CACTI-P: Architecture-Level Modeling for SRAM-based Structures with Advanced Leakage Reduction Techniques

Sheng Li[†] , Ke Chen[‡†], Jung Ho Ahn[§] , Jay B. Brockman[‡] , Norman P. Jouppi[†]
[†]Hewlett-Packard Labs, [‡]University of Notre Dame, [§]Seoul National University
[†]{sheng.li4, kec, norm.jouppi}@hp.com, [‡]{kchen2, jbb}@nd.edu, [§]gajh@snu.ac.kr

***Abstract*—This paper introduces CACTI-P, the first architecture-level integrated power, area, and timing modeling framework for SRAM-based structures with advanced leakage power reduction techniques. CACTI-P supports modeling of major leakage power reduction approaches including power-gating, long channel devices, and Hi-k metal gate devices. Because it accounts for implementation overheads, CACTI-P enables in-depth study of architecture-level tradeoffs for advanced leakage power management schemes. We illustrate the potential applicability of CACTI-P in the design and analysis of leakage power reduction techniques of future manycore processors by applying nanosecond scale power-gating to different levels of cache for a 64 core multithreaded architecture at the 22nm technology. Combining results from CACTI-P and a performance simulator, we find that although nanosecond scale power-gating is a powerful way to minimize leakage power for all levels of caches, its severe impacts on processor performance and energy when being used for L1 data caches make nanosecond scale power-gating a better fit for caches closer to main memory.**

Keywords: Leakage power management, power-gating, circuit modeling, SRAM, cache, manycore processor

I. Introduction

Static random-access memory (SRAM) based structures, including caches, lookup tables, and storage buffers, occupy ever-increasing die area of high-performance VLSI chips such as microprocessors, and thus contribute a significant portion of total chip standby power. Moreover, as Moore's Law has driven CMOS technology well into the nanoscale regime, controlling SRAM leakage power becomes more and more challenging; despite all the efforts and progress in technology including silicon on insulator (SOI), strained-silicon, and multi-gate devices, the leakage power density still increases proportionally as technology advances [14]. Therefore, managing SRAM leakage power has become increasingly important to meet the power budget.

Almost all SRAM-based arrays in modern microprocessors are equipped with advanced leakage power management logic. Figure 1 shows the power breakdown of the Intel Xeon Tulsa and Penryn processors [4], [13]. As shown in the figure, when no leakage control techniques are used, the leakage power of the last level caches occupies 63% and 56% of the total leakage power of Xeon Tulsa and Core 2 Penryn processors respectively, corresponding to 30% and 20% of the total power of the processors. With advanced leakage power management, the leakage power of the last level caches are reduced by more than 6×.

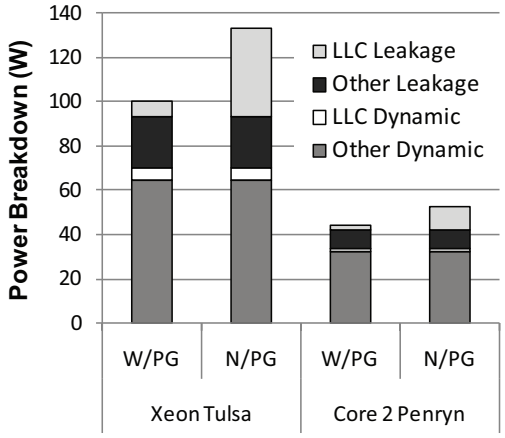


Fig. 1. Power breakdown of two modern processors. Leakage power of last level caches is shown for both with and without leakage power reduction techniques.

The use of SRAM leakage power management schemes affects chip design at the system level. The designer needs to consider not only the power reduction efficiency but also the area and timing overhead for adding the leakage power management unit. More importantly, the wakeup time and wakeup energy of the designs and their impact at the system level must be carefully evaluated.

Our ability to propose, design, and evaluate new power management techniques and their system implications is currently limited by the availability and quality of appropriate system-level tools. At the circuit level, [3], [7], [11] proposed sleep transistor designs for generic circuit blocks. Real cache and RAM implementations with power management units have been demonstrated in modern industry designs [4], [13], [16] . However, there are no architecture-level tools that can perform fast yet accurate evaluation of power control techniques for memory arrays. CACTI [18] was the first tool to provide rapid power, area, and timing estimates of RAM structures for computer architecture research, and following a series of improvements [1], [15] has remained the most popular. McPAT [9], [10] uses the same analytical methodology to simultaneously model power, area, and timing of multicore and manycore processors.

In this paper, we introduce CACTI-P, an extension to CACTI, that provides the first in-depth model of leakage power management techniques for SRAM-based arrays. It models the detailed design of the power management unit including different circuit design styles and sleep transistor sizing. Thus, CACTI-P can correctly evaluate the power, area, and timing overhead of adding the power management units, including the penalties of wakeup latency and wakeup energy.

This material is based upon work supported by the Department of Energy under Award Number DE - SC0005026. The disclaimer can be found at http://www.hpl.hp.com/DoE-Disclaimer.html

Ke Chen and Jay Brockman are partially supported by the C2S2 Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity.

Jung Ho Ahn was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0003683).

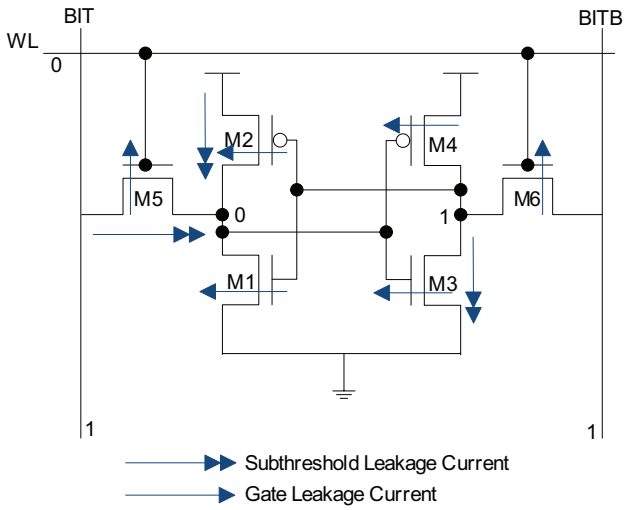


Fig. 2. Leakage current paths in an SRAM cell in an idle state. Bitlines are precharged to V_{cc}.

