Appears in the Proceedings of the 47th Int'l Symp. on Microarchitecture (MICRO-47), December 2014

Enabling Realistic Fine-Grain Voltage Scaling with Reconfigurable Power Distribution Networks

Waclaw Godycki, Christopher Torng, Ivan Bukreyev, Alyssa Apsel, and Christopher Batten

School of Electrical and Computer Engineering, Cornell University, Ithaca, NY
{wg63,clt67,ib264,aba25,cbatten}@cornell.edu

***Abstract*—Recent work has shown that monolithic integration of voltage regulators will be feasible in the near future, enabling reduced system cost and the potential for fine-grain voltage scaling (FGVS). More specifically, on-chip switched-capacitor regulators appear to offer an attractive trade-off in terms of integration complexity, power density, power efficiency, and response time. In this paper, we use architecture-level modeling to explore a new dynamic voltage/frequency scaling controller called the fine-grain synchronization controller (FG-SYNC+). FG-SYNC+ enables improved performance and energy efficiency at similar average power for multithreaded applications with activity imbalance. We then use circuit-level modeling to explore various approaches to organizing on-chip voltage regulation, including a new approach called reconfigurable power distribution networks (RPDNs). RPDNs allow one regulator to "borrow" energy storage from regulators associated with underutilized cores resulting in improved area/power efficiency and faster response times. We evaluate FG-SYNC+ and RPDN using a vertically integrated research methodology, and our results demonstrate a 10–50% performance and 10-70% energy-efficiency improvement on the majority of the applications studied compared to no FGVS, yet RPDN uses 40% less area compared to a more traditional per-core regulation scheme.**

I. Introduction

Monolithic integration using a standard CMOS process provides a tremendous cost incentive for including more and more functionality on a single die. This system-on-chip (SoC) integration enables both low-power embedded platforms and high-performance processors to include a diverse array of components such as processing engines, accelerators, embedded flash memories, and external peripheral interfaces. Almost every computing system requires closed-loop voltage regulators that, at first glance, seem like another likely target for monolithic integration. These regulators convert the noisy voltage levels available from the system's environment into the multiple fixed or adjustable voltage levels required by the system, and they are usually based on efficient switch-mode circuits. These regulators have traditionally been implemented off-chip for two key reasons: (1) limited availability of high-speed switching with suitable parasitic losses; and (2) limited availability of integrated energy-storage elements with suitable energy densities. The economic pressure towards monolithic integration has simply not outweighed the potential reduction in efficiency.

Recent technology trends suggest that we are entering a new era where it is now becoming feasible to reduce system cost by integrating switching regulators on-chip. High-speed switching efficiencies have increased with technology scaling, reducing the need for very high-density inductors and capacitors. This trend is evident in industry, especially in Intel's recent Haswell microprocessors which use in-package inductors with on-chip regulators to provide fast-changing supply voltages for different chip modules [24,33]. At the same time, materials improvements such as integrated in-package magnetic materials (e.g., Ni-Fe [43]) and new integrated on-chip capacitor organizations (e.g., deep-trench capacitors [3, 9]) have improved the density of the energy storage elements that are available. The future of on-chip voltage regulation offers interesting opportunities and significant challenges, and this has sparked interest from the circuit research community [2, 18, 20, 21, 25, 27–29, 43, 45] and to a lesser degree in the architecture research community [1, 10, 26, 47–49].

In addition to reduced system cost, one of the key benefits of on-chip regulation is the potential for fine-grain voltage scaling (FGVS) in level (i.e., many different voltage levels), space (i.e., per-core regulation), and time (i.e., fast transition times between levels). Dynamic voltage and frequency scaling (DVFS) is perhaps one of the most well-studied techniques for adaptively balancing performance and energy efficiency. DVFS has been leveraged to improve energy efficiency at similar performance [5, 19, 26, 31, 44], operate at an energy-minimal or energy-optimized point [8, 16], improve performance at similar peak power [4, 14, 30, 32, 37–39], and mitigate process variation [35]. Most of these studies have assumed off-chip voltage regulation best used for coarse-grain voltage scaling. Traditional off-chip switching regulators operate at low switching frequencies due to the availability of large high-Q passives and the desire to reduce parasitic switching losses. They also have longer control latencies due to slow switching speeds and parasitics between the on-chip load and the off-chip regulator, resulting in voltage scaling response times on the order of tens to hundreds of microseconds [7, 34, 36]. On-chip switching regulators can leverage faster control loops and are tightly integrated with the on-chip load enabling voltage scaling response times on the order of hundreds of nanoseconds. Traditional off-chip switching regulators are expensive, bulky, and obviously require dedicated power pins and on-chip power distribution networks limiting the number of independent on-chip power domains; on-chip switching regulators can be located close to each core enabling per-core voltage scaling.

In this paper, we use an architecture and circuit co-design approach to explore the potential system-level benefit of FGVS enabled by integrated voltage regulation and techniques to mitigate the overhead of this regulation. Section II describes our target system: a sub-Watt eight-core embedded processor design implemented in a TSMC 65 nm process using a commercial standard-cell-based ASIC toolflow.

Section III uses architectural-level modeling to explore a new FGVS controller called the *fine-grain synchronization controller* (FG-SYNC+) that exploits the specific opportunities of fine-grain scaling in level, space, and time. Inspired by Miller et al.'s recent work on Booster SYNC [35], FG-

SYNC+ uses a thread library instrumented with hint instructions to inform the hardware about which cores are doing useful work vs. useless work (e.g., waiting for a task or waiting at a barrier). FG-SYNC+ improves upon this prior work in several ways by leveraging the ability of on-chip voltage regulation to provide multiple voltage levels and using additional hints to inform the hardware of how each core is progressing through its assigned work. Booster SYNC improves performance at the expense of increased average power (i.e., the "boost budget"). Other DVFS controllers usually improve energy efficiency at similar performance or improve performance under a conservative peak power limit fixed at design time. FG-SYNC+ has a more ambitious goal of improving performance and energy efficiency while maintaining similar average power. To do this, FG-SYNC+ exploits the fine-grain activity imbalance often found in multithreaded applications. For example, Figure 1 illustrates the activity of an eight-core system on three multithreaded applications and highlights potential opportunities for increasing the voltage of active cores and decreasing the voltage of waiting cores. Note that exploiting this fine-grain imbalance is simply out-of-reach for traditional off-chip regulators.

Section IV uses circuit-level modeling to explore the practical design of an integrated voltage regulator suitable for use by FG-SYNC+. Much of the prior work in this area explores inductor-based regulators, but we argue that carefully designed on-chip switched-capacitor (SC) regulators can potentially mitigate many of the challenges involved in on-chip regulation. We explore designs with a single-fixed voltage regulator (SFVR) and multiple adjustable voltage regulators (MAVR). Unfortunately, per-core voltage regulation can incur significant area overhead and longer responses times than one might expect. This is mostly because each MAVR regulator must be designed to efficiently support the peak power that can be consumed by the fastest operating mode. Our study of FG-SYNC+ enables us to make a key observation: MAVR is significantly over-designed, since all cores can never be in the fastest operating mode. Based on this observation, we propose a new approach called *reconfigurable power distribution networks* (RPDNs). RPDNs include many small "unit cells" shared among a subset of the cores in the design. Each unit cell contains the flyback capacitance and regulator switches required for a SC regulator; the unit cells can be flexibly reconfigured through a switch fabric and combined with per-core control circuitry to effectively create multiple SC regulators "on-demand". This reconfiguration reduces area overhead by avoiding the over-provisioning inherent in MAVR, improves response time when changing the target output voltage by leveraging the adjustable flyback capacitance in addition to the adjustable regulation frequency, and can potentially improve efficiency in leaky processes by reducing flyback capacitance at low current.

In Section V, we describe our detailed evaluation methodology based on a combination of circuit-, gate-, register-transfer-, and architectural-level modeling; in Section VI, we use this methodology to explore the system-level implication of combining FG-SYNC+ with RPDN. These results suggest a promising new approach that can facilitate fine-grain voltage scaling with low-overhead in future multicore processors. In Section VII, we discuss the impact of di/dt noise on RPDN and the implications of scaling RPDN to larger networks, higher power densities, and different technologies.

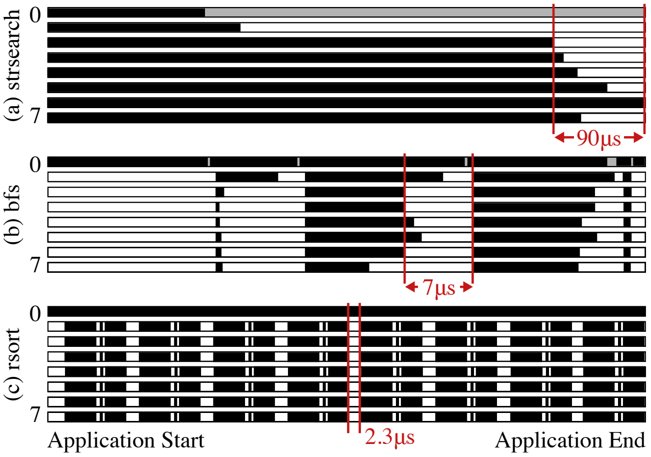


Figure 1. Activity Profile for Select Applications on Eight Cores – Variation in activity across cores produces opportunities for FGVS. Black = active; gray = waiting for join; white = waiting for work.

The contributions of this work are: (1) we propose a new controller called FG-SYNC+ that improves performance and energy efficiency at similar average power; (2) we propose a novel approach to on-chip regulator design based on the idea of reconfigurable power distribution networks; and (3) we use a vertically integrated research methodology to explore the FGVS design space.

II. Target System

Although much of our analysis is applicable to larger high-performance systems, we choose to focus on the smaller low-power systems that will likely be the first to integrate significant on-chip voltage regulation. Our target system is an embedded processor composed of: eight in-order, single-issue, five-stage, RISC cores; private, coherent 16 KB instruction and data L1 caches; and a shared 512 KB unified L2 cache.

