Optimization of On-Chip Switched-Capacitor DC-DC Converters for High-Performance Applications

Pingqiang Zhou, Won Ho Choi, Bongjin Kim, Chris H. Kim and Sachin S. Sapatnekar
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455.
{pingqiang, choi0444, kimx2447, chriskim, sachin}@umn.edu

Abstract—On-chip switched-capacitor (SC) DC-DC converters have recently been demonstrated in silicon for high-performance applications such as multicore processors. The efficiency of the power delivery system using SC converters is a major concern, but this has not been addressed at the system level in prior research. This work develops models for the efficiency of such a system as a function of size and layout of the SC converters, and proposes an approach to optimize the size and layout of the SC converter to minimize power loss. The efficiency of these techniques is demonstrated on both homogenous and heterogenous multicore chips.

I. INTRODUCTION

With on-chip processing moving towards a dominant multicore paradigm, the requirements of on-chip power grids are changing. Temporal and spatial variations in on-chip power demands are particularly acute in multicore processors, and trends show that these challenges will become even more difficult in the future.

Greater integration of on-chip power regulation, based on a single external supply, is imperative in order to ensure supply integrity and serve spatially diverse loads [1], [2]. This is easier said than done, and numerous challenges are faced in integrating on-chip supplies. Inductive power supplies can be impractical since on-die inductors have low quality factors and require large area overheads [2]. As a result, in the recent past, there has been a move towards building on-chip capacitance-based DC-DC converters, since capacitors can achieve higher quality factors with lower areas than inductors. Initial efforts [3], [4] have targeted ultra-low power (several mW) applications, but more recent work has resulted in the ability to drive higher power densities, similar to those encountered in multicore CPUs [5], [6]. For example, through the use of trench capacitors, the work in [6] builds converters that can achieve current densities of 2.3A/mm^2 and 90% efficiency under the experimental conditions in the paper.
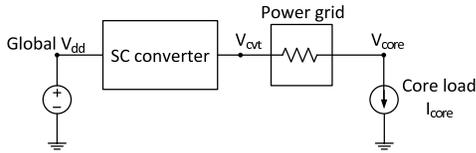

Fig. 1. Schematic of a power delivery system.

Fig. 1 shows a simplified power delivery system including the global V_{dd} supply, a switched-capacitor (SC) converter to convert the input V_{dd} to required voltage supply level, a power grid to distribute the power to local core loads, and a core load. The output of the converters is V_{cvt}, but the exact voltage supply seen by the cores

is downgraded to V_{core} due to losses such as voltage droop (e.g., due to IR drop) in the power delivery network. To overcome these losses and ensure correct core operation, the specification on V_{cvt}, $V_{vdd,dom}$, must be set to

$$V_{vdd,dom} = V_{vdd,core} + V_{droop} + \Delta V \qquad (1)$$

where $V_{vdd,core}$ is the minimum voltage specified at the core load, V_{droop} is the peak voltage droop between V_{cvt} and V_{core}, and ΔV is the peak-to-peak output voltage ripple of the converter. For a core that draws current I_{core}, the power supplied by the converters is:

$$P_{cvt} = I_{core} V_{vdd,dom} \qquad (2)$$

However, the power drawn by the core loads is smaller:

$$P_{core} = I_{core} V_{vdd,core} \qquad (3)$$

The remainder of the power, $I_{core}(V_{droop} + \Delta V)$, is wasted in various parts of the power delivery network.

Prior work on optimizing on-chip capacitive DC-DC converters is very limited. The work in [2] has focused primarily on reducing wasted power *within* the internal design of the converter (i.e., entirely inside the "SC converter" box in Fig. 1) by controlling the the voltage ripple ΔV, optimizing efficiency by choosing the optimal switch width and switching frequency. Under this paradigm, the burden of optimizing the other term for the voltage droop, V_{droop}, (corresponding to the "Power grid" box in Fig. 1) is placed on conventional means for power grid optimization, e.g., grid topology selection and wire widening. The authors in [1] address the problem by suggesting the use of distributed SC converters, which can significantly reduce the voltage droop seen by the local core loads by providing more localized power distribution; however, they have not looked into the efficiency optimization problem.

In this work, we take a novel approach to the problem and consider a more holistic optimization of the DC-DC converter at the system level. We differ from prior efforts in considering not only the internals of the converter but also its context within the system to which it delivers power. In particular, we show that by optimizing the number and layout of the converters for the power domain, it is possible to control the losses due to wasted power in the power grid and enhance the efficiency of the converter. To the best of our knowledge, this is the first work to address efficiency optimization at the system level.

The rest of this paper is organized as follows. In Section II, we present some basic principles of SC converters. This is followed, in Section III, by a description of our proposed models for various components of the power loss as a function of the size and layout of the SC converters in a power delivery system based on SC converters. Next, in Section IV, we present the problem formulation of the efficiency optimization problem, followed by a description of our approaches for solving the problem in Section VI. Finally, in Section VII, the efficiency of our approaches is demonstrated on both homogeneous and heterogeneous multicore chips.

II. SC DC-DC CONVERTERS

A block diagram of a general SC converter system is shown in Fig. 2(a). The system consists of N_{phase} interleaving stages (a typical value of N_{phase} is 32), which reduce the ripple voltage by $1/N_{phase}$

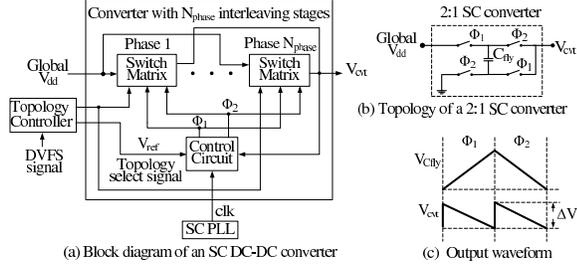compared to an SC converter without any interleaving.

Fig. 2. SC DC-DC converter.

At the core of the system is the switch matrix, one for each phase [7]. This matrix is a reconfigurable arrangement of switches and flying capacitors that is configured in different ways by the "Topology select" signal from the topology controller. Each such configuration provides the ability to produce a different voltage conversion ratio, allowing the converter to generate one of several output voltage levels from the converter [3]: for simplicity, these details are not shown here. The conversion ratio of the converter, $ratio_{cvt}$, is defined as the ratio between the input voltage, which is the external supply voltage, V_{dd}, and the desired output voltage, $V_{vdd,dom}$, which is the specification for the ideal value of V_{cvt}. The control circuit takes these inputs:

- the clock signal clk from a phase-locked loop (PLL)
- the reference voltage for a particular topology V_{ref}
- the feedback voltage V_{cvt} from the converter output

It generates the nonoverlapping clock signals Φ_1 and Φ_2 for the switches in the switch matrix, and may also be used to gate some of the capacitors to control the amount of capacitance that takes part in the charge transfer process [4].