It also models the device-level power-saving technologies including long channel and high-k metal gate devices. CACTI-P enables architects to consistently quantify the cost of leakage power reduction techniques. By combining CACTI-P with a performance simulator, this paper explores nanosecond scale power-gating schemes in future manycore processors. Our results show that nanosecond scale power-gating reduces the leakage power of both L1 data (L1D) and L2 caches significantly but causes degeneration of the system performance and energy-delay product (EDP) because of the wakeup latency and energy overhead. Applying power-gating only to L2 caches is the best option; it improves the system EDP by 18% on average for multiprogrammed workloads composed of the SPEC CPU 2006 benchmark suite [6].

II. Leakage Power and Management Techniques

The leakage power in CMOS is consumed because of leakage current through the transistors which, in reality, function as "imperfect" switches. The magnitude of the leakage current is proportional to the width of the transistor and also depends on the logical state of the device. There are two major leakage mechanisms. The first type of leakage, subthreshold leakage, occurs when a transistor in the off state actually allows a small current to pass between its source and drain. The second type, gate leakage, is the current that leaks through the gate terminal. Figure 2 shows the subthreshold leakage and gate leakage paths in an SRAM cell in the idle state with "0" being stored. Since an SRAM has multiple leakage paths in each cell (as shown in the figure), its leakage power is a serious problem.

The total leakage power dissipation of CMOS circuits can be expressed as shown in Equation (1),

$$P_{leakage} = \underbrace{V_{cc} \times I_{sub}}_{SubthrethodLeakage} + \underbrace{V_{cc} \times I_g}_{GateLeakage} \quad (1)$$

where V_{cc} is the voltage supply, I_{sub} is subthreshold leakage current, and I_g is the gate leakage current. While subthreshold leakage has been a problem since 90nm technologies, the gate leakage has also become a serious challenge at 45nm [5]. Based on the equation there are two approaches to reduce the leakage power: 1) reduce the supply voltage; 2) reduce the leakage current density.

The use of power-gating with sleep transistors is the chief approach to reduce the supply voltage (Although the change of supply voltage also affects the leakage current, the leakage current reaches its full value after the supply voltage is around 50mV [17]). A sleep transistor is either a PMOS or NMOS high V_t transistor that connects an external power supply to an internal circuit power supply which is commonly called a virtual power supply. The sleep transistor is controlled by a power management unit to shift the voltage level of the virtual power supply so that the circuit can be put into different power-saving states when idle. A PMOS sleep transistor (a header switch) is used to control the V_{cc} supply and provides a virtual V_{cc}. An NMOS sleep transistor (a footer switch) controls the V_{ss} supply and creates a virtual ground. Footers, headers, or both can be used in circuit designs to achieve optimal trade-offs between performance and overhead. PMOS transistors are less leaky than NMOS transistors of the same size, leading to a better leakage power reduction. The advantage of a footer switch is its high driving current and hence smaller area.

Long channel devices and Hi-k metal gate devices are the device-level techniques used to reduce intrinsic leakage current density. Equation 2 [17] shows the subthreshold current.

$$I_{ds} = I_{ds0} e^{\frac{V_{gs}-V_t}{n v_T}} (1 - e^{\frac{-V_{ds}}{v_T}}) \quad (2)$$

where I_{ds0} is the current at threshold voltage V_t, which is a constant at a given technology node, V_{gs} is the gate-source voltage, V_{ds} is the drain-source voltage, v_T is the thermal voltage, and n is a constant coefficient. $V_t = V_{t'} - V_{DIBL}$, where $V_{t'}$ is the intrinsic threshold voltage, and V_{DIBL} is the threshold voltage reduction caused by the drain induced barrier lowering (DIBL) effect. Subthreshold current is exponentially dependent on V_t, and V_{DIBL} decreases exponentially with the channel length [19]. Thus, increasing channel length can significantly reduce subthreshold current by increasing V_t. On the other hand, since the saturation current of a device is proportional to channel length, performance decreases linearly with the increase of channel length. The use of 10% longer channel devices can reduce subthreshold current by around 3× with less than 10% performance loss [13].

Equation 3 [12] shows the gate leakage current density.

$$J_{DT} = A E_{ox}^2 exp(-\frac{B(1 - (1 - \frac{V_{ox}}{\phi_{ox}})^{3/2})}{E_{ox}}) \quad (3)$$

where ϕ_{ox} is the barrier height for electrons in the conduction band, A and B are constant coefficients accounting for both the electron properties and ϕ_{ox}, and $E_{ox} = V_{ox}/t_{ox}$ is the gate electric field determined by voltage on the gate dielectric (V_{ox}) and the thickness of gate dielectric (t_{ox}). Equation 3 shows that the gate leakage current density increases rapidly with the increase of E_{ox}. Hi-k gate dielectric has been introduced to keep the same gate capacitance while not decreasing t_{ox}, thus resulting in a relatively small E_{ox} compared to conventional silicon oxide, resulting in significant gate leakage reduction.

III. CACTI-P

CACTI-P is the first modeling framework for SRAM designs with integrated leakage power management. Figure 3 shows a block diagram of the CACTI-P framework. The key components of CACTI-P include:

- power-gating models that perform placement and sizing for sleep transistors. The models also decide the types (PMOS or NMOS) of sleep transistors and compute their area, timing, and power overhead.