We implemented the core and L1 memory system for this design in RTL and used a commercial standard-cell-based CAD toolflow targeting a TSMC 65 nm process to generate layout for one core. Section V describes our research methodology in greater detail. We assume the external supply voltage is 2.2 V and that FGVS should provide up to four voltage levels: 1.0 V for the nominal supply; 0.7 V for a slow, low-power execution mode (*resting mode*); 1.15 V for a fast, high-power execution mode (*sprinting mode*); and 1.33 V for an even faster execution mode (*super-sprinting mode*), all within the acceptable process operating range. Analysis of the placed-and-routed design indicates each core is approximately 0.75 mm^2 and can run at 333 MHz at 1 V. We predict that more aggressive RTL and circuit design could increase this clock frequency by 2× or more.

First-order estimates suggest the full eight-core system would be approximately 6 mm^2. When running a reasonable workload, each core/L1 consumes approximately 20 mW, and when waiting for work or a synchronization primitive, each core/L1 consumes approximately 3 mW. This implies that the power for all eight cores and L1 memory system (excluding the L2 cache) can range from 100–200 mW when doing useful work and that the peak power density of the cores and L1 memory system is approximately 25 mW/mm^2. For all designs we assume (potentially multiple) on-chip phase-locked-loops (PLLs) to enable fast frequency adjustment based on recent low power designs [12, 17]. We will use this target system to help drive the design of FG-SYNC+ and RPDN.

III. FGVS Architecture Design: FG-SYNC+

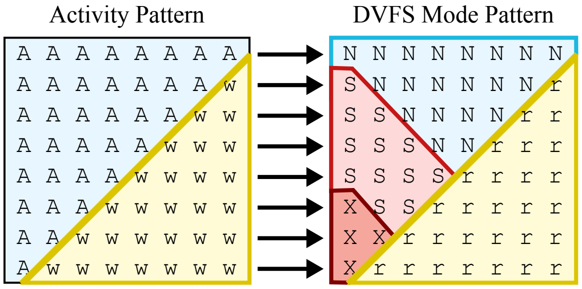
In this section, we explore a new *fine-grain synchronization controller* (FG-SYNC+) using architecture-level modeling. Section V includes more details about our evaluation methodology. After introducing the basic FG-SYNC+ controller, we use three sensitivity studies to understand the implication of varying: (1) the number of voltage levels, (2) the number of voltage domains, (3) and voltage-settling response times. Insights from this section will help motivate our design-space exploration of on-chip voltage regulation in Section IV.

A. Basic FG-SYNC+ Controller

The goal of FG-SYNC+ is to improve performance at the same average power. FG-SYNC+ rests cores that are not doing useful work, creating power slack to sprint cores that are doing useful work. Inspired by previous work on Booster SYNC [35], we instrument synchronization primitives in the threading library with hint instructions to inform the hardware which threads are doing useful work. This elegant approach avoids the need for complex prediction heuristics by exploiting application-level information to efficiently sprint the most critical cores. FG-SYNC+ extends Booster SYNC in two ways: (1) by carefully using the multiple voltage levels available with on-chip regulation and (2) by including hints indicating the progress of each thread.

The hint instructions toggle activity bits in each core. FG-SYNC+ reads these bits every sampling period and uses a lookup table to map activity patterns to DVFS modes. In the example table in Figure 2, if all cores are doing useful work, then FG-SYNC+ runs the entire system at nominal voltage and frequency (first row). As more cores are waiting, FG-SYNC+ rests the waiting cores and uses the resulting power slack to sprint or super-sprint active cores. We design lookup tables offline using our RTL-based energy model to ensure that the power of each configuration will remain below the average power of all cores running at nominal voltage (i.e., with no DVFS). Booster SYNC only provides two voltage levels since it relies on fast switching between two off-chip voltage regulators, thus Booster SYNC improves performance by increasing power consumption. FG-SYNC+'s use of multiple levels enables balancing sprinting and resting cores to improve performance at the same average power.

FG-SYNC+ includes additional "work left" hint instructions embedded in the thread library's `parallel_for` function to inform the hardware how many iterations the core has left to process. This gives FG-SYNC+ insight into the relative progress of each core in a multithreaded application. Without these additional hints, FG-SYNC+ can determine which cores are active but not which of these cores are most critical. The "work left" hint instructions enable FG-SYNC+ to sprint those cores that have the most work to do, potentially reducing the overall execution time.



Activity Pattern	DVFS Mode Pattern
A A A A A A A A	N N N N N N N N
A A A A A A A w	S N N N N N N r
A A A A A A w w	S S N N N N r r
A A A A A w w w	S S S N N r r r
A A A A w w w w	S S S S r r r r
A A A w w w w w	X S S r r r r r
A A w w w w w w	X X r r r r r r
A w w w w w w w	X r r r r r r r

Figure 2. Lookup Table Mapping Activity Pattern to DVFS Modes – FG-SYNC+ uses activity information to rest cores that are waiting, creating power slack to sprint cores that are doing useful work. A = core doing useful work; w = core waiting; r = core resting at 0.7 V; N = core in nominal mode at 1.0 V; S = core sprinting at 1.15 V; X = core super-sprinting at 1.33 V.

B. FG-SYNC+ with Fine-Grain Scaling in Level

We begin our study assuming a system with very fine-grain voltage scaling in space and time: eight voltage domains (i.e., per-core voltage regulation) and instantaneous voltage-settling response time. Then we scale the number of available voltage levels and study the impact on performance and energy efficiency. Note that with one voltage level (1.0 V), FG-SYNC+ is identical to the baseline system with no DVFS since it can neither rest nor sprint.

Supporting two voltage levels enables adding either a rest or a sprint level. Figure 3(a) compares different 2-level FG-SYNC+ controllers running a diverse set of multithreaded applications on our target system. Each controller pairs the nominal level with either the resting, sprinting, or super-sprinting level. The upper-right quadrant in these normalized energy efficiency vs. performance plots has improved performance and energy efficiency compared to the baseline. Points above the isopower line use less power than the baseline, and points below it use more power than the baseline. Figure 3(a) shows that choosing a rest level (0.7 V) improves energy efficiency with no speedup. Some applications even slow down because cores that are waiting for work in a spin-loop respond more slowly to newly-available work. Choosing a sprinting level (i.e., 1.15 V or 1.33 V, similar in spirit to Booster SYNC) increases performance but also significantly increases average power. With only two levels, we are forced to choose between performance or energy efficiency.

Supporting three voltage levels enables adding both a rest and a sprint level. Figure 3(b) compares 3-level and 4-level FG-SYNC+ controllers. FG-SYNC+ can now improve both performance and energy efficiency by resting waiting cores and sprinting active cores. Choosing either sprint (1.15 V) or super-sprint (1.33 V) as our third level improves both performance and energy efficiency, but choosing super-sprint offers greater performance while still staying under the baseline power. Supporting four voltage levels enables adding rest, sprint, and super-sprint levels. FG-SYNC+ gains the ability to super-sprint 1–2 cores or to more evenly sprint 3–7 cores

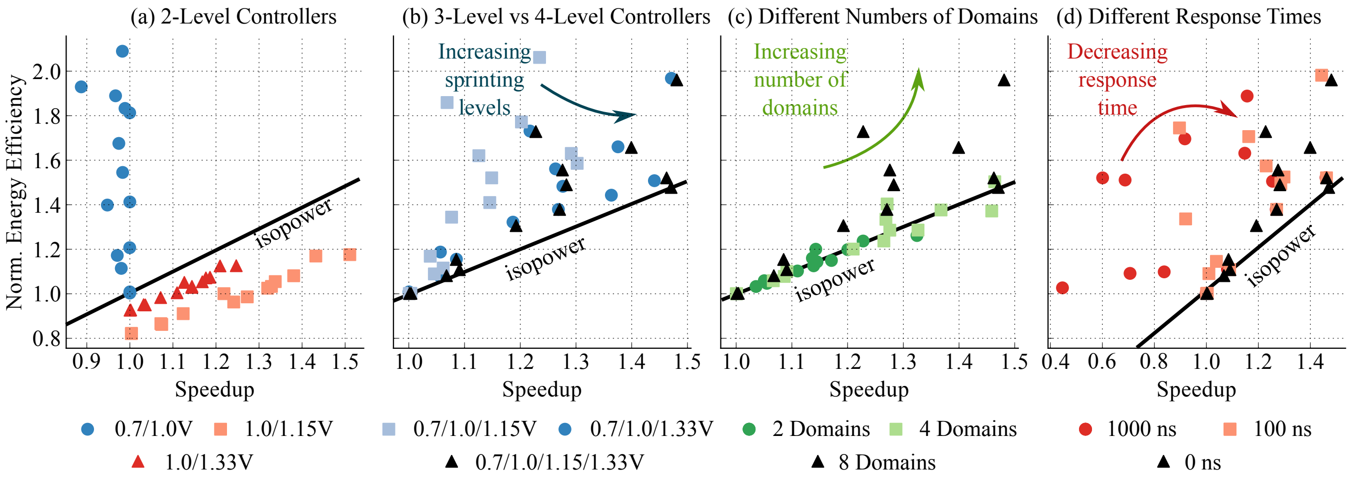


Figure 3. FGVS Exploration in Level, Space, and Time – Normalized energy efficiency and speedup over a baseline system with no DVFS. Points are applications simulated with the given controller. (a) Comparison between 2-level controllers; nominal paired with different resting and sprinting levels. (b) Comparison between different 3-level and 4-level controllers. (c) Using a 4-level controller, comparison between different numbers of voltage domains. (d) Using a 4-level controller with 8 domains, sweep response time per 0.15 V step. The black triangles in plots (b), (c), and (d) represent the same controller with very fine-grain voltage scaling in all three dimensions (i.e., 4-level, 8-domain, 0 ns response time).

(see Figure 2). In short, FG-SYNC+ can use the fourth level to better utilize power slack, further increasing performance and more closely tracking the isopower line.

