# Smart Grid on Chip: Work Load-Balanced On-Chip Power Delivery

Divya Pathak, *Student Member, IEEE*, Houman Homayoun, *Member, IEEE*, and Ioannis Savidis, *Member, IEEE*

***Abstract*—In this paper, a dynamic on-chip power delivery system for chip multiprocessors (CMPs) is proposed, analogous to the smart grid deployed for large-scale energy distribution. The system includes underprovisioned on-chip voltage regulators (VRs) interconnected through a switch network. The peak current rating of the VRs is selected to meet only the average current demand of the cores. A real-time load-balancing algorithm is developed to reconfigure the power delivery network (PDN) by combining the output of multiple VRs when the workload demand exceeds the peak current rating of a single regulator. An operating system level task scheduling heuristic distributes the workloads on the cores such that the required reconfiguration of the PDN is minimized. Simulation results for the proposed power delivery system indicate up to a 44% reduction in the energy consumption of the CMP. In addition, the on-chip footprint of the PDN, including the on-chip VRs and the switching network, is reduced by at least 23%. The proposed cross-layer power management technique is an optimum solution for power-constrained many-core architectures implemented in advanced technology nodes.**

***Index Terms*—Load balancing, on-chip voltage regulation (OCVR), power management, smart grid, work load scheduling.**

## I. Introduction

The power delivery of chip multiprocessors (CMP) faces similar challenges as utility electrical delivery systems: to improve the energy efficiency of the system despite the increasing power demand. Smart grids, with two-way information flow between the consumer and supplier, resolve the imbalance in the demand and supply of electricity to large geographic locations with semipredictable electrical consumption. Despite the increasing power consumption of CMPs utilized for high-performance and exascale computing, the power delivery system for CMPs remains overprovisioned for the worst case power demand. In this paper, a smart, reconfigurable on-chip power delivery network (PDN), is developed with two-way information flow between the processing elements and the on-chip voltage regulation (OCVR) circuits. There are multiple benefits of the proposed solution, including improved energy efficiency and robustness against OCVR failure due to aging or process variation. The main contributions of this paper are as follows.

1) The first paper to apply the concept of smart grid to on-chip power management. The on-chip power distribution is modeled with two-way information flow. The peak current output of the OCVRs supports the average current requirements of the processing elements [1], [2]. In a smart grid, load balancing is achieved through one or all of the following three methods: switch reconfiguration, tie-line addition, and wire upgrade [3]. In the context of on-chip power delivery, an equivalent technique to wire upgrade or tie-line addition is not feasible post-fabrication. In this paper, a high-speed switching (HSS) fabric is proposed to reconfigure the connections between the OCVRs and processing elements to perform dynamic load balancing.
2) Although the on-chip integration of the voltage regulators (VRs) improves the quality of the power delivered through reductions in *IR* drop and $L\mathrm{d}i/\mathrm{d}t$ noise, the parametric and thermal variations, which impact the load circuits, now also influence the OCVRs [4]. In advanced technology nodes, with lower power supply voltages and tighter noise margins, the variation in the regulated voltage provided by the OCVR detrimentally effects and negates the benefits of introducing the voltage regulation circuits on-chip [4]. In this paper, an interconnected power distribution network with OCVRs is developed. The proposed topology is more robust than the conventional radial topology that relies on a single dedicated OCVR for on-chip power distribution. The load circuit is guaranteed regulated power delivery from multiple OCVRs and is, therefore, robust to failure of a single dedicated OCVR.
3) A supply side load-balancing algorithm is developed for dynamic power management. The algorithm is executed on the on-chip power management unit (PMU) and combines the output of the OCVRs to support load currents in excess of the maximum output current supported by the OCVRs. The algorithm is an evolution of the work proposed in [1] for energy-efficient OCVR clustering.
4) A convex energy optimization problem is solved to ensure the reliability of the proposed reconfigurable power delivery system with underprovisioned on-chip VRs. The optimization problem is constrained by the total power budget of the CMP and is limited to the peak current rating of the OCVRs. The feasibility of the solution, determined by solving the optimization problem, is demonstrated through a real-time workload

Manuscript received November 18, 2016; revised March 5, 2017; accepted April 10, 2017. Date of publication June 5, 2017; date of current version August 23, 2017. *(Corresponding author: Divya Pathak.)*

D. Pathak and I. Savidis are with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104 USA (e-mail: divya.pathak@drexel.edu; isavidis@coe.drexel.edu).

H. Homayoun is with the Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030 USA (e-mail: hhomayoun@gmu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2017.2699644

scheduling heuristic. The scheduler is applicable to homogeneous and heterogeneous CMPs.

The rest of this paper is organized as follows: Prior work exploring a reconfigurable PDN (RPDN) is discussed in Section II. The system level simulation and the results of the power consumption profile of multi-application workloads are described in Section III. The proposed power delivery methodology is discussed in Section IV. An energy-efficient workload scheduling heuristic is described in Section V. Simulated results showing the improvement in the figures of merit of the OCVRs and the energy efficiency of the CMP system are provided, respectively, in Sections VI and VII. SPICE simulation of a power grid benchmark to characterize the impact on power supply voltage droop during reconfiguration of the PDN is described in Section VIII. Concluding remarks are provided in Section IX.

