to drive large capacitances include low-swing bus drivers [9] and dual-supply clock distribution [10], [11]. However, previous applications of this technique often came with a large performance penalty [11]. The main issues in using dual supplies in high-performance systems are the cost of level conversion, the delivery of the second supply voltage, and the layout issues associated with it.

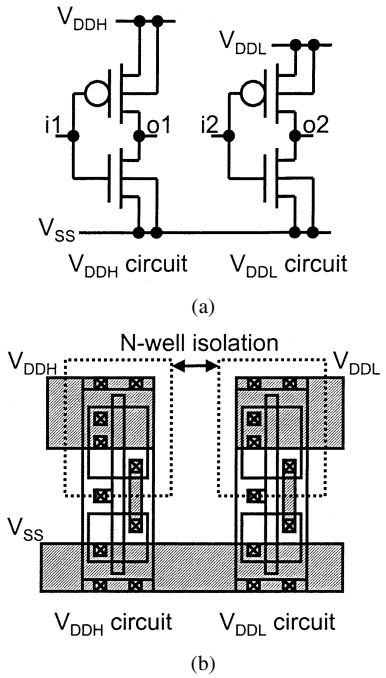


Fig. 1. Conventional dual-supply circuits. (a) Circuit schematic. (b) Layout, indicating necessary n-well isolation.

This paper presents the design of a dual-supply high-performance datapath and analyzes the details of our design from [12]. In Section II, layout issues in a conventional dual-supply design are described and a shared n-well dual-supply technique that is especially suitable for a dual-supply datapath design is proposed. A circuit design of an energy efficient dual-supply ALU is described in Section III. Section IV discusses the area savings with the shared-well technique, and Section V presents a test chip implementation and measured results. Finally, we conclude the paper in Section VI.

II. Layout Issues in Dual-Supply Design

Fig. 1 shows a conventional dual-supply layout style, where high supply (V_{DDH}) and low supply (V_{DDL}) are applied to two

Manuscript received July 1, 2003; revised October 1, 2003. This work was supported by Hitachi Ltd. and MARCO Center for Circuit and System Solutions (C2S2).

Y. Shimazaki is with the Renesas Technology Corporation, Tokyo 187-8588, Japan (e-mail: shimazaki.yasuhisa@renesas.com).

R. Zlatanovici and B. Nikolic are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720-1770 USA (e-mail: bora@eecs.berkeley.edu).

Digital Object Identifier 10.1109/JSSC.2003.822775

0018-9200/04$20.00 © 2004 IEEE

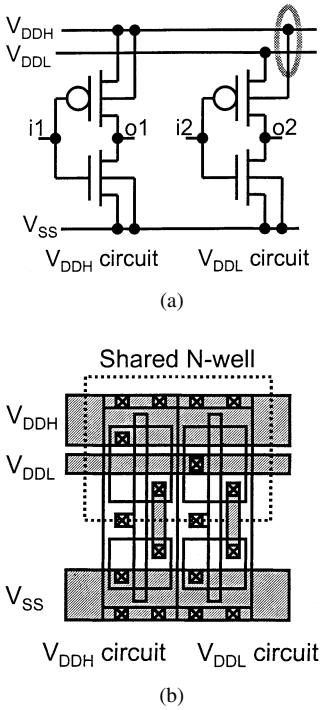


Fig. 2. Shared n-well dual-supply technique. (a) Circuit schematic. (b) Layout view.

neighboring cells. Because the two n-wells are tied to different supply voltages, they have to be separated. Placing the cells in the same row incurs a large layout penalty, which corresponds to the minimum spacing of two wells under different biases. Placing the cells in separate rows achieves the required well separation [6], [7], but results in an impractical layout for the datapath design.

Fig. 2 shows circuit schematic and layout examples of a shared n-well dual-supply technique that is better suited to the datapaths. The power supply is split into V_{DDH} and V_{DDL} rails. The n-well is always tied to V_{DDH}, while the cells are supplied from either V_{DDH} or V_{DDL} by simple via placement. Both V_{DDH} and V_{DDL} cells can be placed in the same row, making this an area-efficient technique with no area overhead in a datapath. The shared-well layout technique provides substantial area savings compared to layout techniques that separate the V_{DDH} and V_{DDL} supply domains, either by row assignment or by well spacing within the same row. The main disadvantages of this style are reduced drive current, increased threshold voltage of the pMOS transistors that affect their driving performance, and possible issues with power routing. However, they can be addressed through careful design.