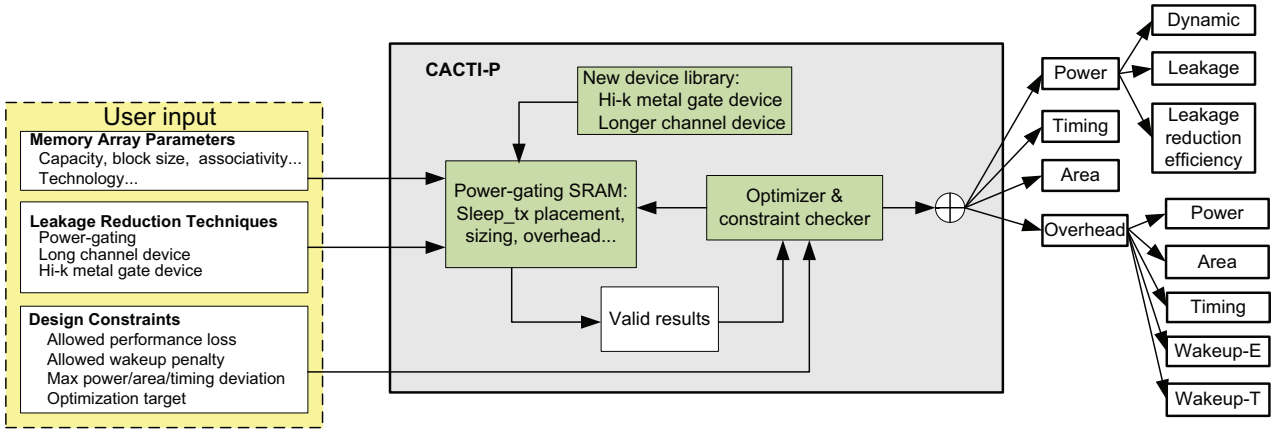


Fig. 3. Block diagram of the CACTI-P framework.

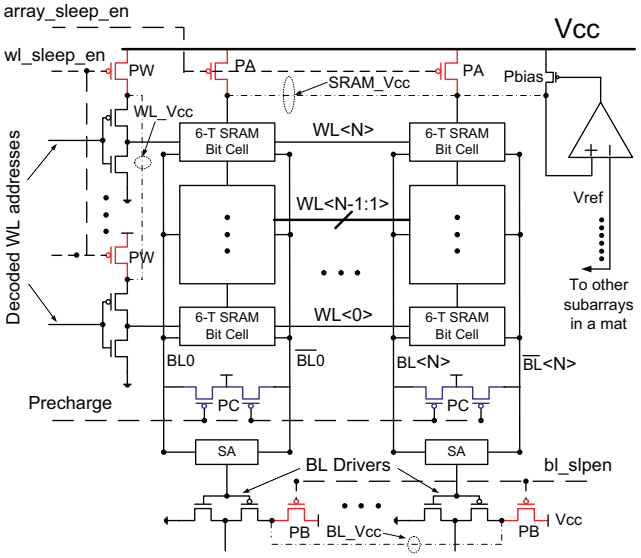


Fig. 4. CACTI-P circuit-level model for multi-mode power-saving states based on the Intel design [16].

- technology models for long channel devices and Hi-k devices to compute leakage reduction when using these techniques.
- an optimizer/checker that performs design space exploration and finds the optimal design that fits in the design constraints for both leakage reduction techniques and overall memory power, area, and timing.

The leakage power reduction techniques, the memory array configurations, and the design constraints are the inputs to CACTI-P. The design constraints related to leakage power reduction include the performance loss and wakeup latency caused by the sleep transistors, the allowed deviation from the best overall latency, and other parameters as shown in Figure 3.

A. Power-gating modeling

Figure 4 shows the modeled power-gating techniques at the SRAM subarray level in CACTI-P. A subarray is the smallest entity inside an SRAM array, and each access involves one or more subarrays. CACTI-P models power-gating at all major potential leakage components inside SRAMs, including memory cell power-gating, wordline power-gating, bitline floating, and bitline I/O power-gating. Power-gating for CAMs and fully-associative caches are also modeled in CACTI-P, including additional searchline and matchline floating using the same modeling techniques as used for bitline floating. CACTI-P incorporates models for state-of-the-art design techniques, including: 1) the placement and topology of sleep transistors, 2) the supply voltage of idle circuit blocks, 3) the types and sizes of sleep transistors, and 4) power-gating granularity, as used in [4], [13], [16].

Power-gating relies on reducing the supply voltage of idle circuit blocks. However, SRAM cells will lose data, if the supply voltage is reduced below the *retention voltage*. CACTI-P models two power-gating states: the sleep (or standby) state and the shut-down state. While the shut-down state reduces supply voltage almost to zero and loses all information stored in the SRAM array, the sleep state reduces the power supply to the minimum retention voltage so that the memory cells retain data while consuming less leakage power. The retention voltage depends on the achievable static-noise-margin of the two inverters in an SRAM cell. To determine the retention voltage of the SRAM cells, designers usually perform extensive on-die measurements, considering variations in the threshold voltage across the chip resulting from the fabrication process variation. CACTI-P chooses the retention voltage (SRAM_v_{ccmin}) based on reported values in industrial designs [4], [5], [13], [16], [20], as shown in Table I. For wordline drivers and bitline drivers that are pure logic gates, the minimum virtual supply voltage WL_V_{ccmin} and BL_V_{ccmin} for sleep states are determined by the sizes and threshold voltages of transistors in the drivers.

Technology (nm)	$V_{ccmin}(V)$	$V_{cc}(V)$
65	0.7	1.1
45	0.65	1.0
32	0.6	0.9
22	0.55	0.8

TABLE I
RETENTION VOLTAGE OF SRAM CELLS FROM 65NM TO 22NM TECHNOLOGY NODES.