A switch matrix topology is shown in Fig. 2(b), with a 2:1 conversion ratio. Fig. 2(c) (top) shows that during Φ_1, the flying capacitor C_{fly} is connected to the input global V_{dd} to get charged, and during Φ_2, the charge stored in C_{fly} is transferred to the load and its voltage drops by ΔV as it is discharged. This is reflected as the output voltage at the output, V_{cvt} of the converter in Fig. 2(a), as shown in Fig. 2(c) (bottom) in Φ_2. Note that another switch matrix is connected to the output during Φ_1 (and is charged during Φ_2), which results in the voltage ripple observed in the V_{cvt} waveform.

Note that the signals Φ_i are generated by a relatively low-frequency clock ($f_{sw} \approx 100MHz$), which is distinct from the multi-GHz clock used by the multicore processor.

III. POWER LOSS ANALYSIS

Efficiency is one of the key design metrics for the on-chip DC-DC converters [2], [8]. We now analyze the inefficiency and power loss in a SC converter. Our analysis is based on [2], [7], [9], as well as from conversations with designers. Some items in this section are taken from the literature, while others are freshly derived.

For each converter, let f_{sw} be the switching frequency of the converter, $C_{sw} = C_{fly} \times N_{phase}$ be the total amount of flying capacitance, and ΔV be the output ripple of the converter.
(1) Conduction loss: This corresponds to the power loss in the switches as the flying capacitors are charged. For each converter, the conduction loss is modeled as:

$$P_{cond} = M_{sw} \frac{I_{out}^2}{N_{phase}} \frac{R_{on}}{W_{sw}} \qquad (4)$$

where M_{sw} is a constant determined by the converter topology (Table I), I_{out} is the total current delivered by the converter, R_{on} is the switch resistance per unit width, and W_{sw} is the switch width.

Conversion ratio	M_{sw}	γ	M_p	M_{topo}
1:1	1	1	0	1/2
4:3	7/3	2/3	$3/8\alpha$	8/9
3:2	1	1	$1/3\alpha$	9/8
2:1	2	2	$1/4\alpha$	2

TABLE I
M_{sw}, γ, M_p AND M_{topo} FOR DIFFERENT TOPOLOGIES [7]. α IS THE RATIO OF THE PLATE CAPACITANCE TO ITS EFFECTIVE CAPACITANCE.

For a given topology, W_{sw} is proportional to f_{sw} and C_{sw}:

$$W_{sw} = \sigma\gamma f_{sw} \frac{C_{sw}}{N_{phase}} \qquad (5)$$

where σ is a fitting coefficient, and γ is topology-dependent (Table I). In an SC converter supporting DVFS, the switch size may be adjustable, where some of a set of parallel switches are turned on to achieve the desired switch size [9].
(2) Gate-drive loss of the switches: The switches in a converter are implemented using transistors. These transistors must be very wide in order to minimize conduction losses, and therefore the power loss in driving their gate nodes can be modeled as:

$$P_{sw} = N_{phase} \cdot N_{sw} \cdot f_{sw} \cdot (C_{gate}W_{sw}) \cdot V_{dd}^2 \qquad (6)$$

where N_{sw} is the number of switches used in one particular topology and C_{gate} is the per-unit-width gate capacitance of the switches.
(3) Parasitic loss: This is the loss from the bottom-plate parasitic capacitance of the flying capacitors. The loss can be estimated as:

$$P_{para} = M_p f_{sw} C_{sw} V_{dd}^2 \qquad (7)$$

where M_p is a parameter that depends on the internal structure of a topology (Table I). This loss component depends on the particular type of the capacitance technology. Deep trench capacitors typically have superior efficiency compared to MIM and CMOS capacitors.
(4) The load power loss: The load power loss $I_{core}(V_{droop} + \Delta V)$, described in Section I, can be separated into two parts:
(4a) The part determined by the voltage ripple, ΔV, is

$$P_{L1} = I_{core}\Delta V \qquad (8)$$

In each cycle, the energy a topology can deliver is given by $M_{topo}C_{sw}N_{phase}\Delta V$, where M_{topo} is determined by the topology (Table I), because with the same amount of flying capacitance C_{sw}, different topologies can deliver different amount of power to the output. When switching at frequency f_{sw}, the current a converter can provide is

$$I_{out} = M_{topo} \cdot f_{sw} \cdot C_{sw} \cdot N_{phase} \cdot \Delta V \qquad (9)$$

i.e., $$\Delta V = \frac{I_{out}}{M_{topo} f_{sw} C_{sw} N_{phase}} \qquad (10)$$

From Equation (10), we can see that with the same output current I_{out}, the voltage ripple ΔV is inversely proportional to the size of charge-transfer capacitance C_{sw}.
(4b) The power loss associated with the voltage droop, V_{droop}, is

$$P_{L2} = I_{core}V_{droop} \qquad (11)$$

Note that the voltage droop changes as we alter the number and locations of the converters on the chip, since the distance between the converters and the utilization points (cores) changes.
(5) Control circuit and clock network: The control unit generates the nonoverlapping clock signals for the switches used in the converter. This unit includes a voltage comparator, DLL and control logic. The power loss of the clock network arises from the wire capacitance, the clock buffers inserted for the wires, and the clock loads. The power losses from control unit P_{ctrl} and clock network P_{clock} are both dependent on the number of used converters N_{cvt}. We use a penalty term for these two items in the objective formulation, as stated in Section V.
(6) Clock sources: The clock source is implemented as a simple PLL with relaxed frequency ($\approx$ 100MHz) and jitter (less than tens of ps) requirements compared to the main PLL for the on-chip circuit. Thus, the power consumption of the clock source is $P_{clksrc} = P_{PLL}$,

where P_{PLL} is the power consumption of one PLL [10].

(7) Topology controller: This generates the signals that provide DVFS directives to reconfigure the topology in each converter to set the conversion ratio that provides the desired voltage output level. The topology controller is a small combinational logic block and its power consumption is in the order of μW, which is ignored here.

IV. OPTIMIZATION FORMULATION

In the scenario studied here, it is safe to assume that the switching frequency f_{sw} and interleaving stages N_{phase} are fixed for the converters. Based on the analysis in Section III, the components of power loss can be divided into four categories.

The first component, which depends on the parameters of the converter, is the power consumption of the conduction loss/gate-drive loss of the switches/parasitic loss/part of load loss P_{L1}, and is determined by the C_{sw} and the global V_{dd}, as:

$$P_1 = P_{cond} + P_{sw} + P_{para} + P_{L1} \tag{12}$$

For each converter, we can change the total flying capacitance, C_{sw}, to tune the voltage ripple ΔV, according to Equation (10). A larger C_{sw} results in smaller ΔV, and can therefore reduce the *load power P_{L1}* (Equation (8)) and *switch conduction loss P_{cond}* (Equations (4) and (5)). On the other hand, the *gate switching loss P_{sw}* (Equations (5) and (6)) and *parasitic loss P_{para}* (Equation (7)) increase with C_{sw}. An optimal value of C_{sw} balances these conflicts.