Fig. 3 shows an example of a shared n-well datapath cell. The power rail is split into two, V_{DDH} and V_{DDL} rails, and they are embedded in the cell layout. This layout exploits local interconnect technology, available in this process, for the contact to V_{DDH} rail. Since the width of the rails is thinner than the conventional one, their resistance increases. Although the dual-supply technique reduces the total power, the length of the rows between power straps has to be limited to avoid increased IR drop. Added space between the two supply rails would affect cell layout density for small cell heights. However, in datapaths, the cell height is usually determined by architectural and performance requirements; therefore, the datapath circuit cells, as in this case, can be made tall enough to avoid any loss in density.

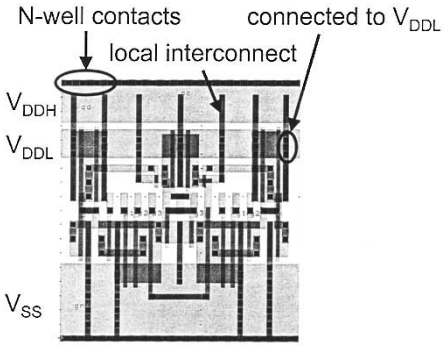


Fig. 3. Layout example of a shared n-well cell connected to V_{DDL}.

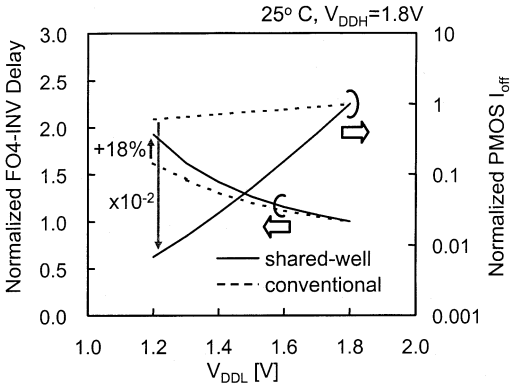


Fig. 4. Delay degradation and leakage reduction of shared n-well V_{DDL} cells compared to conventional layout technique.

In the shared n-well technique, the delay of V_{DDL} circuits is additionally increased due to negative back-biasing of pMOS transistors whose well is tied to V_{DDH}. Fig. 4 shows the simulated fanout-of-4 inverter (FO4-INV) delay and subthreshold current of a pMOS transistor. As the V_{DDL} voltage decreases, the delay increases, resulting in a speed degradation of 18% at 1.2 V, compared to a conventional, nonback-biased V_{DDL} circuit. Since this increase in delay significantly affects the performance of conventional CMOS logic, domino logic is the preferred circuit style for this dual-supply approach. As a side benefit, the pMOS leakage is reduced by two orders of magnitude.

III. Dual-Supply ALU Design

Integer execution units are frequently performance critical, with power densities amongst the largest of all units in a microprocessor chip. Fig. 5 shows a block diagram of a 64-bit ALU module implemented in domino logic, employing the proposed dual-supply technique. The ALU module, similar to [13] and [14], consists of an ALU, an output buffer, and an input operand selector. There are six ALU modules in [13], with the cycle time set by the single-cycle operand bypassing loop. Each of the ALUs can take operands either from the register files, cache, or have them forwarded from the outputs of the other ALUs. The ALU can execute arithmetic (ADD/SUB) and logic (AND/OR/XOR) functions.

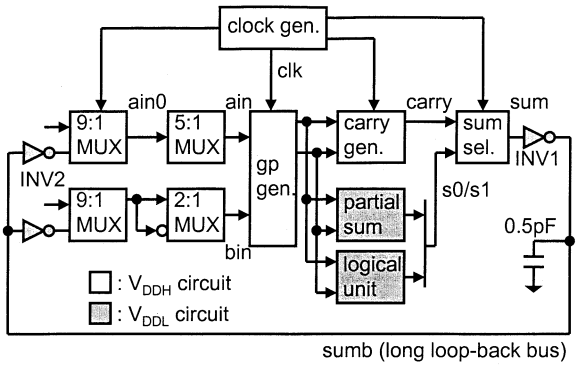


Fig. 5. Block diagram of a 64-bit ALU, indicating portions implemented with high and low supply voltages.

Carry-lookahead adders are commonly used in high-performance datapaths. A common implementation for the carry-lookahead adder is a full Kogge–Stone parallel-prefix tree [15]. In full parallel-prefix trees, all carries are computed in parallel. They are characterized by a regular structure, where each of the parallel prefix operators has the same structure, same number of inputs, and same fanout. Radix-2 trees merge two carries at each level and radix-4 trees merge four input carry signals at each level. Radix-4 trees have lower logic depths, but require more complex gates at each level. At low overall fanouts they are faster than radix-2 trees [16]. A full radix-4 tree is shown in Fig. 6(a), where G/P are generate and propagate signals, G4/P4 and G16/P16 are group generate and group propagate signals at each stage, and G64 is a final carry signal. Sparse trees compute only a subset of final carries and select the correct sums based only on these carries. For example, [13] computes only the odd carries and selects the outputs sums using them.