CACTI-P also models bitline floating that reduces the leakage power caused by the access transistors inside an SRAM cell. Bitline floating is modeled by assuming that bitline precharge transistors (PCs in Figure 4) are turned off when the SRAM is in the sleep state and turned back on once the SRAM returns to the active state. Thus,

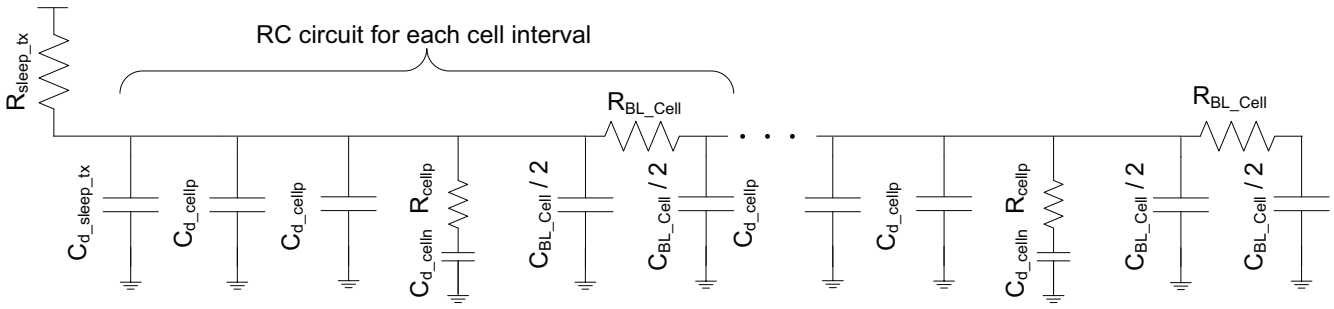


Fig. 5. Array wakeup RC circuit model.

leakage through the access transistors is reduced by dynamically lowering the bitline voltage in the standby state with bitline floating. The voltage on the floating bitline is determined by the capacitance of the bitline and the sleep time of the bitline before it wakes up again.

As shown in Figure 4, sleep transistors are modeled for memory cells, wordline drivers, and bitline I/O drivers. For arrays, CACTI-P assumes one sleep transistor (PA) for each bitline inside a subarray. CACTI-P assumes a sleep transistor (PW) for each wordline driver and a sleep transistor (PB) for each bitline I/O driver. As shown in Figure 4, the groups of the sleep transistors (connected by dotted lines in the Figure 4) in the arrays, wordline drivers, and bitline I/O drivers are modeled as distributed sleep transistor networks (DSTNs) [11]. DSTN achieves smaller overhead than traditional module-based [7] and cluster-based [3] sleep transistor designs. For the interconnect drivers (not shown in the figure), CACTI-P uses DSTNs only for sleep transistors at the same stage and within the same link group (address, data-in, and data-out links) to ease routing complications. Using these assumptions, CACTI-P performs automatic sleep transistor sizing inside each DSTN. Note that although PMOS transistors are used as sleep transistors as shown in Figure 4, CACTI-P models both NMOS and PMOS transistors as sleep transistors and selects the better option considering overall results in power, area, and timing during the automatic optimization process.

CACTI-P sizes the sleep transistors based on two design constraints set by the user: the allowed performance loss when the circuit is active and the allowed wakeup latency. CACTI-P first sizes the DSTN-based sleep transistors using Equations 4 and 5 [11].

$$W_{st} = (1+\beta)\frac{MSSC(ckt)L}{N\delta\mu C_{ox}(V_{cc}-V_{tL})(V_{cc}-V_{tH})} \quad (4)$$

$$\delta = 1-\frac{T_{pd}}{T_{pd-ST}} \quad (5)$$

where δ is the performance loss because of the extra load of the sleep transistors when the circuit block is active, and T_{pd-ST} and T_{pd} are the circuit propagation delay with and without sleep transistors, respectively. β is a factor accounting for the impact of metal wires in the DSTN ranging from 0.05 to 0.5 [11]. N is the number of charging/discharging paths (or the number of total sleep transistors) in the DSTN. L, C_{ox}, and μ are the channel length, gate capacitance, and the carrier mobility of the sleep transistor, respectively. V_{tL} is the threshold voltage of the circuit block, and V_{tH} is the threshold voltage of the sleep transistor. CACTI-P assumes that sleep transistors are all built with low standby power (LSTP) devices [14] with very high V_{tH} to reduce the leakage power caused by sleep transistors when circuit blocks are in the sleep state. $MSSC$ is the maximum simultaneous switching current of the circuit block in the active state. CACTI-P calculates $MSSC$ for different circuit types using Equations 6, 7, and 8.

$$MSSC(array) = n \times 2I_{sat,cell} \quad (6)$$

$$MSSC(WLdirver) = I_{sat,WLdriver} \quad (7)$$

$$MSSC(BLdriver) = n \times I_{sat,BLdriver} \quad (8)$$

where n is the number of switching bits in each access to the subarray, and $n \leq N_c$, where N_c is the number of columns in a subarray. $I_{sat,cell}$ is the saturation current of SRAM cell transistors, and $2I_{sat,cell}$ indicates that both inverters in the SRAM cell contribute to the total switching current. The DSTN can significantly reduce the total sleep transistor area (width) for a given performance loss. For example, the sleep transistors in wordline drivers are connected through the local network. During a normal operation, there is only one active wordline in a subarray at any time, with the $MSSC$ as shown in Equation 7. All connected transistors in the network act as the charging path of the active wordline, which significantly reduces the size of a sleep transistor by N times (N is the number of rows inside a subarray in this case), compared to the traditional module-based [7] and cluster-based [3] sleep transistor designs, while still satisfying the design constraint of allowed performance loss.