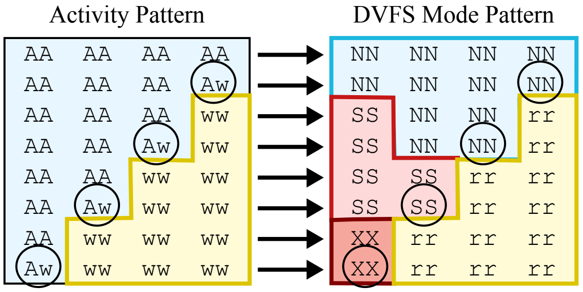
These results motivate supporting at least three levels to benefit from FGVS. For the remainder of this work, we will assume supporting four levels. Note that systems like Booster that use off-chip regulators will find it costly to support more than two levels, since this would require either more resources for additional power pins and on-chip power networks or poorer quality regulation.

C. FG-SYNC+ with Fine-Grain Scaling in Space

We now explore how FG-SYNC+ performs with fewer than eight voltage domains. With two domains, cores 0–3 and 4–7 are grouped into quads; with four domains, neighboring cores are grouped into pairs. All cores in a group must scale their voltages together. Therefore, a core waiting for work cannot rest unless all other cores in the same group are also waiting. If a core sprints, the whole group must sprint as well. In contrast, cores in the 8-domain system can independently scale voltage and frequency. Note that with one voltage domain, FG-SYNC+ cannot sprint without significantly increasing the average power over the baseline; therefore it cannot offer a performance benefit at the same average power.

Figure 3(c) compares FG-SYNC+ with 2–8 voltage domains. Each controller has four voltage levels and instantaneous voltage-settling response time. With two domains, FG-SYNC+ can improve energy efficiency by resting one quad given that all cores in the quad are waiting for work, and active cores in the other quad can be sprinted for modest performance gains. We cannot super-sprint an active quad without exceeding the average power of the baseline. FG-SYNC+ with four domains significantly improves: (1) energy efficiency by enabling more cores to rest in pairs and (2) performance by enabling a single pair to super-sprint when all other pairs are resting.

Having eight domains (i.e., per-core voltage regulation) enables FG-SYNC+ to independently optimize each core's voltage and frequency for its activity. Figure 3(c) shows that having eight domains significantly improves energy efficiency over four domains. Compare the lookup table for four domains in Figure 4 with the lookup table for eight domains in Figure 2. Notice that for the circled pairs in the 4-domain table, FG-SYNC+ must choose between running a waiting core at a non-resting level or running an active core at the resting level, sacrificing either performance or energy-efficiency. In this study, if a domain has at least one active core we choose to run the entire quad at the best voltage level for that active core. This prioritizes performance over energy efficiency for the 2- and 4-domain configuration. The 8-domain configuration does not need to make this trade-off and is able to improve both performance and energy efficiency.

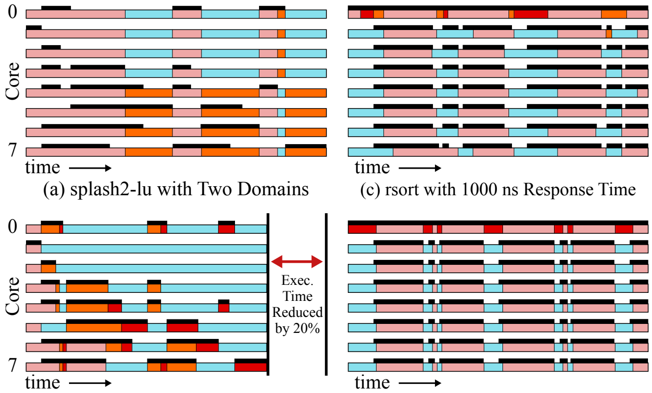


AA	AA	AA	AA
AA	AA	AA	Aw
AA	AA	AA	ww
AA	AA	Aw	ww
AA	AA	ww	ww
AA	Aw	ww	ww
AA	ww	ww	ww
Aw	ww	ww	ww

NN	NN	NN	NN
NN	NN	NN	NN
SS	NN	NN	rr
SS	NN	NN	rr
SS	SS	rr	rr
SS	SS	rr	rr
XX	rr	rr	rr
XX	rr	rr	rr

Figure 4. Example Lookup Table Mapping Activity Patterns with Four Voltage Domains to DVFS Modes – Compare and contrast with eight domains in Figure 2. In the circled pairs, FG-SYNC+ with four domains must choose between either running the waiting core at a non-resting level (energy-inefficient) or running the busy core at rest (lower performance).

Figure 5(a,b) illustrates the impact of coarser voltage domains on application performance for the *SPLASH-2 LU factorization* benchmark. Rows represent cores, black strips represent core activity, and colors represent DVFS modes. In Figure 5(a), FG-SYNC+ is heavily constrained and is forced to inefficiently run a quad at nominal or sprint, even though few cores in the quad are actually doing useful work. In Figure 5(b), per-core voltage regulation enables FG-SYNC+ to



(a) splash2-lu with Two Domains (c) rsort with 1000 ns Response Time

(b) splash2-lu with Eight Domains (d) rsort with 100 ns Response Time

Figure 5. Application Activity Plots for FG-SYNC+ – Rows show controller decisions per-core (rest = blue, nominal = pink, sprint = orange, super-sprint = deep red). Horizontal black strips above cores show when that core is active. (a,b) illustrates the impact of using multiple voltage domains over the full execution of SPLASH-2 LU factorization; (c,d) illustrates the impact of faster voltage-settling response times over a small excerpt from the execution of radix sort.

independently rest waiting cores, sprint several active cores, or even super-sprint one or two active cores, resulting in an execution time reduction of 20%.

This study motivates very fine-grain regulation in space to improve performance and energy efficiency. For the remainder of this work, we will assume per-core voltage regulation.

D. FG-SYNC+ with Fine-Grain Scaling in Time

We assume that after FG-SYNC+ makes a decision to change the voltage and frequency of a domain, it must wait until the voltage has settled before making a new decision. As in [11, 26], we assume that cores continue running even during voltage transitions. When scaling voltage up, the change in frequency must *lag* the change in voltage, and when scaling voltage down, frequency must *lead* the voltage. Both of these constraints ensure that the design always meets cycle time constraints. Together, this means that if FG-SYNC+ decides to scale the voltage up, the new frequency will not take effect immediately, and FG-SYNC+ cannot make a new decision until the new voltage has settled. Note that there is additional energy overhead during this transition as the core slowly scales voltage while staying locked at the frequency of the lower DVFS mode. For example, when scaling from (0.7 V, frequency *f1*) to (1.0 V, frequency *f2*), transition energy is paid during the time the core runs at *f1* and is transitioning voltages between 0.7 V and 1.0 V.

Figure 3(d) illustrates the impact of these overheads on performance and energy efficiency. We use a simplistic model where we linearly increase the voltage-settling response time per 0.15 V step. First note that a relatively slow 1000 ns response time increases the likelihood that FG-SYNC+ will not be able to adjust to fine-grain activity imbalance and indeed the inability to rapidly make controller decisions leads to sharp slowdowns over the baseline. A response time of 100 ns allows FG-SYNC+ to adjust quickly to fine-grain activity imbalance in our applications; this is fast enough to closely track the results for ideal (0 ns) response time.

Energy efficiency is balanced as we scale to finer-grain response times. On one hand, slow response times actually *improve* energy efficiency because cores spend more time waiting at lower (and more energy-efficient) DVFS modes until voltage settles, while still doing useful work; on the other hand, fast response times also improve energy efficiency by enabling FG-SYNC+ to quickly switch to more energy-efficient modes in response to fine-grain activity imbalance.

Figure 5(c,d) illustrates the performance overhead of slow response times more clearly, comparing a partial execution of a *radix sorting* application kernel with 1 µs and 100 ns voltage-settling response times. The performance overhead of slow response time can be seen in Figure 5(c) from the delay between the time that the core becomes active (black on the activity strip) and the time that FG-SYNC+ raises the core frequency (color change from blue to red). In Figure 5(d), the faster 100 ns response time enables FG-SYNC+ to quickly adapt to fine-grain core activity, causing the black activity strips and FG-SYNC+ decisions to "line up".

E. FG-SYNC+ Summary

There are several important insights from this study: (1) to improve both performance and energy efficiency at the same average power, at least three levels are required and four levels results in additional benefits; (2) increasing the granularity of voltage scaling in space results in increased performance and energy; (3) systems require voltage settling response times on the order of 100 ns to exploit fine-grain activity balance.

IV. FGVS Circuit Design: RPDNs

The three primary types of step-down voltage regulators are linear regulators, inductor-based switching regulators, and capacitor-based switching regulators. These regulators can be evaluated based on four key metrics: (1) *integration complexity*, i.e., does the regulator require extra non-standard fabrication steps?; (2) *area overhead and power density*, i.e., how much regulator area is required to deliver a certain amount of power?; (3) *power efficiency*, i.e., ratio of the output power to the supplied input power; and (4) *response time*, i.e., how fast can the target output voltage be adjusted?

Linear voltage regulators (also called linear dropout (LDO) regulators) are an example of a non-switching regulator. LDOs use a power MOSFET as a variable resistor, with a high-gain amplifier wrapped in a feedback configuration to reduce output resistance. At first glance, the lack of energy storage elements seems to imply LDOs will have much lower area overheads. However, a large decoupling capacitor is still required because the feedback loop in LDOs has limited bandwidth. As such, 10–15% of the chip area must be reserved for decoupling capacitance to maintain supply integrity for processor cores and logic during large current steps [20]. In addition, the maximum achievable power efficiency is the ratio of the output/input voltages since the LDO effectively acts as an adjustable resistance. This means that LDOs are highly inefficient for large voltage drops.

Inductor-based switching voltage regulators (also called Buck converters) are the traditional off-chip regulators of choice due to the potential for high power efficiency over wide voltage and current ranges; they also have excellent voltage regulation capabilities. However, in a fully on-chip buck converter, the efficiency is severely limited by the size and parasitics of the inductor. Reduction of these parasitics is the key to an efficient buck converter as shown in recently published work [2,21,25,27]. These designs have reasonable efficiencies only for relatively low step-down ratios, which makes them less suitable for the wide dynamic range required for FGVS. These regulators also provide relatively low power densities on the order of 0.2 W/mm^2. Unfortunately, solutions with higher power densities require magnetic materials, complicated post-fabrication steps, or interposer chips [43,46].