In this design, every fourth carry is calculated in the tree, as shown in Fig. 6(b). While the full radix-4 tree suffers from a large number of complex carry-merge gates, the sparse implementation significantly reduces the gate and wire count, but increases the complexity of the sum computation. All sums are precomputed in a lookahead fashion, implemented in complex domino and static gates.

Fig. 7 shows a comparison of the area-delay performance of full and sparse trees. The critical path of each of the adders is modeled using a linear delay model [17], and each point in the graph is obtained by minimizing the delay under total transistor width constraints. The total transistor width closely approximates the area of the design in fixed-height datapaths, and is also proportional to the total energy.

The critical path of the ALU goes through the adder carry path and is implemented in the V_{DDH} domain. The partial sum generation and the logical unit are noncritical and are supplied from V_{DDL}. The sparse tree adder is particularly suitable for a dual-supply implementation, where the complex sum precompute gates can be placed in the V_{DDL} domain. In this implementation, they are in the critical path when V_{DDL} is lowered to 1.2 V. While the V_{DDH} gates can freely drive V_{DDL} gates, returning to V_{DDH} domain requires level conversion. Domino level converters, similar to [18], are used in the sum selectors and the 9:1 multiplexers. The driving transistors are made wide

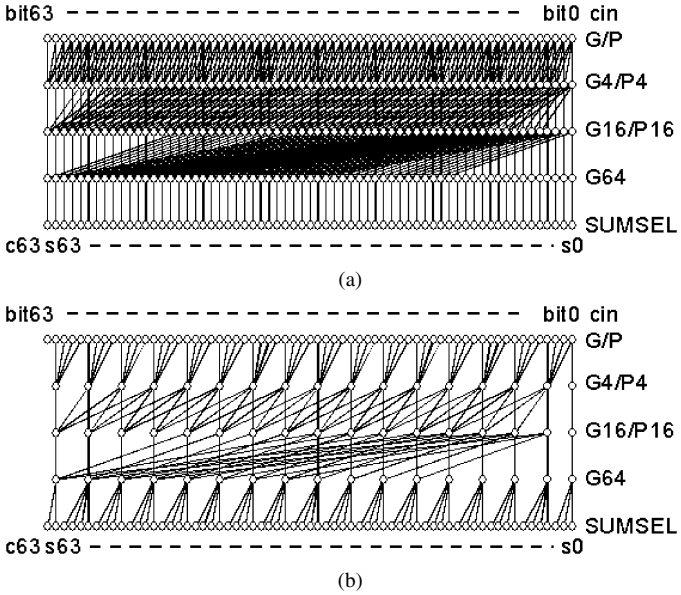


Fig. 6. Radix-4 64-bit adder trees. (a) Full tree. (b) Sparse tree with a sparseness of 4.

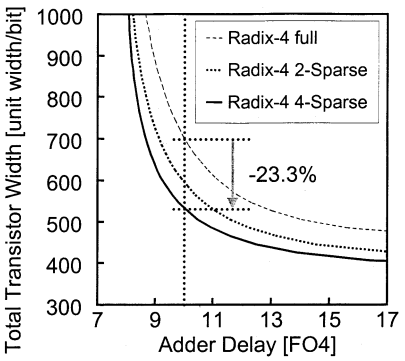


Fig. 7. Comparison of area-delay performance of full and the trees with sparseness levels of 2 and 4.

enough to overcome any contention with the keeper. Detailed circuit schematics of the output buffer and the 9:1 multiplexer are shown in Fig. 8(a). Since the loop-back bus *sumb* has a large capacitive load, placing its driver in the V_{DDL} domain saves power. Since *sum* is a monotonic rising signal, the delay of INV1 does not increase, because the nMOS transistor in the V_{DDL} circuit does not suffer from the negative back-biasing.

In dual-supply design, coupling noise is one of the critical issues. In order to increase the noise immunity, the receiver INV2 is placed in close proximity of the 9:1 multiplexer. The output of INV2, which is a V_{DDL} signal, is converted to V_{DDH} signal by the 9:1 multiplexer. The level conversion is performed very fast, because a domino circuit is a precharge-type structure, where the precharge level is not determined by the input signal but by a precharge control signal *pc*, which is generated from the clock signal. The circuit schematic of the sum select gate is shown in Fig. 8(b). Since the carry generator is a single-rail domino circuit, complementary signals are not available. In order to generate carry complements and to reduce the delay overhead due to level conversion of V_{DDL} signals, a complementary function generator, similar to the one in [19], and the domino level converter are incorporated in one circuit.