The second and third components are, respectively, the power consumption of part of load loss P_{L2}, and the sum of the power loss in the control circuit and clock network.

$$P_2 = P_{L2} \tag{13}$$
$$P_3 = P_{ctrl} + P_{clock} \tag{14}$$

Both P_2 and P_3 are determined by the number and layout of the converters. Changing the granularity of the capacitance through more fine-grained distributed converters placed over the chip (as opposed to a single centralized converter) can help reduce the voltage droop seen by the core loads, therefore reduce the loss P_{L2} [1]. However, using a larger number of converters implies higher cost for the hardware implementation due to higher losses in the control circuit and clock network. Therefore, it is necessary to explore the number and layout of the DC-DC converters to determine an optimum.

The last component, corresponding to the loss of the clock sources is fixed and given by

$$P_4 = P_{clksrc} \tag{15}$$

At the system level, the efficiency of the power delivery system η is defined as the ratio between *power delivered to the load* and *total power extracted from the input V_{dd} supply*, i.e.,

$$\eta = \frac{P_{core}}{P_{core} + P_1 + P_2 + P_3 + P_4} \tag{16}$$

where P_{core} is defined in Equation (3). To increase the efficiency, we minimize the sum of P_1 through P_4, which constitute the power wasted during power delivery. Further, since P_4 is a fixed quantity, to improve the overall efficiency of the power delivery system using SC converters, we should optimize the objective function:

$$\text{minimize} \quad P_1 + P_2 + P_3 \tag{17}$$

The variables in the optimization problem are

- the number of converters used, N_{cvt},
- the capacitance of each used converters C_{sw}, and
- the locations of the converters.

The optimization is subject to the following constraints:

1) The supply voltage at each core load must meet a lower bound:

$$V_{core} \geq V_{vdd,core} \tag{18}$$

2) Since the voltage ripple constraint must limit $\Delta V \leq \Delta V_{max}$, Equation (10) provides a bound on C_{sw}:

$$C_{sw} \geq \frac{I_{out}}{M_{topo} f_{sw} N_{phase} \Delta V_{max}} \tag{19}$$

3) To control the capacitance resource used, we require that:

$$\sum C_{sw} \leq C_{max} = C_{unit} \cdot Area_{max} \tag{20}$$

where C_{unit} is the capacitance density, and $Area_{max}$ is the maximum available area for the converters.

V. MINLP FORMULATION

Fig. 3(a) presents a schematic of the on-chip power delivery network for a multicore processor. The on-chip power delivery network consists of a global V_{dd} supply, on-chip DC-DC converters, the power grid, and core loads. The voltage supplied to the power grid controlled by a set of on-chip SC converters, which can be placed at a list of predefined candidate locations on the chip.
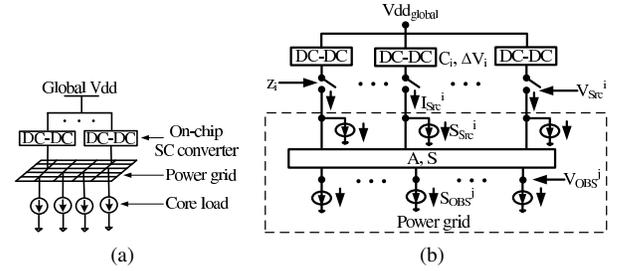

Fig. 3. (a) Model of power delivery network (b) Network macromodel with m candidate converters and n observation nodes.

In the following sections, we show that the optimization problem in Section IV can be formulated as a mixed-integer nonlinear program problem (MINLP), by introducing 0–1 integer variables z_is, with $z_i = 1$ denoting that a converter is placed at candidate location i. We first macromodel the power grid in Section V-A, and then present the complete MINLP formulation in Section V-B.

A. Macromodeling of the power grid

The power grid may have millions of nodes, but we are only interested in OBS, the selected n observation nodes of the core loads, and Src, the m predefined candidate connection nodes for the SC converters. Therefore, we build a macromodel whose ports are these $n + m$ nodes, and abstract away all of the other nodes in the network using the macromodeling approach [11]. Therefore, Fig. 3(a) is transformed to the model shown in Fig. 3(b).

The DC analysis of a V_{dd} power grid is formulated as:

$$Gv = i \tag{21}$$

where G is the conductance matrix for the interconnected resistors, v is the vector of node voltages, and i is the vector of current loads. The equations for the power grid are given as

$$\begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} \begin{bmatrix} V \\ U \end{bmatrix} = \begin{bmatrix} -J_1 + I \\ -J_2 \end{bmatrix} \tag{22}$$

where U and V are voltages of the internal nodes and ports, J_1 and J_2 are current sources connected at ports and internal nodes, and I is the vector of current flowing into the macromodel through the ports. The macromodel of the power grid including only the port nodes (cores' accessing nodes OBS and the candidate nodes for the converters Src) is given by

$$I = AV + S \tag{23}$$

where $A = G_{11} - G_{12}G_{22}^{-1}G_{21}$, and $S = J_1 - G_{12}G_{22}^{-1}J_2$. By partitioning the ports into sets Src and OBS, this can be rewritten as

$$\begin{bmatrix} I_{Src} \\ I_{OBS} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} V_{Src} \\ V_{OBS} \end{bmatrix} + \begin{bmatrix} S_{Src} \\ S_{OBS} \end{bmatrix} \tag{24}$$

where (I_{Src}, V_{src}) and (I_{OBS}, V_{OBS}) are the (current,voltage) values at the Src and OBS ports. Since $I_{OBS} = 0$, we have:

$$V_{OBS} = T \cdot V_{Src} + B \tag{25}$$

where $T = -A_{22}^{-1}A_{21}$, and $B = -A_{22}^{-1}S_{OBS}$. Further,

$$I_{Src} = A_{11}V_{Src} + A_{12}V_{OBS} + S_{Src} = A'V_{Src} + S'_{src} \tag{26}$$

where $A' = A_{11} + A_{12}T$ and $S'_{src} = S_{Src} + A_{12}B$.

From Equations (25) and (26) we can see that the current vector of the Src ports I_{Src} and voltage vector of the OBS ports V_{OBS} are linear functions of the voltage vector of the Src ports V_{Src}.

B. MINLP Formulation

Using the macromodel shown in Fig. 3(b), the optimization problem described in Section IV is equivalent to finding the optimal z_i assignments, and for each used converter i (with $z_i = 1$), determining its size C_i and voltage ripple ΔV_i.