## II. Related Work

Recent work has attempted to improve the energy efficiency of multi-core and many-core systems by reconfiguring the PDN dependent on the power demand of the work load. An RPDN using switched capacitor VRs (SCVRs) and cross bar switches to serve eight cores is proposed in [5]. The RPDN consists of 32 cells, where each cell is an SCVR capable of supporting two voltage step down conversions (2:1 and 3:2). The simulation results indicate that the RPDN offers 40% energy savings as compared with a configuration with per core voltage regulation. The SCVRs offer a power conversion efficiency (PCE) of 80%. This paper does not address the inferior voltage regulation offered by the SCVRs. The SCVR topology is selected over a buck converter as it is assumed that scaling the SCVR involves reducing the size of the capacitor and MOS switch for smaller load currents with moderate losses. Multiphase buck converters offer the same advantage with superior voltage regulation. A buck converter-based fully integrated VR (FIVR) [6] offers a PCE of 90% and, therefore, better energy efficiency. The SCVR offers a voltage transition time in the range of 1000 ns. The voltage transition time for the FIVR is 500 ns (0 V to 1 V). The selection of the SCVR is, therefore, not optimal as OCVRs with improved PCE, output voltage regulation, and voltage transition time are available. The area, power, and transient time of the switching network in the RPDN are also not analyzed. Scalability of the power network is achieved by partitioning the RPDN to serve clusters of cores. With a minimum of four cores per cluster (to support the four voltage levels of the cluster), the number of switches required in the RPDN scales as $N \times M^2$, where $M$ is the number of cores in a cluster and $N$ is the number of clusters in the system. Each cluster requires a dedicated power grid. The area overhead and power overhead due to the switching network and the power grid are not analyzed for a many-core platform.

A run-time reconfigurable VR network of buck converters is described in [7]. Applying dynamic voltage and frequency scaling (DVFS), the lowest energy consumption across a range of voltages and frequencies is determined by solving an integer linear programming (ILP) problem. The timing penalty to set the switching network is not quantified, and the ILP is solved for discrete DVFS timing penalties ranging from 5% to 15%. An off-chip buck converter (LTC3816) SPICE model is used instead of an OCVR, although the OCVR offers an order of magnitude faster voltage response time under DVFS [8].

In [9], an on-chip RPDN that includes dynamic voltage and frequency scaling is proposed on a two tier 3-D IC. The cores are grouped according to voltage demand, which is determined from a lookup table comprised of the power envelops (maximum power consumption per time slot) per control cycle and the associated voltage levels obtained from tracking and predicting the power signature using an autoregression algorithm. The power signatures for SPEC CPU2000 benchmarks are obtained through simulations using *Wattch*. Extraction of the power phase is completed through singular value decomposition of the covariance matrix of power signatures per time slot and per workload.

A workload scheduler is developed to minimize the power slack (the difference between a predetermined power threshold and power envelop per time slot). The penalty of MIPS/Watt due to time multiplexing of the connections between the core and VRs is not analyzed in [9]. In addition, an underlying assumption of the workload scheduling algorithm is that in any given time slot, there is always a subgroup of cores available with positive power slack, which does not always hold true. Since the maximum number of cores that are served by a single VR is limited, if the total driving capability of the system for all VRs is less than the total number of cores, no plausible connections between cores and VRs exists that meets the total current demand. In addition, the network switching time is three times the time slot considered by the power management algorithm, which significantly impacts the performance when switching the workload from one subgroup to another. The power loss in the VRs and switching network is not accounted for when determining the power savings of the RPDN proposed in [9].

Clustering of VRs to boost the energy efficiency of the system is proposed in [10] and [11], but this paper ignores the variation in the PCE of the low dropout (LDO) VRs due to dynamic voltage and frequency scaling. As shown in [12], ignoring the variation in the PCE of the VRs leads to suboptimal workload mapping and, therefore, a large penalty on the energy savings possible with DVFS.

Recent work on RPDNs does not provide an analysis of the penalty in the response time of the power delivery system due to the search and decision time needed to flip the requisite number of switches and reconfigure the connections between the cores and the VRs. In addition, all prior work considers overprovisioned VRs designed for worst case power consumption. The RPDN proposed in this paper is novel, as the PDN is designed to supply the typical power demand of the cores, with the capability to support the peak power demand if necessary. The PDN configuration is managed dynamically without the overhead of solving any off-line or online linear or nonlinear programming optimization problem. In addition to the circuit implementation of the RPDN, a task scheduling algorithm is also developed to minimize the reconfiguration of the RPDN.