Once CACTI-P finishes sizing the sleep transistors, it computes the wakeup latency T_{wakeup} and energy E_{wakeup} of the circuit blocks with power-gating. Figure 5 shows the equivalent RC circuit during the wakeup time of cell array. When the subarray enters/exits power-gating states, all paths in the DSTN charge/discharge simultaneously. Thus, there is no charge sharing between SRAM columns, and only one column in the subarray needs to be considered when computing the wakeup latency. The wakeup time is the intrinsic delay of the RC circuit, computed using the Elmore-delay model as in Equation 9. The wakeup energy of a single column is computed using Equation 10, where C_{ckt} and C_{st} are the total capacitances for the charge/discharge path of the subarray circuit and sleep transistors respectively as shown in Figure 5. ΔV is the voltage swing from V_{ccmin} to V_{cc}. Note that the wakeup energy is the total energy for a complete sleep and wakeup cycle.

$$\tau_{wakeup} = \sum_{i=1}^{N} C_i \cdot \sum_{j=1}^{i} R_j \quad (9)$$

$$E_{wakeup} = (C_{ckt}+C_{st})\Delta V V_{cc} \quad (10)$$

The overall wakeup latency, energy, and area overhead, considering all power-gating circuitry, are described in Equations 11, 12, and 13, respectively. While the wakeup latency takes the maximum of the four power-gated circuits, the wakeup energy is the sum of the total energy consumed during the wakeup process. The area overhead is

Technology (nm)	Subthreshold leakage reduction
65	267%
45	274%
32	283%
22	304%

TABLE II
SUBTHRESHOLD LEAKAGE CURRENT DENSITY REDUCTION (IN PERCENTAGE) BY USING LONG CHANNEL DEVICES FROM 65NM TO 22NM TECHNOLOGY NODES.

the sum of all sleep transistors. Unlike power-gating, bitline floating does not need sleep transistors and thus causes no area overhead.

$$T_{wakeup} = Max(\tau_{wakeup,array}, \tau_{wakeup,WLdriver}, \tau_{wakeup,BLdriver}, \tau_{wakeup,BLfloating}) \quad (11)$$

$$E_{wakeup} = \sum(E_{wakeup,array}, E_{wakeup,WLdriver}, E_{wakeup,BLdriver}, E_{wakeup,BLfloating}) \quad (12)$$

$$A_{wakeup} = \sum(A_{wakeup,array}, A_{wakeup,WLdriver}, A_{wakeup,BLdriver}) \quad (13)$$

After computing overall wakeup latency, CACTI-P will check whether the wakeup latency is over the limit of the allowed wakeup penalty set by the user. If the constraint is violated, CACTI-P will increase the size of the sleep transistors (or the precharge transistors in the bitlines) in the power-gating circuit with the longest wakeup latency and re-calculate the overall wakeup latency. CACTI-P repeats this process until the wakeup latency constraint is met.

B. Long channel and Hi-k metal gate devices

CACTI-P models both long channel and Hi-k devices as technology-level techniques for leakage power reduction. Hi-k device parameters are obtained from Intel's data [5] and MASTAR [14]. CACTI-P assumes all devices are Hi-k devices after 45nm, which reduces the gate leakage by more than 10×, and leads to a total gate leakage of less than 5% of the total subthreshold leakage.

The long channel devices have a 10% longer channel than normal devices at the same technology node with approximately a 3× subthreshold leakage reduction. To accurately model the system implications of long channel devices, we use MASTAR [14] to obtain the exact leakage current reduction values for technology nodes from 65nm to 22nm as shown in Table II. CACTI-P first computes the best access time of an SRAM structure with normal devices on the critical path and long channel devices on non-critical paths. Then during the optimization process, CACTI-P replaces normal devices on the critical path with the long channel devices, starting from the circuit section consuming the smallest portion of the total access latency. After each replacement, CACTI-P checks whether the current total access time is still within the allowed deviation from the best access time specified by the user. This optimization and replacement will stop when the access time penalty is out of the range of the allowed deviation from the best access time, and the results before the violation are considered as the final results. The tuning process for both power-gating and device optimization only incurs less than 10% of computation time of the previous CACTI on modern servers.

IV. VALIDATION

The primary focus of CACTI-P is the accurate modeling of power, area, and timing of SRAM structures with leakage power reduction techniques. Modeled overheads include the wakeup latency/energy and area penalty. We compare the output of CACTI-P against the published data of modern industrial cache designs including a 16MB 16-way associative L3 cache of the 65nm Xeon Tulsa processor [13], a 6MB 24-way associative L2 cache of the 45nm Penryn processor [4], and a 128Kb subarray of Intel's 32nm SRAM design [16]. The configurations for the validations are based on the published data of the target RAMs and caches in [4], [13], [16] including the leakage power reduction techniques, cache capacity, associativity, technology, and bus width. We assume that the allowed power-gating induced performance loss is 5%. The validation targets include different SRAM arrays under technology generations from 65nm to 32nm. Thus, the validation stresses CACTI-P in a comprehensive and detailed way as well as tests its ability to accurately cover multiple technology generations.

Table III shows a comparison of the leakage power for validations against the target SRAM cache designs. Differences between the total leakage power generated by CACTI-P and reported data are 5%, 8%, and 14% for the Xeon L3 cache, the Penryn L2 cache, and the Intel 32nm SRAM subarray, respectively. CACTI-P significantly improves the accuracy of leakage power modeling by 5.2× on average compared to the previous versions of CACTI [1] and brings the error percentage of leakage modeling to the common error range (10 ∼ 20%) of the previous versions of CACTI [1] for dynamic power, area, and timing.

Overhead	32nm SRAM	Model result	Error
$Time_{wakeup}$	0.16 ns	0.14 ns	-12.5%
$Area_{overhead}$	< 2%	1.7%	<-15%

TABLE IV
VALIDATION OF POWER-GATING OVERHEAD RESULTS OF CACTI-P AGAINST AN INDUSTRIAL SRAM DESIGN [16] WITH POWER-GATING.