Capacitor-based switching voltage regulators (also called switched-capacitor (SC) regulators) work by alternately switching a set of capacitors with a given divide ratio from series (charge up) to shunt configuration (discharge). This switching must be fast enough to maintain the output voltage across a load. SC regulators are capable of excellent efficiencies, however, they can only support certain discrete voltage divide ratios (e.g., 3:1, 2:1, 3:2) and usually require more than eight phases to reduce ripple losses [41]. The regulation and output voltage range shortcomings of SC converters are balanced by their potential for higher power densities of 0.8–2 W/mm^2 [9, 29, 42] when using energy-dense on-chip capacitors. Note that in contrast to Buck converters, the energy density of MOS, MIM, and deep trench capacitors is sufficient to avoid the need for any off-chip or in-package energy storage elements. Due to the nature of operation, half of the capacitance in a SC regulator is always seen between the regulator output and ground thereby acting as an effective decoupling capacitance [29]. This means that an explicit decoupling cap may not be necessary, which can further reduce the area overhead of SC regulators. Unlike buck converters which are fundamentally impossible to scale for smaller loads without incurring prohibitive losses, SC regulators can be scaled by simply adjusting the size of the capacitor and the switches. Consequently, SC regulators can be easily subdivided into modules that can be added together in parallel based on demand. All of the above reasons motivate our interest in exploring on-chip SC regulators for FGVS.

Based on the results from Section III, a 4-level, 8-domain FG-SYNC+ configuration provided the best performance and energy efficiency. In the remainder of this section, we will consider three different integrated voltage regulator designs suitable for use with the target system described in Section II: (1) a baseline design which uses a single integrated fixed-voltage regulator (SFVR); (2) multiple adjustable voltage regulators (MAVR) with one regulator per core; and (3) a new approach based on a reconfigurable power distribution network (RPDN).

A. SFVR: Single Fixed-Voltage Regulator

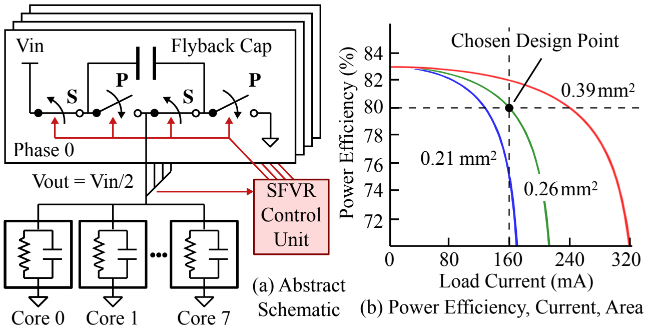
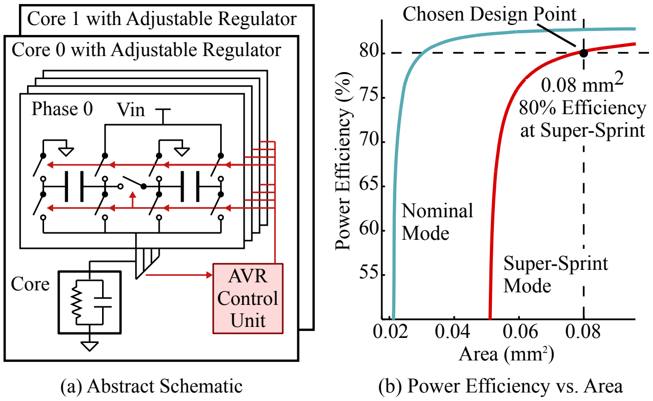


Figure 6. SFVR – (a) 2:1 topology converts Vin to Vin/2 for eight cores; S = switches closed during serial mode; P = switches closed during parallel mode; control unit monitors Vout to regulate the switching frequency; 16 phases are included to reduce ripple (only four phases shown for simplicity). (b) for a fixed voltage, power efficiency varies as a function of output current and flyback capacitance area.

A single fixed-voltage regulator (SFVR) provides a good baseline to compare against more sophisticated regulation schemes. Figure 6(a) illustrates a basic 2:1 switched-capacitor design. In *series mode*, the flyback capacitor is connected in series with the load (cores), and the input voltage source charges up the flyback capacitor. In *parallel mode*, the flyback capacitor is connected in parallel with the load, and the input voltage source is disconnected. In parallel mode, the flyback capacitor acts as an energy source that is discharged to supply power to the cores. As the converter switches between the series and parallel modes, the output voltage will gradually converge to half the input voltage. Faster switching frequencies reduce voltage ripple but decrease efficiency due to switching losses. The switching frequency is also used for fine-grain control of the output voltage. An SFVR control unit monitors the output voltage and adjusts the switching frequency in order to keep the output voltage constant across load current variations. Realistic SC regulators almost always include support for switching multiple phases of the signal in parallel to further minimize ripple. Larger flyback capacitors require more area, but can enable slower switching frequencies and therefore higher efficiencies for a given output voltage and load current. Figure 6(b) illustrates this trade-off using an analytical circuit-level model described in more detail in Section V. For a fixed output voltage, as the regulator area increases, the curve moves to the right and broadens, indicating that (1) higher efficiencies can be achieved for the same output current and (2) higher output current can be achieved for the same efficiency. For our TSMC 65 nm process, we explored a variety of different SFVR designs and ultimately chose a configuration that can provide 80% efficiency at 1 V with an area of 0.26 mm^2 (4% of the core/L1 area). It may be possible to further reduce the area overhead by re-purposing the mandatory on-chip decoupling capacitance as flyback capacitance [29].

B. MAVR: Multiple Adjustable-Voltage Regulators

To enable fine-grain voltage scaling in space and level, we require multiple adjustable voltage regulators (MAVR). Given the voltage levels from Section III, we use the more complicated flyback capacitor topology shown in Figure 7(a). For a 2.2 V input, MAVR achieves the highest efficiency at the following discrete voltage ratios: 1.0 V@2:1 = 82.7% and 1.33 V@3:2 = 80%. Adjusting the switching frequency enables the two remaining target voltage levels: 0.7 V@2:1 = 62% and 1.15 V@3:2 = 75%. Figure 7(b) illustrates the efficiency vs. area trade-off in MAVR. A regulator that must sup-



(a) Abstract Schematic

(b) Power Efficiency vs. Area

Figure 7. MAVR – (a) one adjustable voltage regulator (AVR) per core, only two shown for simplicity; control unit can configure flyback capacitance to convert Vin to Vin/2 and 3Vin/2; other intermediate voltages are possible by adjusting the regulation frequency; 16 interleaved phases are included to reduce ripple (only four phases shown for simplicity). (b) for a fixed output voltage and current, power efficiency varies as a function of area; area for nominal is over-provisioned for the sake of high efficiency at super-sprint.

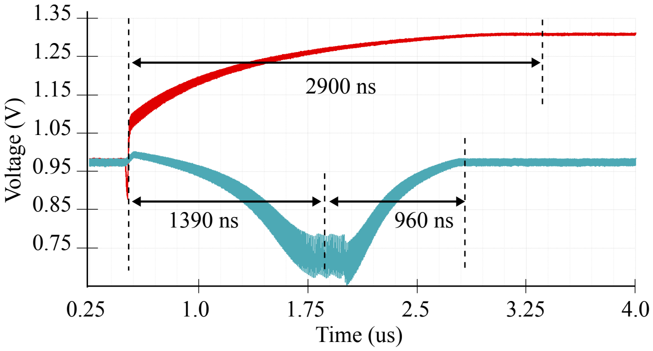


Figure 8. MAVR Transient Response – Transistor-level transient simulation of MAVR design with fixed capacitance per core.

port both nominal and super-sprinting modes requires significantly more area compared to a regulator that only supports the nominal mode. Super-sprinting is simply not possible if the regulator area is less than 0.05 mm^2 since the regulator cannot switch fast enough to provide the required output current. For our TSMC 65 nm process, we explored a variety of different MAVR designs and ultimately chose a per-core regulator area of 0.08 mm^2 which allows efficient voltage regulation from resting to super-sprint. Due to the high output power variation between core operating modes, each AVR control unit must handle significantly larger switching frequency variation than its SFVR counterpart. In order to keep the output voltage stable at low power, the AVR control unit's feedback loop must be slow enough to avoid voltage overshoots; this in turn leads to long voltage-settling response times. Figure 8 uses detailed transistor-level simulations to illustrate the response time of various operating mode transitions for a single regulator in MAVR. Section V describes the methodology used for this analysis in more detail. Most transitions take several microseconds, with the nominal to super-sprint transition taking 2.9 µs. While MAVR does enable FGVS, it does so with high area overhead and long response times.

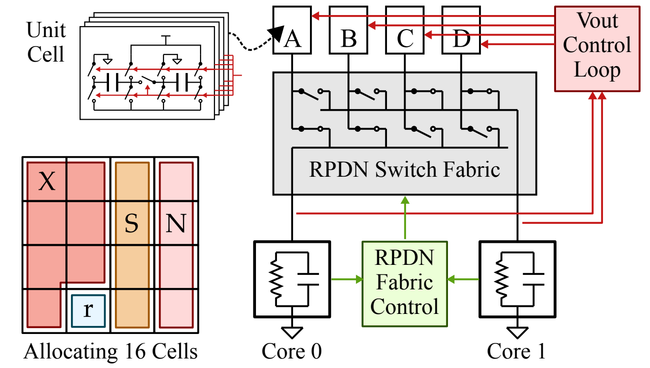


Figure 9. RPDN With Adjustable Unit Cells – All unit cells are designed for an adjustable output voltage from 0.7–1.33 V. Key RPDN blocks for two cores are shown. Diagram on the bottom left shows how unit cells are allocated across resting (r), nominal (N), sprinting (S), and super-sprinting (X) cores in the sub-RPDN.