TABLE I
ARCHITECTURAL PARAMETERS OF THE CORE

| Parameter | Value |
|---|---|
| Core clock frequency | 2.4 GHz |
| Power supply voltage ($V_{dd}$) | 1 V |
| Issue, commit width | 2 |
| INT and FP instruction queue | 16 entries |
| Load and Store Queue | 16 entries |
| INT and FP Physical Register File | 48 entries |
| ROB size | 48 |
| L1 cache | 32KB, 4-way |
| L2 cache | 256KB |

## III. POWER DISSIPATION IN CMPs

The behavior of each application differs with regard to the utilized resources of a computing system. While some applications are CPU-intensive, others are memory-intensive. The power dissipation pattern, therefore, varies across different applications. In addition, applications exhibit differing power dissipation behavior in different execution phases. The power dissipation pattern of workloads must, therefore, be accounted for while designing the power delivery system. The analysis of the typical power consumption profile of different workloads offers insight on the circuit level implementation of the OCVRs that provide regulated power to the CMP system. The simulation methodology to determine the power consumption of SPEC CPU benchmarks and the corresponding simulated results are described, respectively, in Sections III-A and III-B.

### A. Simulation Methodology

A 16-core CMP in a 45-nm technology is modeled using a processor architectural simulator [13]. McPAT [14] is integrated in the simulator to analyze the power consumption of the core. Each core has a two-way issue and out-of-order execution unit. The microarchitectural parameters of the core used in simulations are summarized in Table I.

A set of 49 applications from the SPEC CPU2000 and SPEC CPU2006 benchmark suites are studied to determine the power dissipation behavior of the core. Each benchmark is simulated at four timing intervals to cover execution phases with different power consumption profiles. The simulations are run for 10K cycles per time interval, and the power consumption is sampled cycle by cycle.

### B. Analysis of the Power Dissipation Behavior

The statistical variation of the power consumption per cycle of different SPEC CPU benchmarks with single phase forwarding is shown through a box plot in Fig. 1. The interquartile range for all the studied benchmarks falls approximately an order of magnitude below the peak power consumption of 5.73 W reported through McPAT. The number of outliers beyond 5$\sigma$ coverage for each benchmark is an insignificant fraction of the sample size.

The consolidated power consumption and power variation histograms of the studied benchmarks are shown in Fig. 2. The dashed line in Fig. 2(a) delineates the average power consumption across all benchmarks, which is approximately 0.55 W. The power dissipation of the applications is between 0.3 W and 0.5 W for approximately 65% of the execution time. The studied benchmarks spend more than 78% of the run time consuming less than the average power. The variation in power dissipation between any two clock cycles averages 0.2 W, as shown in Fig. 2(b). The power variation is less than 0.1 W for about 90% of the time. The characterization of the power for the SPEC CPU2000 and SPEC CPU2006 benchmarks is summarized in Table II. The peak power $P_{\text{peak}}$ of 5.73 W reported by McPAT is never consumed. The maximum power consumption of 4.75 W is consumed for a small percentage ($7.5 \times 10^{-5}$%) of the run time across all workloads.

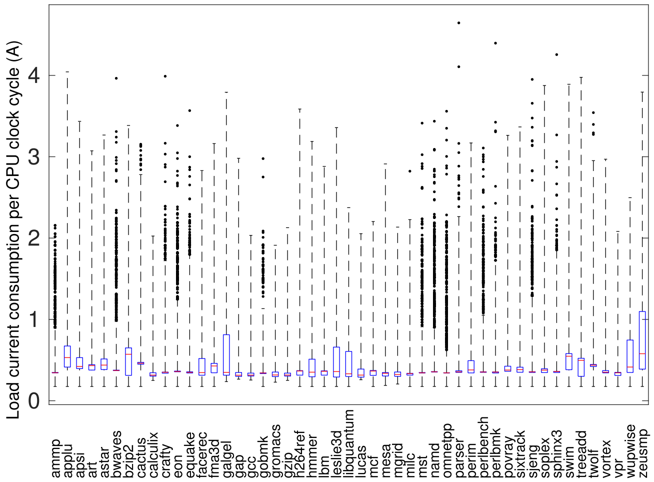
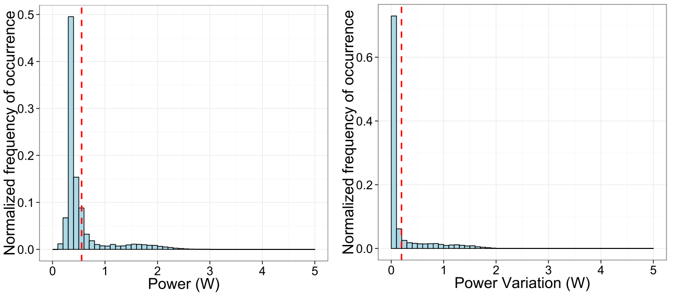


Fig. 1. Statistical analysis of per cycle power consumption of SPEC CPU benchmarks.



(a) (b)

Fig. 2. Histogram of (a) power dissipation and (b) power variation per cycle, for the 49 SPEC CPU2000 and SPEC CPU2006 benchmarks.