We rewrite P_1 (Equation (12)), the power loss associated with the converter and the global V_{dd} supply, as:

$$P_1 = \sum_{i=1}^{m} \left(e_1 e_3 I_{Src}^i \Delta V_i + e_2 V_{vdd,dom}^2 C_i \right) \tag{27}$$

where

$$
\begin{aligned}
e_1 &= \left(\frac{1}{N_{phase}M_{topo}} + \frac{M_{sw}R_{on}}{\sigma\gamma} \right) \frac{1}{f_{sw}} \\
e_2 &= f_{sw} \left(N_{sw}C_{gate}f_{sw}\sigma\gamma + M_p \right) \cdot ratio_{cvt}^2 \\
e_3 &= N_{phase}M_{topo}f_{sw}
\end{aligned}
$$

Using Equation (25), P_2, the power loss in the grid, and P_3 are:

$$
\begin{aligned}
P_2 &= \underbrace{\sum_{i=1}^{m}(V_{Src}^i(I_{Src}^i - S_{Src}^i))}_{\text{Power supplied to the macromdel}} - \underbrace{\sum_{j=1}^{n}(V_{OBS}^j S_{OBS}^j)}_{\text{Power delivered from the macromodel}} \\
&= \sum_{i=1}^{m} \left(V_{Src}^i(I_{Src}^i - S_{Src}^{'i}) \right) - \sum_{j=1}^{n}(B^j S_{OBS}^j) \tag{28}
\end{aligned}
$$

$$P_3 = P_{ctrl} + P_{clock} = c \cdot \sum_{i=1}^{m} z_i \tag{29}$$

where c is penalty weight for control circuit and clock network, $V_{vdd,dom}$, V_{Src}^i, I_{Src}^i, C_i, ΔV_i are the continuous variables and z_is are the 0–1 integer variables in the optimization problem.

Then we can transform the optimization problem defined in Section IV into a MINLP formulation as

$$
\begin{aligned}
\text{min.} \quad P_1 + P_2 + P_3 &= \sum_{i=1}^{m} \left(e_1 e_3 I_{Src}^i \Delta V_i + e_2 V_{vdd,dom}^2 C_i \right) \\
&+ \sum_{i=1}^{m} \left(V_{Src}^i(I_{Src}^i - S_{Src}^{'i}) \right) - \sum_{j=1}^{n}(B^j S_{OBS}^j) + c\sum_{i=1}^{m} z_i \tag{30}
\end{aligned}
$$

subject to

$\forall j \in$ OBS:

$$V_{OBS}^j = \sum_{i=1}^{m}(T_{ji} \cdot V_{Src}^i) + B^j \geq V_{th}^j \tag{31}$$

$\forall i \in$ Src:

$$I_{Src}^i = \sum_{k=1}^{m}(A'_{ik} \cdot V_{Src}^k) + S_{Src}^{'i} \tag{32}$$

$$0 \leq I_{Src}^i \leq M \cdot z_i \tag{33}$$

$$I_{src}^i = e_3 \cdot \Delta V_i \cdot C_i \tag{34}$$

$$0 \leq C_i \leq M \cdot z_i \tag{35}$$

$$0 < \Delta V_i \leq \Delta V_{max} \tag{36}$$

$$V_{Src}^i + \Delta V_i \leq V_{vdd,dom} \tag{37}$$

and

$$\sum_{i=1}^{m} C_i \leq C_{max} \tag{38}$$

Here, V_{th}^j is the minimum required voltage at the observation nodes of each core, and M is a large positive number.

Constraints (31) are transformed from Equation (18), to specify the minimum voltage for each core load. Constraints (32) are from Equation (26), and Constraints (34) from Equation (10). Constraints (33) are structured to ensure that the current I_{src}^i is zero when no converter connected to candidate port i, while Constraints (35) ensure that converter size C_i is zero when I_{src}^i is zero, both through the use of M. Constraints (36) and (38) are from Equations (19) and (20), and Constraints (37) set the bound for the Vdd supply.

We can observe that there are nonlinear (actually non-convex) terms in the objective function (30) and constraints (34) are also nonlinear. Therefore, the above optimization problem is a MINLP.

VI. HEURISTIC APPROACHES

As stated in [12], "MINLP problems are difficult to solve precisely, because they combine all the difficulties of both of their subclasses: the combinatorial nature of mixed integer programs (MIP) and the difficulty in solving nonconvex (and even convex) nonlinear programs (NLP). Because subclasses MIP and NLP are among the class of theoretically difficult problems (NP-complete), so it is not surprising that solving MINLP a challenging and daring venture."

Therefore, in our work we explore heuristic approaches to solve the optimization problem. For the objective function in Equation (30),

- $P_2 + P_3$ is determined by the number/layout of the converters
- P_1 is determined by the converter design, i.e, the size of converters C_i, and $V_{vdd,dom}$, the V_{dd} supply. From Equation (1) we can see that $V_{vdd,dom}$ is determined by the voltage droop in the power grid and the ripple in the converters.

Therefore, we may optimize the power loss in two steps. We first optimize $P_2 + P_3$, the power in the distribution network, by finding the optimal number and layout of the converters. We present two heuristic approaches in Section VI-B for this step. Next, we optimize P_1 to determine the optimal size of each used converter C_i, which is presented in Section VI-C.

A. An approximation for the voltage ripple

We introduce the approximation that all converters have the same voltage ripple. In other words,

$$\Delta V_i = \Delta V \ \forall \ i \text{ such that } z_i = 1.$$

The impact of this assumption is that by Equation (34), the current delivered by a converter i is proportional to its capacitance C_i, which is a reasonable assumption.

We justify this approximation as follows. In Equation (27), let P_1^i be the contribution of the ith converter to P_1. If $z_i = 1$,

$$P_1^i = e_1 e_3 I_{Src}^i \Delta V_i + e_2 V_{vdd,dom}^2 C_i \tag{39}$$

According to Equation (34), P_1^i is equivalent to

$$P_1^i = e_1 \frac{(I_{Src}^i)^2}{C_i} + e_2 V_{vdd,dom}^2 C_i \tag{40}$$

If we minimize P_1^i locally by setting $\partial P_1^i / \partial C_i = 0$, we get

$$C_i = \frac{I_{Src}^i}{V_{vdd,dom}} \sqrt{\frac{e_1}{e_2}} \tag{41}$$

Therefore, according to Equation (34) we can see that

$$\Delta V_i = \frac{I_{Src}^i}{e_3 C_i} = \frac{V_{vdd,dom}}{e_3} \sqrt{\frac{e_2}{e_1}} \tag{42}$$

Since e_1, e_2, and e_3 are constants, and $V_{vdd,dom}$ is common to all the converters, ΔV_is can be assumed to be the same among the used

converters if they are locally optimized. Therefore, in the following discussion, we assume $\Delta V_i = \Delta V$ for each used converter.

If all C_is were free variables, allowed to take any value, this would not be an approximation. However, according to Equation (38), the C_is are not unconstrained, therefore this is an approximation.

B. Optimizing Converter Number/Layout

As stated earlier, the number and layout of the converters also affects the efficiency of the power delivery system. Distributing the converters with finer granularity and optimized layout over the chip can help improve the efficiency loss by reducing the voltage droop seen by the local core loads, when placing the converters closer to the utilization points. However, there is an overhead associated with the power loss in the control units and clock network.