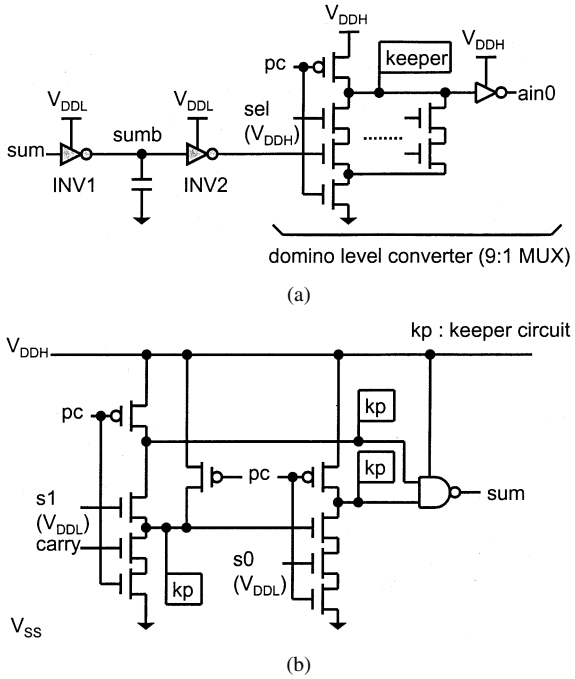


Fig. 8. Level conversion embedded in domino logic. (a) Output buffer and 9:1 multiplexer. (b) Sum select.

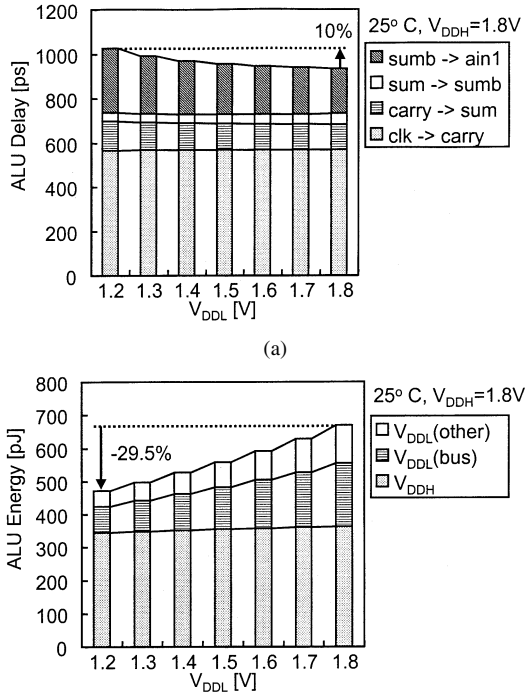


Fig. 9. Simulated performance of dual-supply circuits compared to single supply. (a) Delay. (b) Energy.

The adder circuit implementation uses footless domino gates, with delayed precharge signals. All the timing edges are soft except the two hard edges at the GP block and at the sum selector.

The forwarding interconnect in multi-issue integer execution units is long with a high fanout load; the output buffer has a large power consumption. Fig. 9 shows the performance of the dual-supply ALU compared to the one operated from the V_{DDH}. Lowering the V_{DDL} to 1.2 V lowers the bus driver supply, resulting in 56% energy reduction with 22% delay increase. However, this delay penalty corresponds to only 8% cycle time increase for the complete ALU module. As a result of the dual-supply technique and the domino level converter, 29.5% energy reduction with only 10% delay penalty can be estimated.

IV. Layout Comparison

The width of the bit-slice in the datapath equals 18 metal-1 (M1) tracks. Cell height is chosen to equal to the bit-slice width. Since the carry is computed only for every fourth bit, the sum precompute cells and buffers are placed in unused rows, resulting in a very dense layout. The plan of one bit-slice is shown in Fig. 10(a), with actual cell dimensions indicated. Each box indicates the function of the cell, its width and the voltage level that is being supplied from. The total length of the datapath equals 313 M1 tracks.

To evaluate the advantages of shared-well technique, it is compared to two conventional layout styles. In Fig. 10(b), the V_{DDH} and V_{DDL} domains are separated by a spacing S. This minimum isolation space for two wells under different biases equals four M1 tracks. If this spacing is chosen, the minimum adder width would be 355 M1 tracks, which presents a 13% increase over the shared-well technique. However, this spacing is commonly used for vertical V_{DDH} and V_{DDL} power straps. If the width of the strapping metal is chosen to be five tracks wide, the total adder width would be 397 tracks, presenting a 27% increase. Alternatively, a bit-slice can be split into two rows, as illustrated in Fig. 10(c), with V_{DDH} and V_{DDL} supply lines being shared between neighboring bit-slices. To accommodate the reduced cell height, the area of the cells would have to be slightly increased. Cell dimensions in Fig. 10(c) account for this increase. For example, INV2 cell width is increased from 19 to 40 M1 tracks with height reduced from 18 to nine tracks. The overall adder width would be 401 tracks, an increase of 28%.