TABLE II
COMBINED POWER DISSIPATION CHARACTERISTICS OF SPEC CPU2000 AND SPEC CPU2006 BENCHMARKS

| Parameter | Minimum | Average | Maximum |
|---|---|---|---|
| Power dissipation (W) | 0.175 | 0.555 | 4.755 |
| Power dissipation variation (W) | 0 | 0.195 | 4.333 |

## IV. PROPOSED INTERCONNECTED POWER DELIVERY NETWORK

The power consumption characteristics listed in Table II indicate that the VR rating is overprovisioned for the majority of the run time of the workloads. Significant work has been

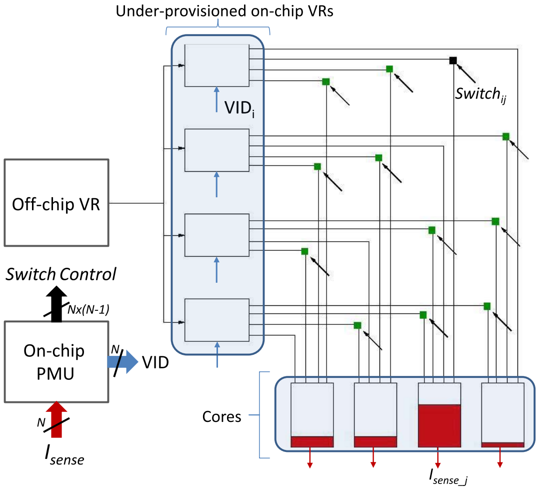


Fig. 3. Proposed interconnected on-chip PDN with run-time VR clustering through a switching fabric.

**Algorithm 1** Load-Balanced Power Delivery With Run-Time OCVR Clustering to Support Higher Than Average Load Current Consumption

```
Inputs:
Current consumption sensed from the core: $I_{sense\_x}$
Current threshhold: $\triangle I$
Voltage level applied to the core: $V_x \in [V_{dd\_1}, V_{dd\_2}, ..., V_{dd\_m}]$
where x ∈ [1,...,N], m = number of DVFS levels, N = number of OCVRs/Cores
Switch matrix: $Switch_{N,N-1}$
Constraints:
$t_{switch} + t_{PMU} < t_{core}$
$\sum_{x=1}^{n} V_x \cdot I_{sense\_x} < \mathrm{N} \cdot V_{dd\_m} \cdot I_{avg}$
Append array CORE_RED with core id x where $I_{sense\_x} \geq I_{avg}$ - $\triangle I$
Append array CORE_GREEN with core id y where $I_{sense\_y} < I_{avg}$ - $\triangle I$
if $length(CORE\_RED) > 0$ then
  call OCVR_CLUSTER ▷ Reconfigure the PDN by clustering the output of OCVRs
else
  call OCVR_DECLUSTER ▷ Reconfigure the PDN by de-clustering the output of OCVRs
procedure OCVR_CLUSTER
  for each i in min(length(CORE_RED), length(CORE_GREEN)) do
    Demand(i) ← CORE_RED($I_{sense\_i}$) + CORE_GREEN($I_{sense\_i}$)
    if Demand(i) $\leq 2 \cdot I_{avg}$ then
      $V_{CORE\_RED(i)} \leftarrow V_{CORE\_GREEN(i)}$ ▷ Align the $V_{dd}$ levels of the two cores whose OCVR outputs are being combined
      $Switch_{CORE\_RED(i),CORE\_GREEN(i)} \leftarrow 1$ ▷ Close the switch so that CORE_RED(i) is served by the OCVR connected to CORE_GREEN(i)
      Delete CORE_RED(i) and CORE_GREEN(i) from the respective arrays
procedure OCVR_DECLUSTER
  for each nonzero element in sparse matrix Switch do
    if $I_{sense\_i} + I_{sense\_j} < 2 \cdot I_{avg}$ - $\triangle I$ then
      $Switch_{i,j} \leftarrow 0$ ▷ Open the switch connecting core i with OCVR j
```

done to optimize the core configuration, work load mapping, and dynamic/static clustering of the cores served by the off-chip VRs, but the energy and area loss incurred due to the integration of overprovisioned VRs has been overlooked. In the proposed PDN, on-chip VRs with a maximum output current equal to the average current consumption $I_{\mathrm{avg}}$ of the load circuits (cores) are considered. A technique to deliver currents higher than $I_{\mathrm{avg}}$ is described in this section.

The block representation of the proposed interconnected PDN is shown in Fig. 3. For a CMP system consisting of $N$ cores, $N$ OCVRs provide the regulated power. As described in [12], providing a single OCVR to an optimally sized cluster of cores yields similar benefits in energy savings as per core DVFS, while also reducing the number of OCVRs as compared with the number of cores. In the proposed PDN, if the OCVR serves a cluster of $n$ cores, the power rating of the OCVR is $\sum_{i=1}^{n} I_{\mathrm{avg},i}$. The output of each OCVR is connected to the inputs of an HSS fabric. The $N$ outputs of the HSS fabric are connected to the local PDN grid of the $N$ cores or core clusters. The HSS fabric is controlled by the PMU. The interconnected PDN shown in Fig. 3 provides increased service reliability as compared with the conventional radial topology used for on-chip power distribution [10]. In addition, the interconnected network provides opportunity to balance the load current through reconfiguration of the switches. The current sensors placed in each core are constantly monitored by the PMU. When the sum of the currents sensed from all cores within a cluster $I_{\mathrm{sense}}$ reaches a threshhold $\triangle I$ below $I_{\mathrm{avg}}$, the PMU configures the HSS to source additional current from the OCVRs that are operating at the same power supply voltage level (under DVS controlled by the PMU).