1) How significant is the converter area?: At this point, it is useful to consider some technology numbers to determine the area overheads of the SC converters. To compute this, we assume that the SC converters are fabricated using deep-trench capacitors. In [6], the reported power density of deep-trench capacitors is 200nF/mm^2. A typical core has the current of $\sim$ 1A. According to Equation (9), if we use a 2:1 converter (with $M_{topo} = 2$) to deliver this amount of current with ripple $\Delta V = 5$mV, $N_{phase} = 32$ and $f_{sw} = 100$Mhz, then the required amount of capacitance is 31.25nF, which transforms to 0.156mm^2. Considering that the typical size of a core is of several mm^2, we may ignore the area effect of the converters when optimizing the layout of the converters. Of course, we can extend our general methodology described in this section to deal with other kinds of capacitors such as the MIM capacitor, by considering the area effect in exploring the granularity of the converters, but this is a topic for future work.

2) MILP-based Approach: In this section, we present an MILP-based approach by reducing the MINLP problem in Section V through a natural approximation and relaxation process.

We proceed under the assumption that for each used converter, $\Delta V_i = \Delta V$, and define

$$V_{vdd,local} = V_{vdd,dom} - \Delta V \qquad (43)$$

From Equation (37) we can see that

$$V_{Src}^i \leq V_{vdd,local} \qquad (44)$$

The power loss due to voltage droop, P_2, shown in Equation (28), can be relaxed as

$$P_2 = \sum_{i=1}^{m}(V_{Src}^i I_{Src}^i) - \sum_{i=1}^{m}(S_{Src}^{'i}V_{Src}^i) - \sum_{j=1}^{n}(B^j S_{OBS}^j)$$

$$\leq V_{vdd,local}\sum_{i=1}^{m}I_{Src}^i - \sum_{i=1}^{m}(S_{Src}^{'i}V_{Src}^i) - \sum_{j=1}^{n}(B^j S_{OBS}^j) \qquad (45)$$

Essentially, since $I_{src}^i = 0$ when $z_i = 0$, the substitution in the first term means that $V_{Src}^i = V_{vdd,local}$. In the above expression, $\sum_{i=1}^{m}I_{Src}^i$ is the total current delivered to the cores, and therefore, a constant. We can see that by relaxation we can transform the nonlinear cost function P_2 to be linear.

In fact, in our experiments using all approaches, we find that V_{Src}^i is nearly equal for every converter i, so that (44) is in practice an equality, confirming the validity of the minimizing the relaxed P_2.

Since $\sum_{j=1}^{n}(B^j S_{OBS}^j)$ is a constant, it is unchanged under any optimization. Then the relaxed power loss $(P_2 + P_3)$ can be minimized by solving the following MILP problem:

$$\text{min.} \quad V_{vdd,local}\sum_{i=1}^{m}I_{Src}^i - \sum_{i=1}^{m}(S_{Src}^{'i}V_{Src}^i) + c\sum_{i=1}^{m}z_i \qquad (46)$$

subject to the linear constraints in Equations (31), (33) and (44).

Note that I_{Src}^i is substituted with V_{Src}^i according to Equation (32), so this MILP formulation has m 0-1 integer variables (z_is), $m + 1$ continuous variables ($V_{vdd,local}$ and V_{Src}^is) and $3m + n$ constraints.

3) Greedy Approach: Considering that MILP can be expensive for a large number of integer variable z_is, we propose a greedy approach to reduce the run-time complexity of solving the optimization problem with a large set of candidate locations for the converters. The idea is to explore different granularity of converters: from one converter for each core, to a single lumped converter for all the cores.

For a chip with l cores, the inputs of the greedy approach include

1) A list of cores $\Re = \{C_0, \ldots, C_l\}$. Core C_i has peak current I_i and minimum required voltage supply V_{vdd,C_i},
2) A adjacency graph G_0 representing the neighbor relationships among the l cores; if a layout is provided instead, this information can be generated using Voronoi diagrams.
3) A list of all candidate locations $\Psi = \{\psi_1, \ldots, \psi_m\}$ for the converters on the chip (Fig. 5 shows part of the candidate set that are used by the converters).

The edge weight w_{ij} of an edge between vertices i and j in the adjacency graph is calculated as the increase in the power loss from combining two converters V_i and V_j into a single converter, V_{ij}. This quantity is the total change in the power loss $P_2 + P_3$, which includes:

1) the change in power loss from voltage droop [Equations (1), (2) and (11)] $\Delta P_{L2} = \Delta V_{vdd,dom} \cdot \sum_{i=1}^{l}I_i$
2) the change in power loss from the control circuit ΔP_{ctrl}
3) the change in power loss from the clock network ΔP_{clock}

i.e, $w_{ij} = \Delta P_{L2} + \Delta P_{ctrl} + \Delta P_{clock}$ where ΔP_{L2} is non-negative because voltage droop tends to increase with fewer converters, $\Delta P_{ctrl} = -P_{ctrlr}$ because the number of converters is reduce by one after combining two converters into one, and ΔP_{clock} is determined by the locations of the converters V_i, V_j and V_{ij}. Note that w_{ij} can be negative in our approach.

Our approach to optimizing the converter design is iterative in nature, and the overall scheme is illustrated in the left half of Fig. 4. We begin with a design with one individual converter for each core. The top right box in Fig. 4 shows an example of the given adjacency graph G_0 for the l cores. In G_0, each node V_i represents the converter for core C_i.
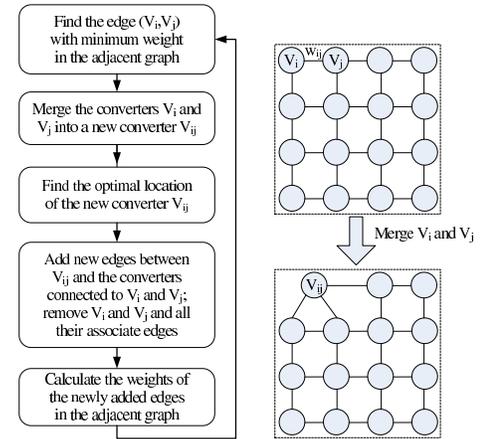

Fig. 4. Outline of the proposed approach to explore different granularity of converters.

The principle behind our method is to begin with the adjacency graph, allowing each core to have its own converter. Then we contract edges in the graph to reduce the number of converters by merging the adjacent converters. Starting from a given adjacency graph G_0 with l converters, at each iteration we greedily merge the neighboring converters V_i and V_j with minimum edge weight w_{ij}, so as to minimize the possible increase of power loss at the next level of converter granularity. When merging two neighboring converters V_i and V_j, two nodes in the adjacent graph is merged into one new

node, and the weights of the edges between this new node edge and its neighbours are updated as stated earlier.