V. Test Chip and Measurement Results

A microphotograph of the test chip is shown in Fig. 11. The chip uses a 1.8-V general-purpose 0.18-μm CMOS process, with one-poly five-metal (1P5M) layers and local interconnect technology. The chip includes six ALU modules, to simulate the loading conditions of a six-issue integer execution unit, control circuitry, clock drivers, and test circuitry. An additional capacitance is added to simulate the cache and register file load. The size of the ALU module is 200×760 μm^2, while the overall chip size is 2 mm $\times$ 1.5 mm.

The test circuitry is included on the chip in order to perform the delay measurements around 1 ns. The measurement principle is shown in Fig. 12. The clock is generated by a ring oscillator with variable supply voltage, V_{DD2}. A built-in data generator feeds data into the ALU and the results are stored in two registers clocked by two different clock signals, skewed by a small delay δ. A comparator compares the two stored values, and if they differ, the early register was not able to latch the data.

In order to ensure the same loading and consistent timing during frequency measurements, the loop is broken at the first

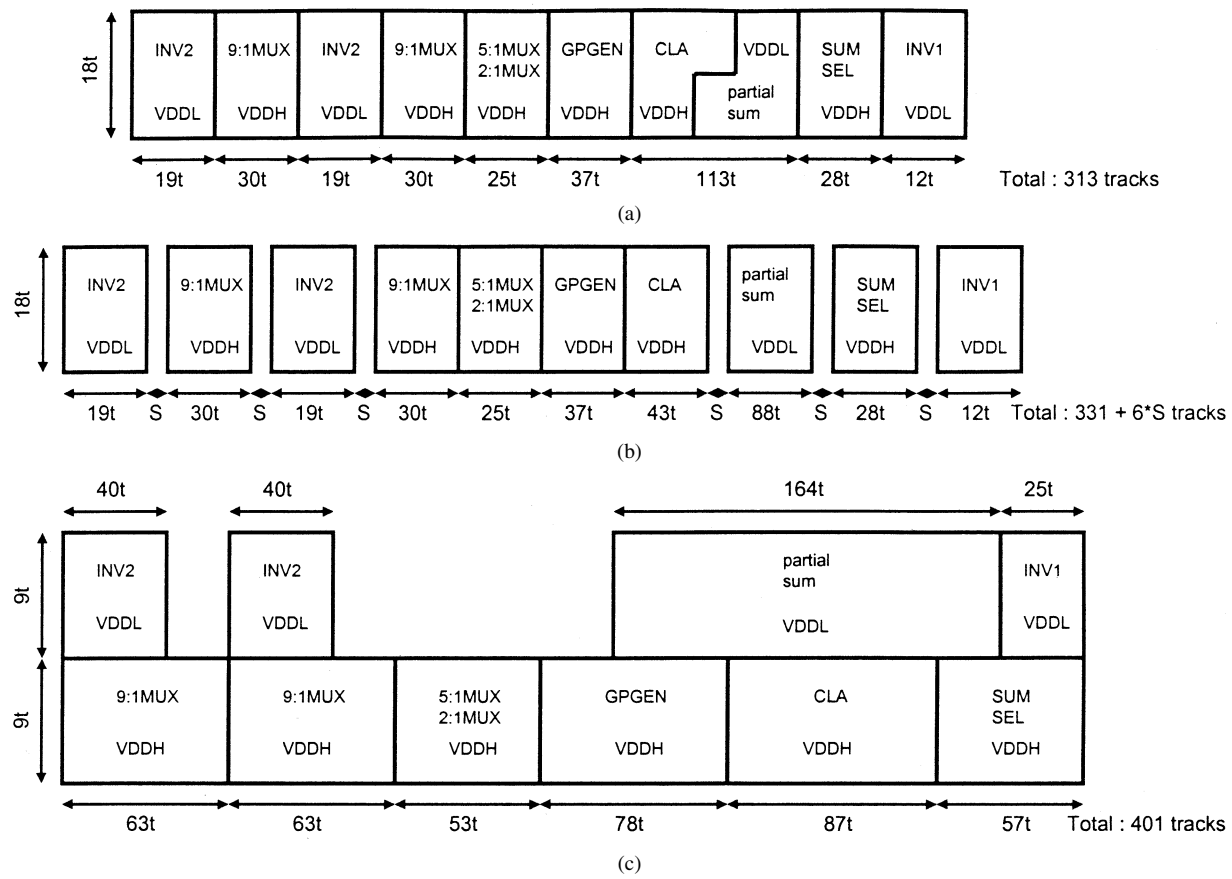


Fig. 10. Bit slice organizations. (a) Shared-well dual-supply technique. (b) Shared row with well spacing. (c) Separate V_{DDH} and V_{DDL} rows.