The high-speed switches are controlled by logic within the PMU that operate on two system parameters, the $V_{\mathrm{dd}}$ levels and the total load current sensed from each core cluster. The analysis of the power consumption of the workloads provided in Section III indicates that the probability of the load current demand exceeding $I_{\mathrm{avg}}$ is 22%. As a result, there are always more than one cluster of cores operating at or below $I_{\mathrm{avg}}$. The PMU is provisioned to add at least one additional OCVR to serve a cluster (or core) requiring current higher than $I_{\mathrm{avg}}$. An available OCVR is ensured if the number of DVFS levels is less than the number of core clusters in the CMP system. The sum of the decision time of the PMU and the time to reconfigure the switches must be less than or equal to the load current transient response time (current slew rate) of an OCVR with a current rating of $I_{\mathrm{peak}}$ to ensure an uninterrupted power supply to the core (or cluster of cores). The switching control of the HSS fabric is described by Algorithm 1.

Algorithm 1 is executed on the on-chip PMU. The inputs to the PMU are the operating power supply voltage and the load currents sensed for each core served by a dedicated OCVR. The PMU controls the $N \times (N-1)$ HSS matrix ($\mathrm{Switch}_{N,N-1}$) to dynamically configure the connections between the $N$ OCVRs and $N$ cores. Two dynamic arrays, *CORE_RED* and *CORE_GREEN*, are maintained with the identification label of the cores, which are consuming, respectively, more than $I_{\mathrm{avg}}$ and less than $I_{\mathrm{avg}} - \triangle I$ currents. If the length of the array *CORE_RED* is nonzero, the *OCVR_CLUSTER* routine is executed, which combines the outputs of the OCVRs with total output current less than two times $I_{\mathrm{avg}}$ (maximum output current of the OCVR for a given core with average load current consumption of $I_{\mathrm{avg}}$). The output voltage of the combined OCVRs is matched before activating the corresponding switches in the HSS matrix. If the dynamic array *CORE_RED* is empty, the routine *OCVR_DECLUSTER* is executed, which deactivates the switches that combined the outputs of the OCVRs. The two constraints in the execution of Algorithm 1 are that the time taken to reconfigure the connections between the OCVRs and cores is less than one core clock cycle ($t_{\mathrm{switch}} + t_{\mathrm{PMU}} < t_{\mathrm{core}}$)

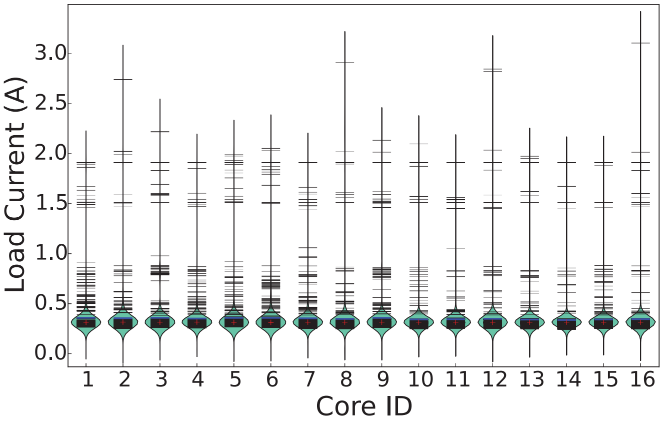


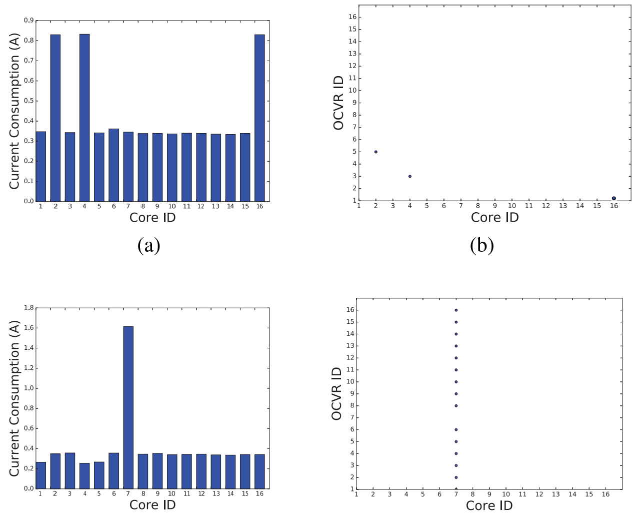
Fig. 4. Load current variation of a 16-core CMP system with 16 OCVRs each with a peak rating of 0.6 A. The stochastic model for load current consumption across the cores is based on SPEC CPU2000 and SPEC CPU2006 benchmark power trace analysis. The variation in load current is used as an input to Algorithm 1.

and the combined power consumption of all the $N$ cores in the system ($\sum_{x=1}^{n} V_x \cdot I_{\text{sense\_}x}$) is less than the total power delivered by the $N$ OCVRs ($N \cdot V_{\text{dd\_}m} \cdot I_{\text{avg}}$). The workload scheduling heuristic described in Section V assures the constraints are met.