We compute the optimal location, as described in the next paragraph, for the combined converter V_{ij}, and then update the adjacency graph. With l cores, our approach will repeat the merging process $l-1$ times to evaluate all possible levels of converter granularity.

We select the location of a converter V_i from the set of candidate locations Ψ to minimize the nominal output voltage of the converters, minus the voltage ripple part [Equation (1)], i.e.,

$$V_{vdd,local} = V_{vdd,dom} - \Delta V = \max_{i \in \{1,...,l\}} (Vdd_{C_i} + V_{droop,i}) \quad (47)$$

where $V_{droop,i}$ is the voltage drop at core C_i. When evaluating each candidate location, the voltage droop of each core can be obtained from the simulation of the power grid. However, consider that the power grid is typically costly to simulate, to speed up the evaluation process, we assume that the conduction resistance between a core C_i and its converter V_j is linearly proportional to their distance $Dist(C_i, V_j)$, i.e., $V_{droop,C_i} = I_i \cdot R_{unit} \cdot Dist(C_i, V_j)$, where R_{unit} is the unit-distance resistance of the power grid. However, the voltage droop for our final results are validated using a accurate circuit simulator.

C. Optimization of Converter Size

After determining the number and layout of converters using the heuristic approaches in Section VI-B, the second step is to determine C_i for each converter i by optimizing P_1.

Let $I_{total} = \sum_{i=1}^{m} I_{Src}^i$ and $C_{total} = \sum_{i=1}^{m} C_i$, then from Equation (42) we can see that

$$\Delta V = \frac{I_{Src}^i}{e_3 C_i} = \frac{I_{total}}{e_3 C_{total}} \quad (48)$$

so to minimize the power loss P_1 in Equation (27) is equivalent to minimizing

$$
\begin{aligned}
P_1 &= e_1 e_3 \Delta V I_{total} + e_2 V_{vdd,dom}^2 C_{total} \\
&= e_1 I_{total}^2 \frac{1}{C_{total}} + e_2 V_{vdd,dom}^2 C_{total} \quad (49)
\end{aligned}
$$

Using Equation (43), Equation (49) can be further transformed to

$$
\begin{aligned}
P_1 &= e_1 I_{total}^2 \frac{1}{C_{total}} + e_2 (V_{vdd,local} + \Delta V)^2 C_{total} \\
&= e_2 V_{vdd,local}^2 C_{total} + I_{total}^2 (e_1 + \frac{e_2}{e_3^2}) \frac{1}{C_{total}} \\
&\quad + \frac{e_2}{e_3} V_{vdd,local} I_{total} \quad (50)
\end{aligned}
$$

where I_{total} is a constant, and $V_{vdd,local}$ can be found after solving the optimization problem in Section VI-B. The constraints for the above problem is given by Equation (38) and

$$C_{min} = \frac{I_{total}}{e_3 \Delta V_{max}}, \quad (51)$$

which is derived from Equations (36) and (48).

Note that P_1 is a *convex* function of C_{total}. It is easily determined that the optimal solution to the unconstrained problem defined in Equation (50) is given by:

$$C_0 = \frac{I_{total}}{V_{vdd,local}} \sqrt{\frac{e_1 + \frac{e_2}{e_3^2}}{e_2}} \quad (52)$$

However, this value of C_0 may fall outside the bounding constraints (38). If so, from the convexity of the objective function, we can conclude that the optimum must be at the extreme point of the allowable C_{total} interval that is closer to C_0.

Next, the optimal size of C_{total} for the converters, C_{opt}, is

$$C_{opt} = \begin{cases} C_{min} & \text{if } C_0 < C_{min} \\ C_0 & \text{if } C_{min} \leq C_0 \leq C_{max} \\ C_{max} & \text{if } C_0 > C_{max} \end{cases} \quad (53)$$

Then we can calculate the voltage ripple ΔV according to Equation (48) using C_{opt}, and the optimal size of each used converter

C_i can be calculated by Equation (48) because I_{Src}^i is known after solving the optimization problem in Section VI-B.

VII. EXPERIMENTAL RESULTS

Our heuristic approaches described in Section VI are implemented in C++. The MILP problem is solved using CPLEX [13].

A. Test Cases

Our approaches were exercised on two chips, one of which is a homogeneous multicore while the other is a heterogenous multicore processor. The configuration of each chip is described below:
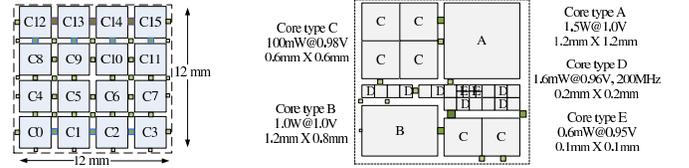

Fig. 5. Two test cases with 16 homogeneous cores (left) and 32 heterogeneous cores (right)

Homogeneous Chip: Our homogeneous test case consists of a chip with one power domain of 16 identical cores, as shown in Fig. 5 (left), which follows the tile-based design for multicore chip [14]. Each core consists of a CPU, L1 I/D cache and L2 cache with area ratio of 2:1:2. The core is $3 \times 3 mm^2$ with a peak current of 1A@0.6V. In our simulations, we model the current ratio among CPU, L1 cache and L2 cache inside each core using guidelines consistent with [15].

Heterogeneous Chip: We also consider a heterogeneous test case consisting of a set of ARM Cortex cores [16]. Simpler versions of such heterogeneous cores are already on the market today [17]. This test case has one power domain of 32 cores as shown in Fig. 5 (right). Core types A through E are, respectively, the A9, A8, A5, M4, and M0 cores.

Table II shows our experimental parameters in the 32nm technology node based on the published literature and PTM [18]. We assume the available converter area to be up to 20% of the total core area.

Individual parameters	Homo16, Hete32	Common parameters	
$Ratio_{cvt}$	2:1, 3:2	f_{sw}	100Mhz
I_{total}	16A, 3.14A	N_{phase}	32
ΔV_{max}	10mV, 20mV	C_{unit}	200nF/mm^2
$Area_{max}$	28.8mm^2, 1.056mm^2	C_{gate}	3fF/μm
C_{max}	5.76 μF, 0.21μF	R_{on}	130$\Omega \cdot \mu$m
N_{sw}	4, 7	c	4.0mW
M_{topo}	2, 9/8	α	0.1%
-	-	σ	512μm/(μF·MHz)

TABLE II
CONFIGURATIONS OF THE TWO CHIPS.

B. Comparison of Heuristic Approaches

We have presented two heuristic approaches for the optimization of the number and layout of the converters in Section VI-B, followed by the optimization of converter size using a closed-form solution. The first heuristic approach (refer to Section VI-B2) *Heuristic-MILP* formulates the optimization as a MILP problem, and the second heuristic approach *Greedy* in Section VI-B3 uses *greedy* strategy to explore the number and layout of converters at different levels of granularity. We compare these two approaches with a manual design approach, which evenly distributes the converters over the chip at different levels of granularity with total number of converters set to 2^k, $k = 0, 1, 2, \ldots, \lfloor \log_2^m \rfloor$, where m is the numbers of candidate locations for the converters.