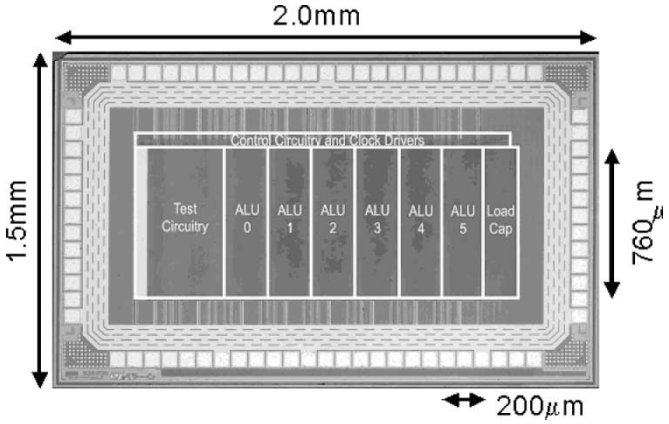


Fig. 11. Die microphotograph.

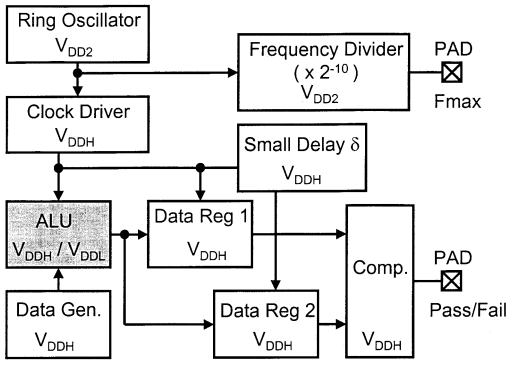


Fig. 12. Block diagram of the test procedure.

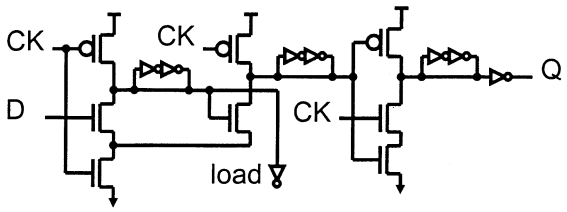


Fig. 13. Test flip-flop with same setup time as GP generator.

hard edge (input of the GP block). To ensure the correct measurement data, a flip-flop is designed such that it has the same input loading and setup time as the original GP gate (Fig. 13). This guarantees that the broken loop measures the exact cycle time of the closed loop. Using this setup, the frequency measurement procedure consists of increasing V_{DD2} (thereby increasing the clock frequency) until the output of the comparator toggles. At that point, the clock frequency is measured easily through a built-in 2^{10} divider. The critical path is $G[19] \rightarrow sum[56] \rightarrow ain[59]$, and the input combination that exercises it is $64'$h00FFFFFFFF80000+$64'$h0000000000080000= $64'$h0100000000000000. Since the ALU uses dynamic gates, it is sufficient to observe just the comparator at the output $(ain[59])$.

To eliminate the test circuitry consumption from the measurements, power is measured by observing the difference in consumption when n or $n+1$ ALUs are clocked.

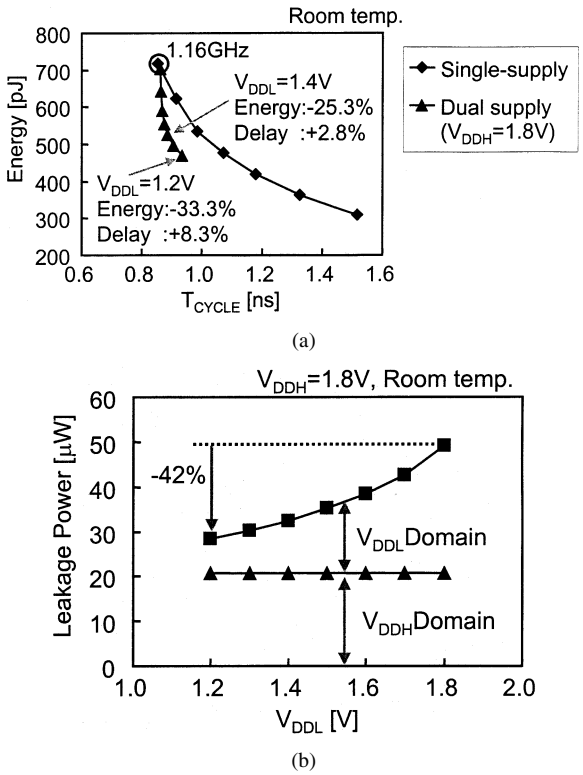


Fig. 14. Measurement results. (a) Energy versus cycle time. (b) Leakage power versus V_{DDL}.