Algorithm 1 is implemented in the Python programming language and is analyzed with the parameters summarized in Table III. The per cycle power consumption of different SPEC benchmarks for a finite number of CPU cycles is used as an input to Algorithm 1. In addition, a stochastic model of the current consumption of the cores in a CMP system is developed based on the statistical parameters captured from the per cycle power consumption analysis of the SPEC CPU benchmarks (see Fig. 2). The dynamic power consumption $P_{\text{dynamic}}$ per core is modeled as a Type IV Pearson distribution [15] given by

$$f(x)dx = \left[1 + \left(\frac{x-\lambda}{a}\right)^2\right]^{-m} \times \exp\left[-\nu \cdot \tan^{-1}\left(\frac{x-\lambda}{a}\right)\right] dx. \quad (1)$$

The parameters $m$, $a$, $\nu$, and $\lambda$ are derived from the skewness and kurtosis exhibited by the per cycle power consumption of the SPEC CPU2000 and SPEC CPU2006 benchmarks and are, respectively, 4.145, 2.277, 0.889, and 0.322. The load current consumption as obtained from the stochastic model for 1000 CPU cycles across 16 cores is shown in Fig. 4. The peak current rating of each OCVR is set to 0.6 A, which is one order of magnitude less than the $I_{\text{peak}}$ value obtained through McPAT. The high-speed switch configuration for two randomly chosen time stamps to support the run-time load current variation on each core is shown in Fig. 5. A case where the current consumption of core 7 exceeds 2.5 times the maximum current output of the OCVR is shown in Fig. 5(c). The algorithm clusters the output of the OCVRs available in the CMP system connected to cores demanding less than the maximum current output of a single OCVR. A statistical load current model with normal distribution is also analyzed through Monte Carlo simulations with a maximum possible value of $I_{\text{peak}}$. The frequency of OCVR clustering increases with a normally distributed load current model; however, a developed workload mapping technique, described in Section V, significantly reduces the occurrence of clustering. The four $V_{\text{DD}}$ levels listed in Table III are selected corresponding to the core configuration provided in Table I.



(a) (b)

(c) (d)

Fig. 5. Simulated results of the implementation of Algorithm 1. A 16-core CMP system with 16 OCVRs each with a peak rating of 0.6 A is considered. The load current across cores is shown for two time stamps in (a) and (c) with the corresponding switches that are active to supply current that exceeds 0.6 A shown, respectively, in (b) and (d).

TABLE III
SIMULATION PARAMETERS FOR ALGORITHM 1

| Parameter | Value |
|---|---|
| Number of cores | 8, 16, 32, 64, 128 |
| $V_{DD}$ levels | 0.7 V, 0.8 V, 0.9 V, 1 V |
| OCVR peak current rating ($I_{avg}$) | 0.6 A |
| Current threshold ($\triangle I$) | 0.1 A |
| Execution duration | 10 million CPU cycles |
| Load current variation | Stochastic model with Pearson type IV distribution based on SPEC CPU benchmarks |

## V. ENERGY-EFFICIENT WORK LOAD DISTRIBUTION WITH UNDERPROVISIONED VOLTAGE REGULATORS

Intelligence is introduced into electric grids for large-scale power delivery and distribution through smart communication and metering technologies [16]. The inputs from the smart technologies are used to solve an optimization problem to minimize power losses through demand or supply side load management. The user (residential or industrial) in large-scale smart grids is described as either controllable or noncontrollable load. Demand side load management is feasible in smart grids as controllable loads are scheduled for off-peak hours to reduce the power consumption during peak hours. In the context of CMPs and many-core systems, controllable loads are those tasks that have a soft deadline, such as nonreal-time applications. The rescheduling of controllable tasks reduces the energy consumption of the CMP system for a given scheduling cycle. Noncontrollable loads are tasks that are constrained to a hard or firm deadline, such as real-time applications.

The algorithm developed in Section IV for run-time OCVR clustering is an example of supply side load management. The on-chip PMU executes reconfiguration of the OCVRs to meet the changing load current demands of the cores. In this section, an energy optimization work load scheduling technique is developed to relax the timing and load current constraints imposed on Algorithm 1, which are given, respectively, by

$$t_{\text{switch}} + t_{\text{PMU}} < t_{\text{core}} \tag{2}$$

$$\sum_{x=1}^{N} V_x \cdot I_{\text{sense\_}x} < N \cdot V_{\text{dd\_}m} \cdot I_{\text{avg}}. \tag{3}$$

The proposed workload scheduling technique is similar to demand side load management performed by an OS level scheduler. The workload scheduling technique effectively addresses the timing constraints for real-time workloads with hard deadlines. The workload, CMP, and power consumption models are adapted from the work done in [17] and are described in Sections V-A through V-C.