Table III shows the results of these approaches. Columns 2–3 show m, the numbers of candidate locations for the converters, and n, the number of observation nodes for the cores. Columns 4–9 show

TABLE III
COMPARISON OF OPTIMIZATION EFFICIENCY, WITHOUT LIMITATION ON # CONVERTERS

Chip	m	n	Manual						Greedy							Heuristic-MILP						
			#cvts	P1	P2	P3	Total	η	#cvts	P1	P2	P3	Total	η	CPU	#cvts	P1	P2	P3	Total	η	CPU
Homo16	56	208	32	763	574	128	1465	86.1	16	773	1024	64	1861	82.5	2.9	47	705	283	188	1176	88.1	370.1
Hetero32	76	203	16	160	277	64	501	86.1	11	157	184	44	385	88.9	1.7	13	157	141	52	350	90.1	362.7
Average					1		1											0.50		0.75		

TABLE IV
COMPARISON OF OPTIMIZATION EFFICIENCY, WITH SAME LIMITATION ON NUMBER OF CONVERTERS

Chip	m	Max. #cvt	n	Manual						Greedy							Heuristic-MILP						
				#cvts	P1	P2	P3	Total	η	#cvts	P1	P2	P3	Total	η	CPU	#cvts	P1	P2	P3	Total	η	CPU
Homo16	56	16	208	16	806	1235	64	2106	81.0	16	773	1024	64	1861	82.5	2.9	16	779	991	64	1834	82.8	360.4
Hetero32	8	8	203	8	160	311	32	503	86.0	8	158	240	32	430	87.8	1.7	8	157	200	32	389	88.8	374.4
Average						1		1				0.80		0.87					0.72		0.82		

the results of manual design, columns 10–16 give the results of the greedy scheme discussed in Section VI-B3, and columns 17–23 show the results of the heuristic approach presented in Section VI-B2. For each approach, we list the total number of converters used, the total power loss (refer to Equation (17)) and its breakdown, P_1, P_2, and P_3, in mW. We also show η, the system-level efficiency of the power delivery system, and CPU, the runtime of these two heuristic approaches in seconds (on a 64-bit 2.5GHz Intel Quad-core platform).

The greedy approach can only find a solution with up to l converters, where l is the number of the cores in the chip (l equals to 16 for the homogeneous chip, and 32 for the heterogeneous chip). Therefore, compared to the manual design it finds worse solution for the homogeneous chip but better solution for the heterogeneous case. In comparison, the heuristic approach based on MILP can perform better than greedy approach: it can reduce P_2 (power loss due to voltage droop) by about 50% and total power loss by 25% compared to the manual design. The system-level efficiency is improved from 86.1% to 88.1% for the homogeneous chip, and from 86.1% to 90.1% for the heterogeneous chip. The runtime of the MILP problem is tractable, it takes only a few minutes for CPLEX to solve these two chips.

As stated before, the manual design and greedy approach have limited search space w.r.t the number of converters, as compared to the MILP-based approach. For a comparison that is more favorable to the limited search space of manual design and greedy approach, and to explore the quality of our approach under stringent constraints, we perform another set of experiments by setting the same upperbound for the available number of converters for these three approaches. The results are presented in Table IV. Column 3 shows the upper bound for number of converters. From the table we can see that compared to manual design, on average, *Greedy* and *Heuristic-MILP* can still improve the results respectively by 13% and 18% in terms of the total power loss. This is because with the same number of converters, the heuristic approaches can search different combinations of the converters. Even for the homogeneous chip, there is still room for improvement because of the unevenly distribution of current within each core and the asymmetry in the power pads shared by different power domains in a single chip.

Fig. 6(a) shows how the power losses P_2, P_3 and the total power loss $P_1 + P_2 + P_3$ change with various number of converters for the homogeneous chip by applying the heuristic approach *Heuristic-MILP*. We can see that as we increase the number of converters from 1 (all the cores connected to a converter) to 30, the power loss P_2 due to voltage droop decreases quickly, with a reduction of more than 20X. This implies that the distributed design of the converters can effectively reduce the IR drop seen by the cores, and therefore, improve the efficiency of the power delivery system. The reduction

in total power loss starts to slow down as we further increase the converter number, and the overhead from the control circuit and clock network begins to dominate the overall power loss. Similar results can be observed for the heterogeneous chip as shown in Fig. 7(a).

Fig. 6(a) shows high power loss (more than 10W) when only a few converters are used. This is because we generated the results with the same wiring resources for different number of converters. The loss number can be reduced by using more interconnect resources through narrowing the pitch of the power grid, but that can cause very high congestion.

For the homogeneous chip, the lowest total power loss is achieved with 47 converters as shown in Fig. 6(b), and the layout is shown in Fig. 5(left). Note that although there is no large difference in the total power loss between the cases using 47 and 56 converters, more routing resource is needed for the clock network when more converters are used, which is not captured by power loss objective function. It is certainly possible to use an enhanced objective that captures this factor, or to determine a reasonable tradeoff by examining the curve. For the heterogeneous chip, the lowest total power loss of is achieved with 13 converters shown in Fig. 7(b), and the layout is shown in Fig. 5(right).

In Section VI, we had proposed heuristic approaches to break the MINLP problem (described in Section V) into two independent sub-problems. In fact, we have another formulation (details not shown due to space limitations) that solves MINLP problem approximately in an iterative way: We start with the initial guess to the MINLP problem provided by the *Heuristic-MILP* and closed-form solution presented in Section VI-C. And we set the integer variables z_is to be the values from the initial guess (i.e., fixing the number and location of the converters).

The iterative process, called *Heuristic-iterative*, consists of two steps:
(1) For fixed z_is, the MINLP problem in Section V-B becomes a NLP, that is solved by CPLEX through sequential linear programming.
(2) We update the number and location of the converters by solving a MILP problem by fixing some variables based on the NLP solution. The key difference between *Heuristic-MILP* and *Heuristic-iterative* is that we allow the converters to have different voltage ripple ΔV_is in *Heuristic-iterative*. Table V presents the results of comparison between *Heuristic-MILP* and *Heuristic-iterative*. We observe that *Heuristic-iterative* can only improves the initial guess provided by *Heuristic-MILP* by a small amount. This implies that our assumption about identical voltage ripple made in Section VI is acceptable in terms of the solution quality.