C. RPDN: Reconfigurable Power Distribution Networks

Based on our insight from Section III, we make the key observation that MAVR is significantly over-provisioned for FG-SYNC+. Each per-core regulator in MAVR must independently support the super-sprinting mode, but only one or two cores will ever be using this mode at any given time. While it might be possible to use thread migration and a fixed assignment of cores to voltage levels [39, 47], thread migration can introduce non-trivial performance and energy overheads. We take an architecture and circuit co-design approach to design reconfigurable power distribution networks (RPDNs) that meet the needs of FG-SYNC+ while reducing area overhead. RPDN allows sprinting cores to effectively "borrow" energy storage from resting cores to avoid over-provisioning the aggregate energy storage.

Figure 9 illustrates a simple example of an RPDN for two cores. The *RPDN control unit* configures the *RPDN switch fabric* to connect *RPDN unit cells* to supply power to each of the cores. In this example, there are four unit cells and each cell is a small switched-capacitor converter capable of 2:1 and 3:2 operation. The RPDN switch fabric is a two-input, two-output crossbar. The RPDN switch fabric is initially configured such that cells A and B supply core 0 while cells C and D supply core 1. If core 0 is waiting while core 1 is active, the RPDN switch fabric can be reconfigured such that cell A supplies low power to core 0 while cells B–D supply high power to core 1. Essentially, core 1 can borrow energy storage from core 0 on-demand.

The design in Figure 9 is greatly simplified to illustrate the basic concept of RPDNs. Our actual RPDN design includes 32 unit cells with eight phases per cell and can power eight cores. Preliminary estimates show that scaling the RPDN switch fabric across all eight cores incurs significant losses. In response, we partitioned the RPDN into two isolated sub-RPDNs. Each sub-RPDN has half of the 32 unit cells to distribute to a four-core partition. Each unit cell uses a multi-level SC regulator design that enables 2:1 and 3:2 step-down conversions, similar to the regulator shown in Figure 7(a), except with only 8 phases. Based on TSMC 65 nm transistor-

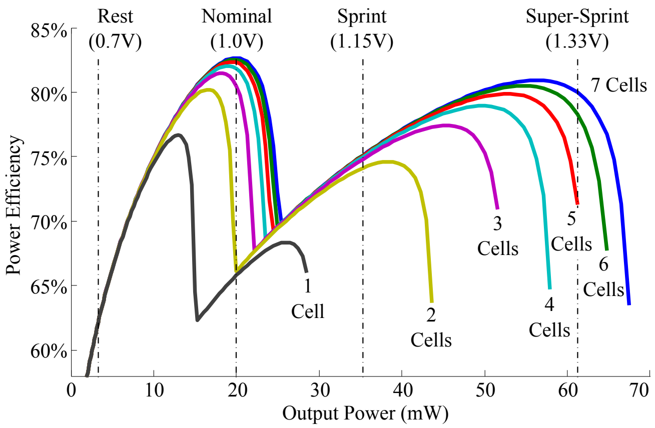


Figure 10. RPDN Power Efficiency vs. Output Power For Single Core – Each cell is 0.011 mm^2. The 7-cell RPDN curve also represents MAVR area. RPDN uses 1,4,4,7 cells for rest, nominal, sprint, and super-sprint, respectively.

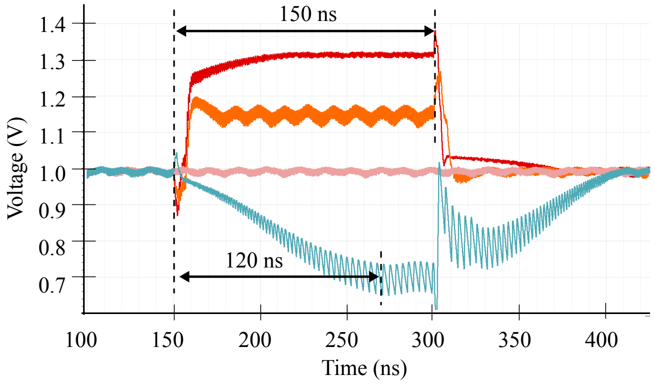


Figure 11. RPDN Transient Response – Transistor-level transient simulation of a sub-RPDN illustrating the benefit of capacitance reallocation. Four cores start at nominal, then three cores move to sprint, super-sprint, and rest and then back to nominal.

level models, the sub-RPDN switch fabric introduces a 0.25–0.75% efficiency degradation with 8% extra converter area.

Figure 10 shows the efficiency vs. output power for a single core as a function of the number of cells allocated to that core for a cell area of 0.011 mm^2. Four cells are required to efficiently support the nominal mode, while seven cells are required to efficiently support the super-sprinting mode. Since the sprinting mode uses a different flyback capacitor topology, it is able to achieve reasonable efficiency with the same number of cells as the nominal mode. Resting mode consumes very little power, so a single cell is sufficient. The inset in Figure 9 shows how unit cells of one sub-RPDN can be allocated to four cores operating in four different modes. The two cores operating in the nominal and sprinting modes are allocated four cells each. The resting core only requires a single cell, so the super-sprinting core “borrows” three cells from the resting core. MAVR must provision for the worst case, so each per-core regulator must include flyback capacitance equivalent to seven cells. RPDN provides an average of just four cells per core, and then uses reconfiguration to create seven-cell regulators for super-sprinting on-demand.

TABLE I. COMPARISON OF SFVR, MAVR, AND RPDN

	PDN Area	Power Efficiency for Vout =			Transient Response (ns)			Voltage Scaling	
	(mm^2)	0.7V	1.0V	1.33V	Min	Typ	Max	Space	Time
SFVR	0.26	n/a	80%	n/a	n/a	n/a	n/a	No	No
MAVR	0.64	62%	82.7%	80%	164	1950	3850	Yes	Yes
RPDN	0.37	62%	81.8%	80%	36	120	226	Yes	Yes

The RPDN architecture offers obvious advantages in terms of area savings. Based on our analytical model described in Section V, we compute the area overhead for SFVR, MAVR, and RPDN to be 4%, 10%, and 6% respectively. This means that RPDN provides area savings of 40% over MAVR when supporting per-core supply regulation across the same number of cores. In addition to reducing area overhead, RPDN also significantly reduces the voltage-settling response time. For resting cores, RPDN uses 15% of MAVR’s area which allows the RPDN control loop to be much faster. Furthermore, when switching between different operating modes, RPDN changes capacitance in addition to the SC divide ratio and switching frequency. This means the RPDN control loop has to make a significantly smaller switching frequency adjustment in order to accommodate the new operating mode. Figure 11 shows the transient response for one sub-RPDN where each core switches to a different operating mode. The response time for the nominal to super-sprint transition takes just 150 ns.

D. *Summary of Power Distribution Networks*

Table I summarizes the trade-offs discussed throughout this section. While on-chip voltage regulation offers the potential for fast and flexible control, it also incurs various overheads. In the case of SFVR, no flexibility is offered. MAVR provides the flexibility for fine-grain voltage scaling, but at the cost of high area overhead and long response times. Finally, RPDN offers an interesting middle ground. RPDN enables the flexibility of MAVR with significantly reduced area overhead and faster response times.

V. EVALUATION METHODOLOGY

We used a vertically integrated evaluation methodology that uses a mix of circuit-, gate-, register-transfer-, and architectural-level modeling. Circuit-level modeling is used to characterize each power distribution network (PDN); gate- and register-transfer-level modeling are used to build accurate area and energy models of our target embedded processor; and architectural-level modeling is used to analyze the system-level impact of each PDN.

A. *Circuit-Level Modeling*

We performed SPICE-level circuit simulations of all SC regulators with Cadence Spectre using models from a TSMC 65 nm process. We also used an analytical SC model to enable faster design-space exploration of regulator efficiency vs. power, supply voltage, and area as well as for estimating the transient response. Our analytical SC model, which computes switching losses, gate drive losses, bottom plate losses, and series resistive losses, is based on prior work by

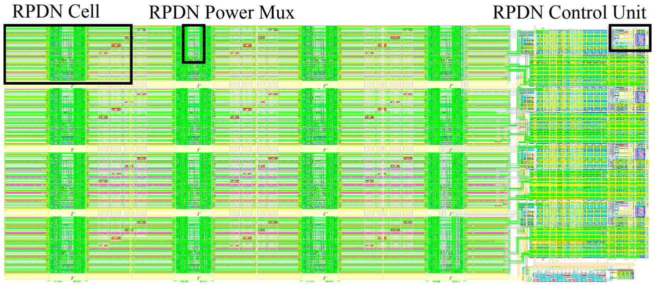


Figure 12. Tapeout-Ready Layout of a sub-RPDN Test Chip – We use the SPICE-level extracted model to validate accuracy of our analytical model for efficiency, area, and transient response.

Seeman [41]. After carefully extracting capacitor density, bottom plate capacitance, and switch parasitics from the Cadence 65 nm PDK, our analytical model and Cadence simulation results match closely in both absolute efficiency and area numbers. To properly account for a realistic RPDN switch fabric and clock distribution overheads, we performed a full-chip layout of a 16-cell sub-RPDN (see Figure 12). SPICE-level simulations were also used to help determine the relationship between voltage and frequency for our cores across different operating modes. We used nine delay stages consisting of multiple FO4 loaded inverters, NAND, and NOR gates connected in a loop, such that the total delay in the loop matches our RTL cycle time for a given voltage. We used the change in delay vs. supply voltage as a model for core voltage-frequency scaling. Finally, we augmented our SC analytical model to account for shunt losses due to capacitor leakage. We used this augmented model to study scaling towards leakier processes, as described in Section VII.

B. Gate- and RTL Modeling

To estimate power, we created an instruction-level energy model derived from a realistic RTL implementation of an in-order, single-issue scalar core and L1 memory system. The RTL model is synthesized and placed-and-routed using a combination of Synopsys DesignCompiler, IC Compiler, and PrimeTime PX with a TSMC 65 nm standard-cell library characterized at 1 V. We then ran a suite of *energy microbenchmarks* that are each designed to measure the energy of a specific instruction. For example, the `addiu` energy microbenchmark warms up the instruction cache (to isolate the energy solely due to the instruction under test) and then executes 100 `addiu` instructions in sequence. We ran this microbenchmark on: (1) the synthesized gate-level design to obtain bit-accurate traces that are fed into PrimeTime power simulations for power estimates, and (2) on the RTL simulator to obtain cycle counts for the execution. Coupled with the cycle time of the placed-and-routed design, we can calculate the energy per `addiu` instruction.