### A. Workload and CMP Model

The real-time workloads are modeled as a set of independent periodic tasks $\tau_i \in \mathcal{T}$ to be scheduled on a subset of cores of a many-core system $\pi_j \in \Pi$ [17]. Each task $\tau_i$ has a hard deadline of $D_i$. Each core $\pi_j$ supports distinct DVFS levels $V_j \in [V_{\text{dd\_1}}, V_{\text{dd\_2}}, \ldots, V_{\text{dd\_}m}]$ and frequencies $f_j \in [f_1, f_2, \ldots, f_m]$. A task $\tau_i$ with a hard deadline $D_i$ requires at most $C_{i,j}$ cycles to execute on a core $\pi_j$ at the highest supported voltage $V_{\text{dd\_}m}$ and frequency $f_m$. The context switching overhead and overhead due to resource sharing amongst tasks that remain unresolved after task partitioning is included in $C_{i,j}$. The computational capacity required by task $\tau_i$ on core $\pi_j$ is defined as $u_{i,j} = (C_{i,j}/D_i)$. The subset of tasks $T_j$ that are executed on core $\pi_j$, therefore, require a total computational capacity of $U_j = \sum_{\tau_i \in T_j} u_{i,j}$ Hz.

### B. Power Model

The power consumption of a processing element $\pi_j$ is approximated as a function of frequency, similar to the work done in [17]. The power consumed by any processing element is given by (4). The $\kappa * f^\alpha$ and $\beta$ terms in

$$P(f) = \kappa * f^\alpha + \beta. \tag{4}$$

represent, respectively, the dynamic and static power consumption of the cores. The model parameters $\kappa$, $\alpha$, and $\beta$ for the Samsung Exynos A15 and A7 processors [17] are used for the validation of Algorithm 2. The power consumption with frequency using the estimated model parameters is shown in Fig. 6.

**Algorithm 2** Real-Time Workload Partitioning and Scheduling on a Many-Core System With Underprovisioned OCVRs

```
Inputs:
Set of real time tasks: $\mathcal{T}$
Set of $N$ cores in the many-core system: $\Pi$
Outputs: Schedulable taskset ($\Theta_j$) on each core $\pi_j \in \Pi$ with assigned voltage
$V_j \in [V_{dd\_1}, V_{dd\_2}, ..., V_{dd\_m}]$ and frequency $f_j \in [f_1, f_2, ..., f_m]$
procedure PARTITION($\mathcal{T}$, $\Pi$)
  for each $\pi_j \in \Pi$ do
    $\Theta_j \leftarrow \emptyset$, U $\leftarrow$ 0
  $\mathcal{T}^{'} \leftarrow$ SORT($\mathcal{T}$ by descending $max_j\ u_{i,j}$)
  for each $\tau_i \in \mathcal{T}^{'}$ do
    $\Pi^{'} \leftarrow$ j: $U_j + u_{i,j} < f_{m,j}$ ▷ Cores on which $\tau_i$ is schedulable
    if $\Pi^{'} = \emptyset$ then return Failed to schedule
    k $\leftarrow$ arg $min_{j \in \Pi^{'}}\ P_j(U_j+u_{i,j})$ ▷ core id on which $\tau_i$ consumes least power
    $\Theta_k \leftarrow \Theta_k \bigcup \tau_i$ ▷ $\Theta_k$ is the schedulable set of tasks on $\pi_k$
    $U_k \leftarrow U_k + u_{i,j}$
return $\Theta$
procedure DVFS($\Theta$)
  while $\sum_{\pi_j \in \Pi} f_j^\alpha > (P_{total} - N \cdot \beta)/\kappa$ do
    for each $\pi_j \in \Pi$ do
      while $\sum_{\tau_i \in \Theta_j} u_{i,f_j} \leq f_m$ do
        $f_j \leftarrow (f_x \mid f_x \in$ F and $f_x \leq f_j)$ ▷ lower the operating frequency to one
of the supported DVFS levels $F = (f_1, f_2, ..., f_m)$
procedure SCHEDULE($\Theta_j$, $f_j$)
  $k \leftarrow$ arg $min_{\tau_i \in \Theta_j}\ D_i$
  $\pi_j \leftarrow \tau_k$
```

### C. Optimal Workload Scheduling

An optimization problem is defined to partition and schedule real-time workloads on a many-core platform. A specific set of constraints unique to the proposed RPDN are considered, which account for the use of underprovisioned on-chip VRs. The objective of the optimization problem is to minimize the energy consumption of the many-core platform, including the power consumed by the OCVRs. The energy consumed by the system in a scheduling period $T_{\text{epoch}}$ is given by

$$\min_{U_j} \sum_{\pi_j \in \Pi} \frac{P(U_j)}{\text{PCE}_{U_j}} \cdot T_{\text{epoch}}, \tag{5}$$

where P$(U_j)$ is the power consumed by the core $\pi_j$ with computational capacity $U_j$ to execute the scheduled task set, and $\text{PCE}_{U_j}$ is the combined PCE of the OCVR(s) supplying current to the core $\pi_j$. The workload scheduling is constrained by the total computational capacity $U_j$ available to execute the taskset on $\pi_j$, where the total capacity must exceed the computational demand of the taskset

$$\text{s.t.} \sum_{\pi_j \in \Pi} U_j \geq \sum_{\tau_i \in T} u_i. \tag{6}$$

In addition, the operating frequency must fall within the supported frequency range of the cores as given by

$$f_{1,j} \leq U_j \leq f_{m,j} \quad \forall \pi_j \in \Pi. \tag{7}$$

The total power consumed by the cores at any time instant must be less than the combined maximum power supported by all OCVRs in the system as described by

$$\sum_{\pi_j \in \Pi} P(U_j) < N \cdot V_{\text{dd\_}m} \cdot I_{\text{avg}}. \tag{8}$$