The major contribution of CACTI-P is not only to perform accurate evaluation of power reduction with leakage power control techniques but also to accurately model the overhead such as the wakeup latency, wakeup energy, and area penalty of sleep transistors when power-gating is applied. As shown in Table IV, the results of the modeled wakeup latency and area overhead (the wakeup energy has not been reported thus cannot be compared) are well aligned with reported data in the SRAM design in [16]. Again, note that the error percentage (<-15%) is within the common error range (10 ∼ 20%) of the previous

Validation Targets	Published $P_{leakage}$	CACTI-P result	CACTI-P Error	CACTI [1] result	CACTI [1] Error
Xeon 16MB 65nm L3	6.6 W	6.92 W	5%	27.2 W	412%
Penryn 6MB 45nm L2	1.7 W	1.84 W	8%	9.65 W	568%
Intel 32nm 128 Kb subarray	5 mW	5.7 mW	14%	36.5 mW	630%

TABLE III
VALIDATION OF LEAKAGE POWER REDUCTION MODELING RESULTS OF CACTI-P AGAINST INDUSTRIAL SRAM DESIGNS WITH POWER MANAGEMENT TECHNIQUES. THE PREVIOUS CACTI [1] RESULTS ARE LISTED AS A REFERENCE TO SHOW THE IMPROVEMENT OF MODELING ACCURACY.

Validation Targets	Published $P_{dynamic}$/Area/Latency	CACTI-P Results	Error
Xeon 16MB 65nm L3	5.4 W / 200 mm^2 / 9 ns	4.75 W / 171 mm^2 / 8.3 ns	-12% /-15% /-8%
Penryn 6MB 45nm L2	1.8 W / 40 mm^2 / 4.9 ns	1.55 W / 34 mm^2 / 4.1 ns	-14% /-16% /-16%

TABLE V
VALIDATION OF DYNAMIC POWER, AREA AND TIMING RESULTS OF CACTI-P AGAINST INDUSTRIAL LAST-LEVEL CACHE DESIGNS.

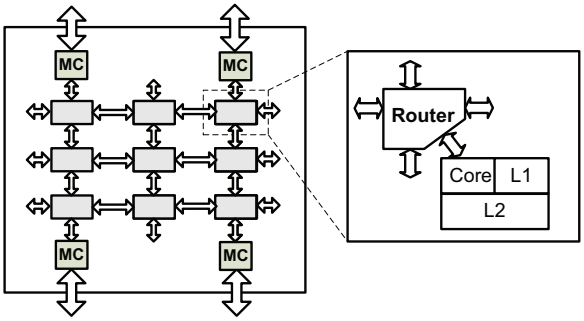


Fig. 6. Manycore system architecture. MCs refer to memory controllers.

versions of CACTI [1] for dynamic power, area, and timing.

CACTI-P also outputs the contributions of various power management techniques to the total leakage power reduction. The use of long channel devices reduces the leakage power by 2.67× and 2.74× for the 65nm Xeon L3 and 45nm Penry L2 caches, respectively. The use of power gating reduces leakage power by 1.52× and 1.97× for the 65nm Xeon L3 and 45nm Penry L2 caches, respectively. These results are in close agreement with the reported data [4], [13], [16]. CACTI-P also accurately provides results of power, area, and timing on overall SRAM arrays. As shown in Table V, the modeled dynamic power, area, and latency numbers track the published numbers well, with an average error around 15% over access time, area, and dynamic power.

The validation of CACTI-P against the industry designs shows that CACTI-P achieves accurate modeling results on leakage power reduction techniques not only across different designs but also across technology generations.

V. CASE STUDY

To demonstrate the utility of CACTI-P, we apply it to the study of the system-level impact of nanosecond scale power-gating caches in a future manycore architecture at 22 nm. We assume throughput-oriented cores with multi-level caches and apply the nanosecond scale power-gating technique to each level of caches. Nanosecond scale power-gating seeks to minimize the leakage power by maximizing the time that the caches stay in the sleep state while the processor is still running. Nanosecond scale power-gating is achieved by putting the caches in the sleep state after each batch of cache accesses. Each cache access only involves one or a few subarrays that are usually a small portion of the entire cache. For nanosecond scale power-gating, all subarrays are in the sleep state when the cache is idle. When the cache is active, the active subarrays enter the sleep state after each cache access unless there are accesses to the same subarrays issued into the cache pipeline. While nanosecond scale power-gating can maximize leakage power reduction, the associated timing and energy penalty may cause the degradation of system performance and waste dynamic energy. In this study we find how the multi-level cache hierarchy can make the best use of nanosecond scale power-gating.