Worst case power is exercised with the input combination that causes all internal dynamic nodes to be evaluated to $0{:}\,64'$hFFFFFFFFFFFFFFFF + $64'$hFFFFFFFFFFFFFFFF = $64'$hFFFFFFFFFFFFFFFE. Leakage is measured at room temperature for the whole chip, with the ring oscillator off, and then divided among the six ALUs and the test circuitry according to their total transistor width.

A simple functionality test is also included on the chip. Several random combinations and their precomputed results are hardwired and compared with the ALU results at runtime using an external low-frequency clock.

With $V_{DDH} = V_{DDL} = 1.8$ V, the chip operates at its nominal frequency of 1.16 GHz, corresponding to 13 FO-4-INV delays. Fig. 14(a) summarizes the effect of the dual-supply operation on circuit speed and energy consumption. The single-supply operation is plotted as a reference where the supply is scaled down to meet the target delay. When the target delay is increased by 2.8%, total energy savings are 25.3% using dual supplies. A delay increase of 8.3% results in an energy savings of 33.3%. In comparison to a single reduced supply operation, the energy savings are 20.2% and 20.9%, respectively. Leakage power reduces by 42% at $V_{DDL} = 1.2$ V, Fig. 14(b), illustrating the effect of the negative back-biasing of pMOS transistors.

VI. Conclusion

A shared n-well dual-supply-voltage technique has been implemented that is suitable for datapath design. An adder circuit using a sparse radix-4 carry tree with sparseness of 4 has been designed to save 23% of energy compared to a conventional radix-4 full tree adder. A low-swing bus and domino level converter can reduce ALU energy with small impact on overall ALU delay, when supplied from a low voltage. The feasibility of high-performance dual-supply design has been demonstrated by a fabrication of a test chip. It operates at 1.16 GHz at the nominal voltage of 1.8 V, and saves 25% energy with 2.8% delay increase, while with 8.3% delay increase, the energy saving is 33%.The design techniques presented in this paper can be used in fine- and coarse-granularity dual-supply designs.

Acknowledgment

The authors thank ST Microelectronics for test chip fabrication.

References

[1] S. Borkar, "Design challenges of technology scaling," *IEEE Micro*, vol. 19, pp. 23–29, July–Aug. 1999.

[2] P. P. Gelsinger, "Microprocessors for the new millennium: Challenges, opportunities and new frontiers," in *IEEE Int. Solid-State Circuits Conf. (ISSCC'01) Dig. Tech. Papers*, San Francisco, CA, Feb. 2001, pp. 22–25.

[3] H. P. Hofstee, "Power constrained microprocessor design," in *Proc. IEEE Conf. Computer Design*, Freiburg, Germany, Sept. 2002, pp. 14–16.

[4] R. W. Brodersen, M. A. Horowitz, D. Marković, B. Nikolić, and V. Stojanović, "Methods for true power minimization," in *Int. Conf. Computer-Aided Design, (ICCAD2002) Dig. Tech. Papers*, San Jose, CA, Nov. 2002, pp. 35–42.

[5] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, pp. 473–484, Apr. 1992.

[6] K. Usami and M. Horowitz, "Clustered voltage scaling for low-power design," in *Proc. Int. Symp. Low Power Electronics and Design (ISLPED)*, Apr. 1995, pp. 3–8.

[7] K. Usami and M. Igarashi, "Low-power design methodology and applications utlizing dual-supply voltages," in *Proc. Asia South Pacific Design Automation Conf.*, Jan. 2000, pp. 123–128.

[8] R. Puri *et al.*, "Pushing ASIC performance in a power envelope," in *Proc. Design Automation Conf. (DAC'03)*, Anaheim, CA, June 2003, pp. 788–793.

[9] Y. Nakagome, K. Itoh, M. Isoda, K. Takeuchi, and M. Aoki, "Sub-1-V swing internal bus architecture for future low-power ULSIs," *IEEE J. Solid-State Circuits*, vol. 28, pp. 414–419, Apr. 1993.

[10] J. Pangjun and S. S. Sapatnekar, "Low-power clock distribution using multiple voltages and reduced swings," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 3, pp. 309–318, June 2002.