Only a subset of instructions are characterized in this way. For instance, `sll` and `srl` are similar enough that we only needed to characterize one of these instructions. Separate energy microbenchmarks are used to quantify the energy per taken-branch versus a not-taken-branch. Similarly, we tested cases for load and store hits and misses separately. In general, we see a range of 60–75 pJ per arithmetic instruction, with higher ranges for long-latency and memory instructions. We used these results to build an energy dictionary containing the energy for every instruction. The energy dictionary can be applied to an RTL or cycle-level trace containing the distribution of dynamic instruction types to produce a total energy and power estimate for the simulation. Furthermore, we can leverage the voltage-frequency relationship derived from our circuit-level modeling to scale the energy and power of the nominal configuration to other voltage-frequency pairs.

C. Cycle-Level Modeling

We use the gem5 simulator [6] in syscall emulation mode to model a multicore processor with eight single-issue in-order cores, each with private 16 KB L1 I/D caches and sharing a 1 MB L2 cache. We have extended gem5 to enable architecture and circuits co-design in several ways.

Multithreading Support in Syscall Emulation Mode – We modified gem5's address space mapping to allow cores to share memory in syscall emulation mode, and added support for a simple multi-threading library that pins threads to cores.

Software Hints – We modified gem5 to toggle an activity bit in each core after executing the new activity hint instruction. We also added support for the "work left" hint instructions to pass thread progress information from parallelized loops to the hardware.

Dynamic Frequency Scaling Support – We modified gem5's clock domains and clocked object tick calculations to support dynamic frequency scaling. Cores can independently scale their frequency, but we centralized control of all clock domains in the FG-SYNC+ controller.

Integration with RTL-Based Power Model – We modified gem5 to capture detailed, per-core instruction counts that occur not only in each DVFS mode, but also in each DVFS mode *transition* to properly account for energy overheads during these transitions. Using these statistics and our energy dictionary, we calculate energy/power for each configuration.

Integration with Circuits – We integrated voltage settling response times from circuit-level simulations for each mode transition into our gem5 frequency scaling framework. We use these response times to delay frequency change events to simulate realistic voltage scaling. We also use our circuit-level analytical SC regulator model (verified with SPICE simulations) to integrate regulator energy efficiencies into our power model to obtain realistic energy and power overheads.

D. Application Kernels and Benchmarks

We use a variety of custom application kernels as well as selected PARSEC, SPLASH-2, and PBBS benchmarks on our architectural-level model to analyze the system-level benefit of various configurations (see Table II).

bfs computes the shortest path from a given source node to every reachable node in a graph using the breadth-first-search algorithm and is parallelized across the wavefront using double buffering. *bilat* performs a bilateral image filter with a lookup table for the distance function and an optimized Taylor series expansion for calculating the intensity weight. *dither* generates a black-and-white image from a gray-scale image using Floyd-Steinberg dithering. Work is parallelized across the diagonals of the image, so that each thread works on a subset of the diagonal. A data-dependent conditional allows threads to skip work if an input pixel is white. *rsort* performs an incremental radix sort on an array of integers. During each iteration, individual threads build local histograms of the data, and then a parallel reduction is performed to determine the mapping to a global destination array. Atomic memory operations are necessary to build the global histogram structure. *kmeans* implements the k-means clustering algorithm. Assignment of objects to clusters is parallelized across objects. The minimum distance between an object and each cluster is computed independently by each thread and an atomic memory operation updates a shared data structure. Cluster centers are recomputed in parallel using one thread per cluster. *mriq* computes a calibration matrix used in magnetic resonance image reconstruction algorithms. *pbbs-dr* is a PBBS application for 2D Delaunay Mesh refinement. Work is parallelized across the bad triangles. Parallel threads move through reserve and commit phases and will only perform retriangulation if all the neighbors they marked were reserved successfully. Newly generated bad triangles are assigned to other threads. *pbbs-knn* is a PBBS application that, given an array of points in 2D, finds the nearest neighbor to each point using a quadtree to speed up neighbor lookups. Quadtree generation is parallelized at each depth of the tree so that each thread works on a separate sub-quadrant. The quadtree is used to find the nearest neighbor to each point in parallel. *pbbs-mm* is a PBBS application for maximal matching on an undirected graph. Work is parallelized across the edges in the graph. Parallel threads move through reserve and commit phases. Threads attempt to mark the endpoints of the assigned edge with the edge ID. In the commit phase, threads will only mark its edge as part of the maximal matching if both endpoints were reserved successfully. *splash2-fft* is a SPLASH-2 benchmark that performs a complex 1D version of a radix-sqrt(n) six-step FFT algorithm. Cores are assigned contiguous sets of rows in partitioned matrices. Each core transposes contiguous sub-matrices from every other core and transposes one locally. *splash2-lu* is a SPLASH-2 benchmark that performs a matrix factorization into a lower and upper triangular matrix. Parallelization is across square blocks of size B and this parameter is picked so that blocks fit in the cache. *strsearch* implements the Knuth-Morris-Pratt algorithm to search a collection of byte streams for the presence of substrings. The search is parallelized by having all threads search for the same substrings in different streams. The deterministic finite automatas used to model substring-matching state machines are also generated in parallel. *viterbi* decodes frames of convolutionally encoded data using the Viterbi algorithm. Iterative calculation of survivor paths and their accumulated error are parallelized across paths. Each thread performs an add-compare-select butterfly operation to compute the error for two paths simultaneously, which requires unpredictable accesses to a lookup table.

VI. Evaluation Results

In this section, we compare the SFVR, MAVR, and RPDN designs. We evaluate the performance, energy efficiency, and power of our applications as they run on our target system with each type of power distribution network (PDN). We choose a 4-level, 8-domain FG-SYNC+ controller based on the FGVS study in Section III. From our circuit-level study in Section IV, we account for realistic voltage-settling response times and regulator power efficiencies in each DVFS mode for varying load currents.

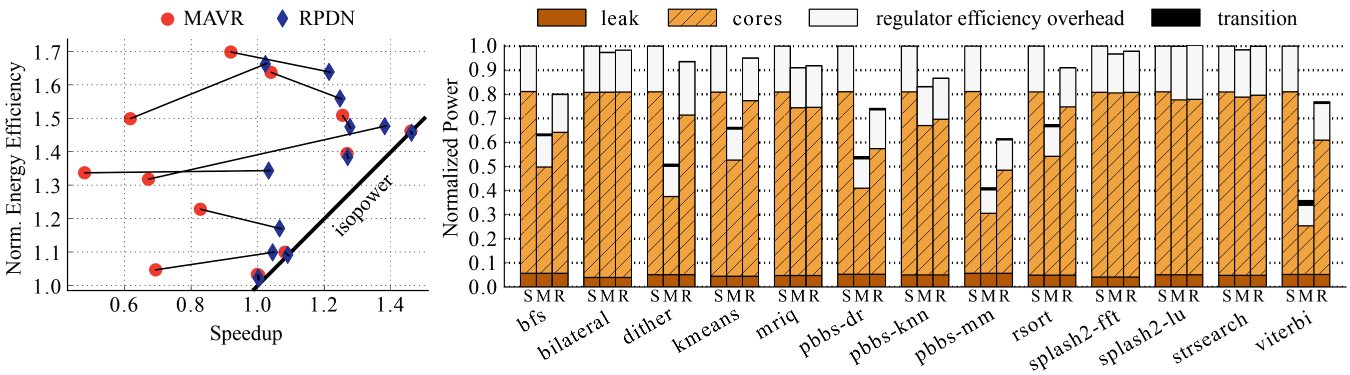
Table II. Application Performance and Energy

		Performance			**Energy**			**Trans**	
App	**DInsts**	**SFVR (µs)**	**MAVR**	**RPDN**	**SFVR (µJ)**	**MAVR**	**RPDN**	**MAVR**	**RPDN**
bfs	58	101	1.04	1.25	17	0.61	0.64	350	693
bilateral	6540	3254	1.00	1.00	767	0.97	0.98	1	3
dither	2762	4121	0.67	1.38	749	0.76	0.68	360	1939
kmeans	483	355	0.69	1.05	74	0.96	0.91	309	970
mriq	8318	8409	1.27	1.27	1640	0.72	0.72	3	6
pbbs-dr	20057	39785	0.92	1.21	7010	0.59	0.61	429	2966
pbbs-knn	645	850	1.26	1.28	158	0.66	0.68	53	63
pbbs-mm	305	714	0.62	1.02	117	0.67	0.60	452	3836
rsort	268	220	0.83	1.07	42	0.81	0.85	336	754
splash2-fft	4146	2226	1.00	1.00	502	0.97	0.98	9	18
splash2-lu	7780	13045	1.46	1.46	2390	0.68	0.69	5	9
strsearch	1434	1101	1.08	1.09	212	0.91	0.92	17	28
viterbi	3522	4465	0.48	1.03	798	0.75	0.74	558	4599

DInsts = dynamic instruction count in thousands; Trans = transitions per ms. MAVR/RPDN performance and energy results are normalized to SFVR.

Figure 13(a) compares MAVR and RPDN energy efficiency and speedup, both normalized to SFVR. The raw numbers are given in Table II. MAVR has sharp slowdowns for many applications, modest speedups for others, and very high energy efficiency across most applications. Notice that sharp slowdowns generally occur for applications with higher transitions per millisecond (e.g., *dither, kmeans, viterbi*, see Table II). This implies that MAVR response times are too slow for FG-SYNC+ to adapt to the fine-grain activity imbalance. RPDN has higher performance as well as higher energy efficiency across most applications, including those with higher transitions per millisecond, all at similar average power compared to SFVR. Notice that the results in Figure 13(a) look very similar to those in Figure 3(d). Table I explains the similarity: the typical MAVR voltage-settling response time is on the order of 1000 ns while the typical RPDN response time is on the order of 100 ns. RPDN's order-of-magnitude faster response time is a key enabler for achieving good performance and energy efficiency when exploiting fine-grain activity imbalance with FGVS.