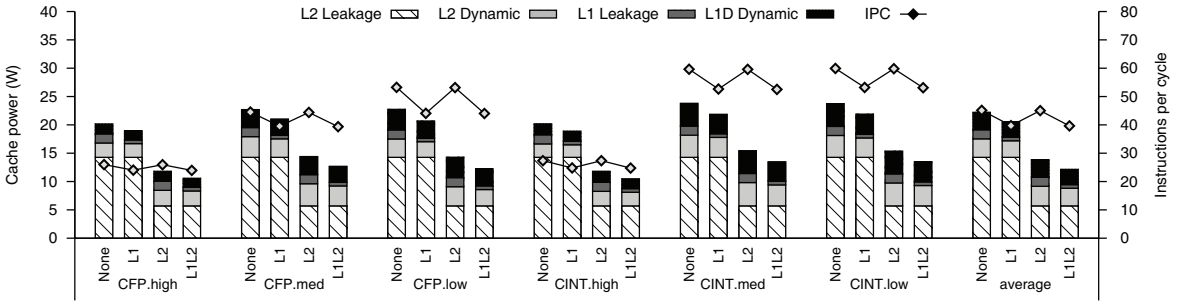
A. Experimental Setup

Figure 6 shows the future manycore architecture we assume targeting high throughput computing. It consists of 64 cores running at 3.5GHz connected by a 2D-mesh on-chip network. Up to 4 threads share an in-order core pipeline like a Niagara processor [8]. Each core has 32KB 4-way set-associative L1 instruction and data caches as well as a 512KB 16-way set-associative L2 cache. All caches have 64B cache lines. An L2 cache is inclusive of L1 caches in the same core. A directory-based MESI protocol is used for cache coherency. The on-chip network employs minimal dimension-order routing and two virtual channels per physical port. There are two memory controllers per edge so that a total of 8 controllers are used, with each memory controller having a single channel of DDR3-1600 DIMM. We use McSim [2], a manycore simulation infrastructure, for performance simulation. By using McPAT [9], [10], the total chip thermal design power (TDP) and area are determined to be 136W and 263mm^2 respectively, given the 3.5GHz frequency. We apply long channel devices and Hi-k devices to both L1 and L2 caches. The L1 instruction (L1I) cache and L1 data (L1D) cache have the same structures, and nanosecond scale power-gating can be applied to both. However, we only apply nanosecond scale power-gating to L1D cache since the L1I cache is seldom idle during processor execution and thus not a good candidate for nanosecond scale power-gating. Thus, we only study the system implications of power-gating on both L1D and L2 caches. Table VI lists the results used in our simulations obtained from CACTI-P on the native access latency/energy, the wakeup latency/energy because of power-gating sleep transistors, and the leakage power with and without power-gating. Note that although the wakeup latency of the L1D cache is much less than one core clock cycle, the cache access time still increases by one cycle since the sleep transistors are controlled by the clock signal and need to be restored to the full supply voltage level before wordline activation starts.

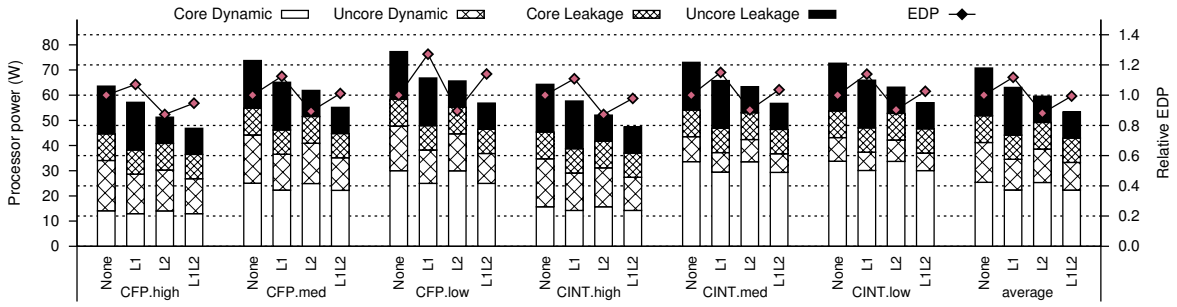
The SPEC CPU 2006 [6] benchmark suite is used in our experiments. Twelve applications are selected from the integer (CINT) and floating-point (CFP) categories and divided into 3 groups for each category, based on their main-memory bandwidth demand (Table VII). The benchmark suite is simulated as consolidated multi-

	Access time (ns)/energy (nJ)	Wakeup time (ns)/energy (nJ)	Leakage power (mW) NP/WP
32KB L1D	0.45 (2 cycles) / 0.07	0.05 (1 cycle) / 0.013	25 / 11
512KB L2	2.3 (8 cycles) / 0.6	0.19 (1 cycle) / 0.18	223 / 89

TABLE VI
CACTI-P MODELING RESULTS OF THE 32KB L1 CACHE AND 512KB L2 CACHE IN THE MANCYCORE PROCESSOR AT 22NM TECHNOLOGY NODE. LEAKAGE POWER INCLUDES BOTH THE RESULTS WITH (WP) AND WITHOUT POWER GATING (NP)



(a) Cache power and IPC (Instruction Per Cycle). Higher is better for IPC.



(b) Processor power and EDP (Energy-Delay Product). Lower is better for power and EDP.

Fig. 7. Cache power, system power, IPC, and relative EDP of the tested manycore system running the SPEC CPU 2006 benchmark suite. None stands for a configuration without power-gating on any cache, which is the baseline. L1, L2, and L1L2 stand for configurations where a power-gating technique is applied to only L1D caches, only L2 caches, and both L1D and L2 caches.

SPEC CPU2006

Set	Applications
CINT	
high	429.mcf, 462.libquantum, 471.omnetpp, 473.astar
med	403.gcc, 445.gobmk, 464.h264ref, 483.xalancbmk
low	400.perlbench, 401.bzip2, 456.hmmer, 458.sjeng
CFP	
high	433.milc, 450.soplex, 459.GemsFDTD, 470.lbm
med	410.bwaves, 434.zeusmp, 437.leslie3d, 481.wrf
low	436.cactusADM, 447.dealII, 454.calculix, 482.sphinx3

TABLE VII
SPEC 2006 APPLICATION MIXES FOR HIGH, MEDIUM, AND LOW MEMORY BANDWIDTH.

programmed workloads, with each hardware thread running one copy of a benchmark. We simulate two billion instructions for each run.

B. Results

Figure 7 shows cache power, processor power, IPC, and relative EDP of the manycore system on the SPEC CPU 2006 benchmark suite. Figure 7(a) shows the dynamic and leakage power of L1D caches and L2 caches as well as the IPC of the 6 application mixes. When no power-gating technique is applied to the L2 caches, L2 leakage power surpasses other cache power components owing to the large capacity of the L2 caches. Power-gating the L2 caches reduces more than half of their leakage power, but makes negligible impacts on the system performance and dynamic energy. As a result, it improves the system-level EDP by 18% on average as shown in Figure 7(b).