VIII. CONCLUSION

In this paper, we study the efficiency of the power delivery system using SC converters at the system level. This work develops
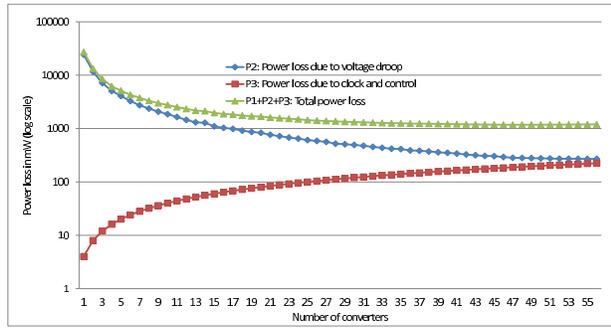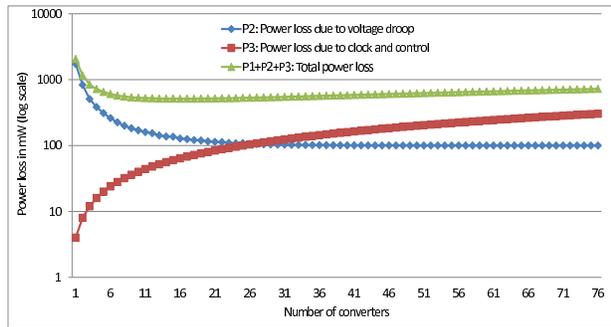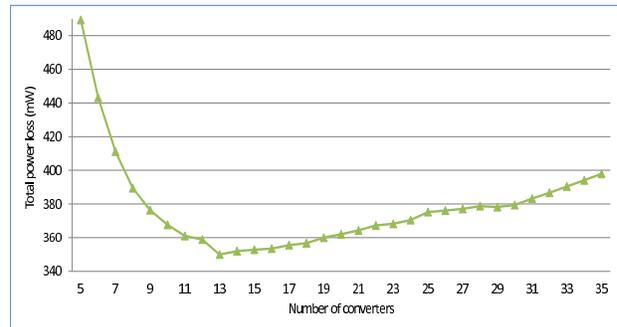
(a) Complete
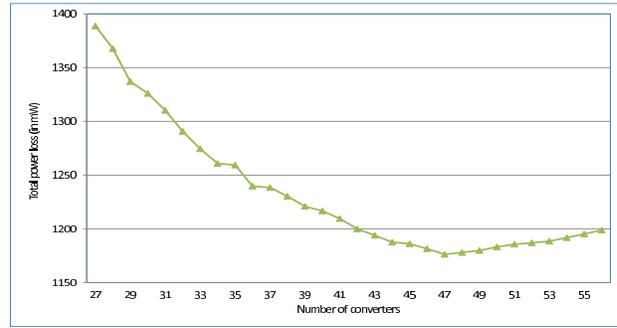
(b) Part

Fig. 6. Power loss vs. # converters for homogeneous chip. The left figure shows the complete graph for P_1, P_2 and the total power loss. The right figure shows part of the total power loss as the number of converters changes from 27 to 56.

(a) Complete

(b) Part

Fig. 7. Power loss vs. # converters for heterogeneous chip. The left figure shows the complete graph for P_1, P_2 and the total power loss. The right figure shows part of the total power loss as the number of converters changes from 5 to 35.

TABLE V
HEURISTIC-MILP vs. HEURISTIC-ITERATIVE

Chip	Heuristic-MILP					Heuristic-iterative				
	#cvts	P1	P2	P3	Total	CPU	#cvts	P1	P2	P3	Total	CPU
Homo16	47	704.9	283.8	188	1176.7	370.1	47	703.6	283.7	188	1175.3	374.9
Hetero32	13	156.7	141.9	52	350.6	362.7	13	156.1	141.7	52	349.8	364.9

models for the efficiency of such a system as a function of size and layout of the SC converters, and the problem is formulated as a mixed integer non-linear program optimization. We then propose heuristic approaches to optimize the size and layout of the SC converter to minimize power loss. The efficiency of these techniques is demonstrated on both homogenous and heterogenous multicore chips. Our current work only considers the deep trench capacitor and in future we would extend our work to deal with other types of capacitors such as CMOS and MIM capacitors, by considering the area effect in exploring the granularity of the converters.

REFERENCES

[1] P. Zhou *et al.*, "Exploration of on-chip switched-capacitor DC-DC converter for multicore processors using a distributed power delivery network," in *CICC*, 2011, pp. 1–4.
[2] H.-P. Le *et al.*, "Design techniques for fully integrated switched-capacitor DC-DC converters," *JSSC*, vol. 46, no. 9, pp. 2120–2131, Sept. 2011.
[3] Y. Ramadass and A. Chandrakasan, "Voltage scalable switched capacitor DC-DC converter for ultra-low-power on-chip applications," in *IEEE Power Electronics Specialists Conference*, 2007, pp. 2353–2359.
[4] Y. Ramadass *et al.*, "A 0.16mm² completely on-chip switched-capacitor DC-DC converter using digital capacitance modulation for LDO replacement in 45nm CMOS," in *ISSCC*, 2010, pp. 208–209.
[5] H.-P. Le *et al.*, "A 32nm fully-integrated reconfigurable switched-capacitor DC-DC converter delivering 0.55 W/mm² at 81% efficiency," in *ISSCC*, 2010, pp. 210–211.
[6] L. Chang *et al.*, "A fully-integrated switched-capacitor 2:1 voltage converter with regulation capability and 90% efficiency at 2.3A/mm²," in *VLSI Symposium*, 2010, pp. 55–56.
[7] Y. K. Ramadass, "Energy processing circuits for low-power applications," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts, 2009.
[8] Z. Zeng *et al.*, "Tradeoff analysis and optimization of power delivery networks with on-chip voltage regulation," in *DAC*, 2010, pp. 831–836.
[9] S. Kudva and R. Harjani, "Fully-integrated on-chip DC-DC converter with a 450x output range," *JSSC*, vol. 46, no. 8, pp. 1940–1951, Aug. 2011.
[10] D. Duarte, N. Vijaykrishnan, and M. Irwin, "A complete phase-locked loop power consumption model," in *DATE*, 2002, p. 1108.
[11] M. Zhao *et al.*, "Hierarchical analysis of power distribution networks," in *DAC*, 2000, pp. 150–155.
[12] M. R. Bussieck and A. Pruessner, "Mixed-integer nonlinear programming," *SIAG/OPT Newsletter: Views & News*, 2003.
[13] "IBM ILOG CPLEX Optimization Studio v.12," available at http://www-01.ibm.com/software/integration/optimization/cplex-optimization-studio/.
[14] S. Bell *et al.*, "Tile64 - processor: A 64-core SoC with mesh interconnect," in *ISSCC*, 2008, pp. 88–598.
[15] S. Vangal *et al.*, "An 80-Tile Sub-100-W TeraFLOPS Processor in 65-nm CMOS," *JSSC*, vol. 43, no. 1, pp. 29–41, Jan. 2008.
[16] "ARM Cortex processors," available at http://arm.com/products/processors/index.php.
[17] ARM Holdings plc, "big.LITTLE Processing," available at http://www.arm.com/products/processors/technologies/biglittleprocessing.php.
[18] "Predictive Technology Model," Device Group at Arizona State University, Available at http://www.eas.asu.edu/~ptm.