Figure 13(b) shows power breakdowns for each application running on our target system with each type of PDN. The results for SFVR include the power of eight cores running the application at nominal voltage, the power lost in the 2:1 SC converter with an 80% conversion efficiency, and leakage power overhead. MAVR consumes significantly less power than SFVR by resting waiting cores, but MAVR has difficulty exploiting this power slack to improve performance due to slow response times. The impact of response time is tightly linked to how often cores transition. A delayed decision is very likely to remain optimal for applications that transition only rarely; these applications have speedups even with slow response times (e.g., *splash2-lu, mriq, strsearch*). Applications with the greatest slowdowns consume the least power



(a) Normalized Energy Efficiency vs Speedup

(b) System-level Power Breakdowns for SFVR, MAVR, RPDN

Figure 13. System-Level Evaluation for SFVR, MAVR, and RPDN – (a) MAVR and RPDN energy efficiency vs. performance normalized to SFVR. Lines connect the same application using both MAVR and RPDN. Raw numbers are given in Table II. (b) Power breakdowns for SFVR, MAVR, and RPDN (S, M, R, respectively) normalized to SFVR. Bar stacks show leakage power, aggregate core power, regulator power efficiency overhead, and transition power overhead.

(e.g., *viterbi*). This indicates that they spend most of their time executing slowly in low-power modes. RPDN closely tracks the average power of SFVR by resting cores and trading power slack for improved performance. RPDN actually achieves *lower* average power compared to SFVR for some applications. The extra power slack is an opportunity for further optimization using a more aggressive online controller. The results for regulator power efficiency overhead show that RPDN does not sacrifice regulator efficiency in exchange for performance. Leakage power and transition power remain fairly small for each type of PDN.

Figure 13 shows that in general, MAVR can offer high energy efficiency but suffers significant slowdowns compared to SFVR due to slow response times. RPDN enables higher performance and energy efficiency at the same or lower average power compared with SFVR, while reducing the area overhead by 40% compared to MAVR. These results suggest RPDN is an attractive option for enabling realistic fine-grain voltage scaling in future embedded systems.

VII. Discussion

In this section, we discuss the impact of di/dt noise, as well as RPDN scalability for higher core counts, higher power densities, and different technologies.

Noise – On-chip power management can potentially diminish di/dt noise issues. One key motivation for moving power management on-chip is to reduce the impact of PCB wires and package parasitics (e.g., bond pads). In addition, using on-chip step-down voltage converters reduces the package supply-plane impedance since lower current is delivered for the same power. Lastly, for on-chip SC regulators, a portion of the flyback capacitance effectively acts as additional decoupling capacitance. While the additional wiring required for RPDN will introduce some parasitic inductance at high frequency, on-chip wires are short compared to board- and package-level wires. Future work can potentially investigate a detailed characterization of di/dt noise for various integrated regulator designs.

Scaling Core Count – Scaling core count by directly scaling the RPDN switch fabric (i.e., each RPDN cell connected to every core) complicates wiring and has high efficiency losses. In this work, we addressed scalability by partitioning the RPDN into two sub-RPDNs, where each sub-RPDN is assigned to a cluster of four cores. We carefully picked our rest, nominal, sprint, and super-sprint levels such that the cells in a single RPDN partition can support any DVFS mode decision made by FG-SYNC+ at high efficiency. When scaling to larger core counts, there may be certain configurations where there simply is not enough intra-partition energy storage to support the desired operating modes. The FGVS controller would need to account for these scenarios and react accordingly. Future work can potentially explore more sophisticated RPDN switch fabrics. For example, one could imagine RPDN fabrics that provide just "nearest neighbor" connectivity or use multiple stages of switching.

Scaling to Higher Power Densities – Our target system has a power density of 25 mW/mm^2 at nominal voltage, but our findings in this paper still hold true for higher power densities. PDN area overhead increases roughly linearly with core power density; this means that high-power, high-complexity cores will still benefit from RPDN area savings over MAVR. For example, a 4× increase in core power density to 100 mW/mm^2 would have 4× larger area overheads for the integrated PDN compared to the core, but the relative area savings of RPDN are the same (i.e., RPDN area overhead of 24% versus MAVR area overhead of 42% still means RPDN saves 40% area compared to MAVR).

Scaling to Different Technologies – Integration of switching regulators on-chip is a recent phenomenon and is a direct result of the impacts of technology scaling. Earlier CMOS generations did not allow for integration of efficient switching regulators due to low-quality switches and low energy-density passives that required high switching frequencies. Analog components generally do not improve with aggressive scaling, but switching voltage regulators are an exception. Better switches in smaller technology nodes such as 28 nm and beyond are likely to improve SC regulator efficiency rather than degrade it. Furthermore, advances in technology that increase capacitor density (e.g., deep-trench capacitors [3, 9]) have the potential to make integrated regulators for high-power systems both relevant and efficient. Finally, as CMOS technology scales, there is potential for increased leakage. In this case, RPDN offers an additional ben-

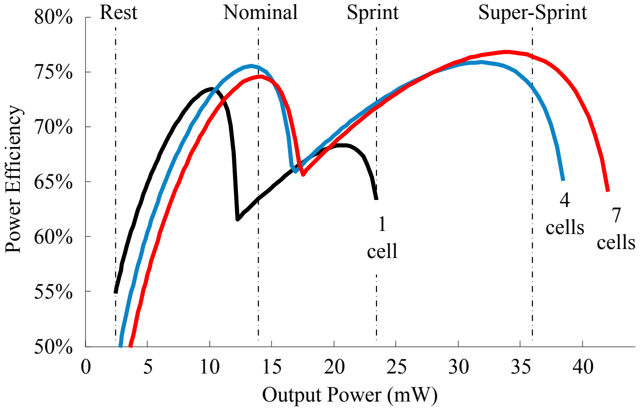


Figure 14. RPDN Power Efficiency vs. Output Power For Single Core in a Leakier 65 nm Process – Each cell is 0.015 mm^2. The 7-cell RPDN curve also represents MAVR area. RPDN tracks the optimal efficiency across all curves, either matching or outperforming MAVR in each operating mode.

efit compared to MAVR. Figure 14 shows energy efficiency vs. output power for a single core as a function of the number of cells allocated to that core (similar to Figure 10), except for a leakier 65 nm process with rest, nominal, sprint, and super-sprint levels scaled accordingly. Notice that in a leakier process at low power, using seven cells results in lower power efficiency compared to using just one cell. RPDN enables reduced flyback-capacitor leakage by using just a single cell when cores are operating in a low-power resting mode.

VIII. Related Work

Most of the previous work in on-chip voltage regulation explores the design space between off-chip and on-chip regulators while focusing on the potential for energy savings at similar performance. [26] conducted sensitivity studies for on-chip Buck regulators and contrasted their lower efficiencies and faster response times with slow, efficient off-chip regulators. The authors took into account overheads for various off-chip and on-chip cases and concluded that on-chip per-core voltage regulation could reduce energy, albeit at large area overhead. The authors determined the optimal supply voltage for each benchmark offline and did not explore online controllers. [31] focused on multiprogrammed applications with an emphasis on saving energy on memory- vs. cpu-bound applications, rather than on exploiting fine-grain activity imbalance in multithreaded workloads.

Other works, such as [47] recognized similar trade-offs as we do in on-chip voltage conversion. The work proposed to use an S-factor-based algorithm to identify workloads with potential for energy savings and then to migrate these threads to a dedicated core powered by an on-chip regulator. With offline training, each application is identified and binned at runtime for migration. This approach was only shown to save energy with multiprogrammed workloads. Taking the same approach for multithreaded applications with fine-grain activity imbalance would likely incur high migration overhead, but the process of identifying workloads offline can complement our work by allowing our controllers to search for complex application-level patterns. In [13], an online learning algorithm for power scaling is proposed that could be implemented on top of our controller to help identify long-term application behaviors.

A switched-capacitor converter is used for dynamic voltage-frequency control in [23]. In this work, the SC converter alternates between different topologies while the core frequency tracks the output voltage ripple. [1, 10] considers Buck converters and focuses on jointly optimizing power consumption of the converter and the core by finding the lowest computational energy point; they considered only steady-state responses. [35] elected to switch cores between two supply rails. This approach requires dedicated supply rails and incurs non-trivial supply pin overhead, which is increasingly important in future technologies [22]. In addition, care must be taken when switching cores between different supply rails by either scheduling power gating events [40], reducing the supply plane impedance, or by increasing decoupling capacitance [15]. Integrated SC converters allow dynamic regulation without the need for large decoupling capacitance while simultaneously relaxing the impedance requirements placed on power-supply routing by utilizing higher off-chip voltages.

IX. Conclusion

Recent trends in technology and the drive to integrate more functionality on chip have generated significant interest in an on-chip voltage regulation, with the goal of reducing cost and enabling fine-grain voltage scaling (FGVS). In this paper we present a new controller, FG-SYNC+, specifically designed for FGVS. Our FG-SYNC+ analysis demonstrates the importance of exploiting fine-grain scaling in level, space, and time. We used insights from our analysis to motivate a new voltage regulation scheme based on the idea of reconfigurable power distribution networks (RPDNs). RPDNs avoid the need to over-provision per-core voltage regulators, thereby reducing regulator area overhead while simultaneously improving performance. Our promising results suggest that there is an important opportunity for architecture and circuit co-design of integrated voltage regulation in future systems.

Acknowledgments

This work was supported in part by NSF CAREER Award #1149464, a Spork Fellowship, and donations from Intel Corporation and Synopsys, Inc. The authors acknowledge and thank Ji Kim for his help in developing the instruction-based energy model, and Derek Lockhart, Shreesha Srinath, and Pol Rosello for their help in writing multicore application kernels.

References

[1] R. A. Abdallah et al. System Energy Minimization via Joint Optimization of the DC-DC Converter and the Core. *Int'l Symp. on Low-Power Electronics and Design*, Jun 2011.