On the contrary, power-gating the L1D caches lowers the L1D leakage power by 2.2×, which is significant, but translates into only a 7% reduction in total cache power since the size of the L1D caches are relatively smaller than the L2 caches. Because L1D caches are directly connected to the core pipelines and more latency critical, nanosecond scale power-gating on L1D caches affects the performance more significantly than power-gating the L2 caches. Moreover, because of the frequent accesses to L1D cache, the wakeup energy also adds an extra 18% to its dynamic power. On average, it lowers the IPC by 11%, which combined with the extra dynamic power leads to a 15% increase in the system-level EDP. The influences of L1D nanosecond scale power-gating and L2 nanosecond scale power-gating on leakage power reduction, IPC, and system EDP are mostly superposed, which means that the impacts of power-gating on each level are rarely dependent. The system configuration that applies nanosecond scale power-gating only to L2 caches gives the best EDP among all the options.

VI. CONCLUSION

Managing SRAM leakage power has become one of the most critical design constraints to meet the chip power budget. CACTI-P is the first architecture-level modeling tool to consider all major leakage power reduction techniques for SRAM-based structures. CACTI-P models both the leakage power savings as well as the power, area, and timing overheads when implementing advanced leakage management techniques. By providing these capabilities, CACTI-P bridges the gap between circuit and device technologies on the one hand and system organizations on the other, enabling architects to perform quantitative research on a broad design space for leakage management of SRAM-based structures for future processors. Using CACTI-P with a performance simulator, we show that nanosecond scale power-gating applied to L2 caches can achieve a good balance between leakage power savings and the degradation on performance and wasting of dynamic energy, which improves the system energy-delay product by 18% on average for multiprogrammed workloads composed of the SPEC CPU 2006 benchmark suite.

REFERENCES

[1] "CACTI 6.5," http://www.hpl.hp.com/research/cacti.

[2] "McSim: a Manycore Simulation Infrastructure," http://scale.snu.ac.kr/mcsim.

[3] M. Anis, S. Areibi, M. Mahmoud, and M. Elmasry, "Dynamic and Leakage Power Reduction in MTCMOS Circuits Using an Automated Efficient Gate Clustering Technique," 2002, pp. 480–485.

[4] V. George, *et al.*, "Penryn: 45-nm Next Generation Intel Core 2 Processor," in *ASSCC'07 IEEE Asian Solid-State Circuits Conference*, 2007.

[5] F. Hamzaoglu, *et al.*, "A 3.8 GHz 153 Mb SRAM Design With Dynamic Stability Enhancement and Leakage Reduction in 45 nm High-k Metal Gate CMOS Technology," *IEEE Journal of Solid-State Circuits*, Jan 2009.

[6] J. L. Henning, "Performance Counters and Development of SPEC CPU2006," *Computer Architecture News*, vol. 35, no. 1, 2007.

[7] J. Kao, S. Narendra, and A. Chandrakasan, "MTCMOS Hierarchical Sizing Based on Mutual Exclusive Discharge Patterns," in *in DAC*, 1998, pp. 495–500.

[8] P. Kongetira, K. Aingaran, and K. Olukotun, "Niagara: A 32-Way Multithreaded Sparc Processor," *IEEE Micro*, vol. 25, no. 2, 2005.

[9] S. Li, J. Ahn, J. B. Brockman, and N. P. Jouppi, "McPAT 1.0: An Integrated Power, Area, and Timing Modeling Framework for Multicore Architectures," HP Labs, Tech. Rep. HPL-2009-206, 2009.

[10] S. Li, *et al.*, "McPAT: An Integrated Power, Area, and Timing Modeling Framework for Multicore and Manycore Architectures," in *MICRO 42: Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, 2009, pp. 469–480.

[11] C. Long and L. He, "Distributed Sleep Transistor Network for Power Reduction," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 12, pp. 937–946, September 2004.

[12] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003.

[13] S. Rusu, S. Tam, H. Muljono, D. Ayers, and J. Chang, "A Dual-Core Multi-Threaded Xeon Processor with 16MB L3 Cache," in *ISSCC*, 2006.

[14] Semiconductor Industries Association, "International Technology Roadmap for Semiconductors (ITRS) / Model for Assessment of CMOS Technologies and Roadmaps (MASTAR) http://www.itrs.net/."

[15] S. Thoziyoor, J. Ahn, M. Monchiero, J. Brockman, and N. Jouppi, "A Comprehensive Memory Modeling Tool and its Application to the Design and Analysis of Future Memory Hierarchies," in *ISCA*, 2008.

[16] Y. Wang, *et al.*, "A 4.0 GHz 291 Mb Voltage-Scalable SRAM Design in a 32 nm High-k + Metal-Gate CMOS Technology With Integrated Power Management," *IEEE Journal of Solid-state Circuits*, vol. 45, pp. 103–110, 2010.

[17] N. Weste and D. Harris, *CMOS VLSI Design: A Circuits and Systems Perspective. 3rd Edition*, 2004.

[18] S. Wilton and N. P. Jouppi, "An Enhanced Access and Cycle Time Model for On-Chip Caches," DEC WRL, Tech. Rep. technical report number 93/5, 1994.

[19] X. Xi, K. M. Cao, H. Wan, M. Chan, and C. Hu, "BSIM4.2.1 MOSFET Model," Department of Electrical Engineering and Computer Sciences University of California, Berkeley, Tech. Rep., 2001.

[20] K. Zhang, *et al.*, "SRAM Design on 65-nm CMOS Technology with Dynamic Sleep Transistor for Leakage Reduction," *JSSC*, vol. 40, no. 4, pp. 895–901, 2005.