[11] R. Krishnamurthy *et al.*, "Dual-supply voltage clocking for 5 GHz 130 nm integer execution core," in *Symp. VLSI Circuits Dig. Tech. Papers*, Honolulu, HI, June 2002, pp. 128–129.

[12] Y. Shimazaki, R. Zlatanovici, and B. Nikolic, "A shared-well dual-supply-voltage 64-bit ALU," in *IEEE Int. Solid-State Circuits Conf. (ISSCC'03) Dig. Tech. Papers*, San Francisco, CA, Feb. 2003, pp. 104–105.

[13] S. Mathew, R. Krishnamurthy, M. Anders, R. Rios, K. Mistry, and K. Soumyanath, "Sub-500 ps 64 b ALUs in 0.18 μm SOI/Bulk CMOS: Design and scaling trends," *IEEE J. Solid-State Circuits*, vol. 36, pp. 1636–1646, Nov. 2001.

[14] E. S. Fetzer *et al.*, "A fully bypassed 6-issue integer datapath and register file on the Itanium-2 microprocessor," *IEEE J. Solid-State Circuits*, vol. 37, pp. 1433–1440, Nov. 2002.

[15] P. M. Kogge and H. S. Stone, "A parallel algorithm for efficient solution of a general class of recursive equations," *IEEE Trans. Comput.*, pp. 786–793, Aug. 1973.

[16] R. Zlatanovici and B. Nikolić, "Power-performance optimal 64-bit carry-lookahead adders," in *Proc. Eur. Solid-State Circuits Conf.*, Sept. 2003, pp. 321–324.

[17] I. Sutherland, B. Sproull, and D. Harris, *Logical Effort: Designing FAST CMOS Circuits*. San Francisco, CA: Morgan Kaufmann, 1999.

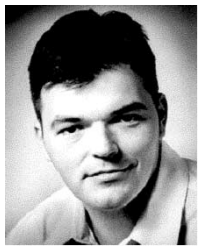
[18] N. Tzartzanis *et al.*, "A 34Word × 64b 10R/6W write-through self-timed dual-supply-voltage register file," in *IEEE Int. Solid-State Circuits Conf. (ISSCC'02) Dig. Tech. Papers*, San Francisco, CA, Feb. 2002, pp. 416–417.

[19] S. Vangal *et al.*, "5-GHz 32-bit integer execution core in 130-nm dual-V_T CMOS," *IEEE J. Solid-State Circuits*, vol. 37, pp. 1421–1432, Nov. 2002.

Yasuhisa Shimazaki (M'99) received the B.E. and M.E. degrees in electrical engineering from Nagoya University, Nagoya, Japan, in 1991 and 1993, respectively. He joined Hitachi, Ltd. as a VLSI circuit engineer upon graduation. In 2000, he studied at University of California, Berkeley as Visiting Industrial Fellow. Since 2003, he has been engaged in development of low-power microprocessors at Renesas Technology Corp. Currently, he is on loan to SuperH (Japan), Ltd., where he is in charge of ultra low-leakage SRAM design. His research interests include low-power and high-speed digital integrated circuits.

Radu Zlatanovici received his B.S and M.S. degrees from Politehnica University Bucharest, Romania in 1999 and 2000 respectively. In 2000 he joined the University of California at Berkeley where he received an M.S. degree in 2002. He is currently working toward his Ph.D. degree at the same university.

He was on the faculty of Politehnica University Bucharest from 1999 to 2000. In 2002 and 2003 he interned at IBM T.J. Watson Research Center, Yorktown Heights, NY working on power—performance tradeoffs for pipelined digital circuits. His research interests include high-speed and low-power arithmetic circuits and design optimization in the power—performance space.



Borivoje Nikolić (S'93-M'99) received the Dipl.Ing. and M.Sc. degrees in electrical engineering from the University of Belgrade, Yugoslavia, in 1992 and 1994, respectively, and the Ph.D. degree from the University of California at Davis in 1999.

He was on the faculty of the University of Belgrade from 1992 to 1996. He spent two years with Silicon Systems, Inc., Texas Instruments Storage Products Group, San Jose, CA, working on disk-drive signal processing electronics. In 1999, he joined the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley as an Assistant Professor. His research activities include high-speed and low-power digital integrated circuits and VLSI implementation of communications and signal-processing algorithms. He is co-author of the book *Digital Integrated Circuits: A Design Perspective*, 2nd ed, Prentice-Hall, 2003.

Dr. Nikolić received the NSF CAREER award in 2003, College of Engineering Best Doctoral Dissertation Prize and Anil K. Jain Prize for the Best Doctoral Dissertation in Electrical and Computer Engineering at University of California at Davis in 1999, as well as the City of Belgrade Award for the Best Diploma Thesis in 1992.