[2] M. Alimadadi et al. A Fully Integrated 660 MHz Low-Swing Energy-Recycling DC-DC Converter. *IEEE Trans. on Power Electronics*, 24(6):1475–1485, 2009.

[3] T. Andersen et al. A 4.6 W/mm^2 Power Density 86% Efficiency On-Chip Switched Capacitor DC-DC Converter in 32-nm SOI CMOS. *Applied Power Electronics Conf. and Exposition*, Mar 2013.

[4] M. Annavaram, E. Grochowski, and J. Shen. Mitigating Amdahl's Law through EPI Throttling. *Int'l Symp. on Computer Architecture*, Jun 2005.

[5] A. Bhattacharjee and M. Martonosi. Thread Criticality Predictors for Dynamic Performance, Power, and Resource Management in Chip Multiprocessors. *Int'l Symp. on Computer Architecture*, Jun 2009.

[6] N. Binkert et al. The gem5 Simulator. *ACM SIGARCH Computer Architecture News*, 39(2):1–7, May 2011.

[7] T. Burd et al. A Dynamic Voltage Scaled Microprocessor System. *IEEE Journal of Solid-State Circuits*, 35(11):1571–1580, Nov 2000.

[8] B. H. Calhoun and A. Chandrakasan. Ultra-Dynamic Voltage Scaling Using Sub-Threshold Operation and Local Voltage Dithering. *IEEE Journal of Solid-State Circuits*, 41(1):238–245, Jan 2006.

[9] L. Chang et al. Fully-Integrated Switched-Capacitor 2:1 Voltage Converter w/ Regulation Capability & 90% Efficiency at 2.3A/mm^2. *Symp. on Very Large-Scale Integration Circuits*, Jun 2010.

[10] Y. Choi, N. Chang, and T. Kim. DC-DC Converter-Aware Power Management for Low-Power Embedded Systems. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 26(8):1367–1381, Aug 2007.

[11] L. T. Clark et al. An Embedded 32-b Microprocessor Core for Low-Power and High-Performance Applications. *IEEE Journal of Solid-State Circuits*, 36(11):1599–1608, Nov 2001.

[12] W. Deng et al. 0.022 mm^2 970 uW Dual-Loop Injection-Locked PLL with -243dB FOM Using Synthesizable All-Digital PVT Calibration Circuits. *Int'l Solid-State Circuits Conf.*, Feb 2013.

[13] G. Dhiman and T. Rosing. Dynamic Voltage Frequency Scaling for Multi-Tasking Systems Using Online Learning. *Int'l Symp. on Low-Power Electronics and Design*, Aug 2007.

[14] J. Donald and M. Martonosi. Techniques for Multicore Thermal Management: Classification and New Exploration. *Int'l Symp. on Computer Architecture*, Jun 2006.

[15] R. G. Dreslinski et al. Reevaluating Fast Dual-Voltage Power Rail Switching Circuitry. *Workshop on Duplicating, Deconstructing, and Debunking*, Jun 2012.

[16] R. G. Dreslinski et al. Near-Threshold Computing: Reclaiming Moore's Law Through Energy-Efficient Integrated Circuits. *Proc. of the IEEE*, 98(2):253–266, Feb 2010.

[17] T. Fischer et al. A 90-nm Variable Frequency Clock System for a Power-Managed Itanium Architecture Processor. *IEEE Journal of Solid-State Circuits*, 41(1):218–228, Jan 2006.

[18] W. Godycki, B. Sun, and A. Apsel. Part-Time Resonant Switching for Light Load Efficiency Improvement of a 3-Level Fully Integrated Buck Converter. *European Solid-State Circuits Conf.*, Sep 2014.

[19] V. Gutnik and A. P. Chandrakasan. Embedded Power Supply for Low-Power DSP. *IEEE Trans. on Very Large-Scale Integration Systems*, 5(4):425–435, Dec 1997.

[20] P. Hazucha et al. Area-Efficient Linear Regulator With Ultra-Fast Load Regulation. *IEEE Journal of Solid-State Circuits*, 40(4):933–940, Apr 2005.

[21] P. Hazucha et al. A 233-MHz 80–87% Efficient Four-Phase DC-DC Converter Utilizing Air-Core inductors on Package. *IEEE Journal of Solid-State Circuits*, 40(4):838–845, Apr 2005.

[22] G. Huang et al. Compact Physical Models for Power Supply Noise and Chip/Package Co-Design of Gigascale Integration. *Electronic Components and Technology Conf.*, May 2007.

[23] R. Jevtic et al. Per-Core DVFS with Switched-Capacitor Converters for Energy Efficiency in Manycore Processors. *IEEE Trans. on Very Large-Scale Integration Systems*, PP(99), 2014.

[24] D. Kanter. Haswell's FIVR Extends Battery Life. Microprocessor Report, The Linley Group, Jun 2013.

[25] W. Kim, D. Brooks, and G.-Y. Wei. A Fully-Integrated 3-Level DC-DC Converter for Nanosecond-Scale DVFS. *IEEE Journal of Solid-State Circuits*, 47(1):206–219, Jan 2012.

[26] W. Kim et al. System-Level Analysis of Fast, Per-core DVFS Using On-Chip Switching Regulators. *Int'l Symp. on High-Performance Computer Architecture*, Feb 2008.

[27] S. S. Kudva and R. Harjani. Fully-Integrated On-Chip DC-DC Converter With a 450X Output Range. *IEEE Journal of Solid-State Circuits*, 46(8):1940–1951, Aug 2011.

[28] H.-P. Le et al. A Sub-ns Response Fully Integrated Battery-Connected Switched-Capacitor Voltage Regulator Delivering 0.19 W/mm^2 at 73% efficiency. *Int'l Solid-State Circuits Conf.*, Feb 2013.

[29] H.-P. Le, S. R. Sanders, and E. Alon. Design Techniques for Fully Integrated Switched-Capacitor DC-DC Converters. *IEEE Journal of Solid-State Circuits*, 46(9):2120–2131, Sep 2011.

[30] J. Lee and N. S. Kim. Optimizing Throughput of Power- and Thermal-Constrained Multicore Processors using DVFS and Per-Core Power-Gating. *Design Automation Conf.*, Jul 2009.

[31] H. Li et al. VSV: L2- Miss-Driven Variable Supply-Voltage Scaling For Low Power. *Int'l Symp. on Microarchitecture*, Jan 2003.

[32] D. Lo and C. Kozyrakis. Dynamic Management of TurboMode in Modern Multi-core Chips. *Int'l Symp. on High-Performance Computer Architecture*, Feb 2014.

[33] A. J. Martinez et al. Haswell: The Fourth-Generation Intel Core Processor. *IEEE Micro*, 34(2):6–20, Mar/Apr 2014.

[34] R. Miftakhutdinov. An Analytical Comparison of Alternative Control Techniques for Powering Next-Generation Microprocessors. TI Application Note, 2005.

[35] T. N. Miller et al. Booster: Reactive Core Acceleration For Mitigating the Effects of Process Variation and Application Imbalance in Low-Voltage Chips. *Int'l Symp. on High-Performance Computer Architecture*, Feb 2012.

[36] J. Park et al. Accurate Modeling and Calculation of Delay and Energy Overheads of Dynamic Voltage Scaling in Modern High-Performance Microprocessors. *Int'l Symp. on Low-Power Electronics and Design*, Aug 2010.

[37] N. Pinckney et al. Limits of Parallelism and Boosting in Dim Silicon. *IEEE Micro*, 33(5):30–37, Sep/Oct 2013.

[38] A. Raghavan et al. Utilizing Dark Silicon to Save Energy with Computational Sprinting. *IEEE Micro*, 33(5):20–28, Sep/Oct 2013.

[39] K. K. Rangan, G.-Y. Wei, and D. Brooks. Thread Motion: Fine-Grained Power Management for Multi-Core Systems. *Int'l Symp. on Computer Architecture*, Jun 2009.

[40] V. Reddi et al. Voltage Emergency Prediction: Using Signatures to Reduce Operating Margins. *Int'l Symp. on High-Performance Computer Architecture*, Feb 2009.

[41] M. Seeman. *A Design Methodology for Switched-Capacitor DC-DC Converters*. Ph.D. Thesis, EECS Department, University of California, Berkeley, May 2009.

[42] M. Seeman et al. A Comparative Analysis of Switched-Capacitor and Inductor-Based DC-DC Conversion Technologies. *Workshop on Control and Modeling for Power Electronics*, Jun 2010.

[43] N. Sturcken et al. An Integrated Four-Phase Buck Converter Delivering 1 A/mm^2 with 700 ps Controller Delay and Network-On-Chip Load in 45-nm SOI. *Custom Integrated Circuits Conf.*, 2011.

[44] D. Truong et al. A 167-Processor Computational Platform in 65-anm CMOS. *IEEE Journal of Solid-State Circuits*, 44(4):1130–1144, Apr 2009.

[45] M. Wens and M. Steyaert. Fully-Integrated CMOS 800-mW Four-Phase Semi-Constant On/Off-time Step-Down Converter. *IEEE Trans. on Power Electronics*, 26(2):326–333, Feb 2011.

[46] J. Wibben and R. Harjani. A High-Efficiency DC-DC Converter Using 2 nH Integrated Inductors. *IEEE Journal of Solid-State Circuits*, 43(4):844–854, Apr 2008.

[47] G. Yan et al. AgileRegulator: A Hybrid Voltage Regulator Scheme Redeeming Dark Silicon for Power Efficiency in a Multicore Architecture. *Int'l Symp. on High-Performance Computer Architecture*, Feb 2012.

[48] Z. Zeng et al. Tradeoff Analysis and Optimization of Power Delivery Networks with On-Chip Voltage Regulation. *Design Automation Conf.*, Jun 2010.

[49] P. Zhou et al. Exploration of On-Chip Switched-Capacitor DC-DC Converter for Multicore Processors Using a Distributed Power Delivery Network. *Custom Integrated Circuits Conf.*, Sep 2011.