### D. Real-Time Workload Scheduling Heuristic

A heuristic is described in this section that performs the real-time workload scheduling on the cores for the optimization problem developed in Section V-C. The heuristic consists of three procedures: *PARTITION*, DVFS, and *SCHEDULE*. The

TABLE IV

PARAMETERS OF THE CMP CORES DERIVED FROM THE SAMSUNG EXYNOS 5410 BIG.LITTLE ARCHITECTURE

| **Parameter** | **big core (A15)** | **LITTLE core (A7)** |
|---|---|---|
| Nominal voltage (V) | 1.16 | 1.225 |
| Nominal frequency (MHz) | 1600 | 1200 |
| OCVR maximum current rating ($I_{avg}$ in mA) | [700 to 1000] | [90 to 120] |
| Power model parameters [$\alpha_j$, $\kappa_j$ (mW/$MHz^3$), $\beta_j$ (mW)] | [2.63, 2.91× $10^{-6}$, 146.49] | [3.28, 1.00× $10^{-8}$, 34.24] |
| Number of cores in homogeneous CMP system | [16] | [0] |
| Number of cores in heterogeneous CMP system | [4] | [4] |

*PARTITION* procedure is an evolution of the marginal-power heuristic (M-PWR) developed in [17]. Optimal workload partitioning is achieved by incrementing the load on each core such that the constraint given by (7) is not violated. The tasks $\tau_i \in \mathcal{T}$ are first sorted in decreasing order of the maximum computational demand $u_{i,j}$ on core $\pi_j \in \Pi$. A task is assigned to a core if the scheduling of the task results in the least increase in the power consumption. The output from the procedure is a scheduled taskset $\Theta_j$ on each core.

The DVFS procedure reduces the operating frequency and the voltage of the cores until the constraint given by (8) is satisfied. The right-hand side of (8) is a constant value equal to the total power $P_{\text{total}}$ of the CMP. Expressing the total power consumed by the cores with the power model given by (4) in (8) provides a limit to the operating frequency of the cores raised to the power $\alpha$, as given by (10). The use of the DVFS procedure results in the optimal frequency of operation for each core by solving the bounded knapsack problem. The deadline of each task in the taskset $\Theta_j$ is analogous to the value of the item in the knapsack. The required computational demand at a given frequency $f_j$ on processor $\pi_j$ is $u_{i,f_j}$. The weight added to the knapsack is analogous to $u_{i,f_j}$. The objective of the knapsack problem is to maximize the number of tasks executed on a core without violating the task deadline. The procedure lowers the operating frequency of each task until constraints

$$\sum_{\tau_i \in \Theta_j} u_{i,f_j} \leq f_m \tag{9}$$

$$\sum_{\pi_j \in \Pi} f_j^{\alpha} \leq (P_{\text{total}} - N \cdot \beta)/k \tag{10}$$

are satisfied. Once the operating frequency of each task in $\Theta_j$ is determined, the *SCHEDULE* procedure schedules the task sets on each core based on an earliest deadline first policy.

### *E. Evaluation of the Real-Time Workload Scheduling Heuristic*

The expected task scheduling from the execution of Algorithm 2 is analyzed for both homogeneous and heterogeneous CMP platforms. The CMP includes processing elements based on models of the ARM A15 and A7 cores integrated in the Samsung Exynos 5410 platform [18]. The parameters used in constructing a 16 core homogeneous and an eight core heterogeneous CMP are listed in Table IV. The DVFS levels applied to the cores are listed in Table V. The variation in the power consumption of the core with frequency, based on the power model given by (4) and validated in [17], is shown in Fig. 6.

TABLE V

VOLTAGE AND FREQUENCY PAIRS USED BY THE DFVS PROCEDURE IN ALGORITHM 2. THE NOMINAL VOLTAGE AND FREQUENCY OF EACH CORE ARE LISTED IN BOLD

| **big core (A15)** | | **LITTLE core (A7)** | |
|---|---|---|---|
| Frequency (MHz) | Voltage (V) | Frequency (MHz) | Voltage (V) |
| 1800 | 1.250 | **1200** | **1.225** |
| 1700 | 1.200 | 1100 | 1.125 |
| **1600** | **1.162** | 1000 | 1.100 |
| 1500 | 1.137 | 900 | 1.037 |
| 1400 | 1.100 | 800 | 0.987 |
| 1300 | 1.062 | 700 | 0.950 |
| 1200 | 1.025 | 600 | 0.950 |
| 1100 | 1.000 | 500 | 0.950 |
| 1000 | 0.962 | 400 | 0.950 |
| 900 | 0.925 | 300 | 0.950 |
| 800 | 0.900 | 200 | 0.950 |

TABLE VI

PARAMETERS TO GENERATE REAL-TIME PERIODIC TASKSETS

| **Parameter** | **Value** |
|---|---|
| Number of tasks ($N_t$) | [32, 48, 64, 80] |
| Task utilization range | [0.1 to 0.9] |
| Task period range in seconds ($T_i$) | [10 to 100] |
| Taskset utilization factor ($\rho$) | [0.1 to 1] |

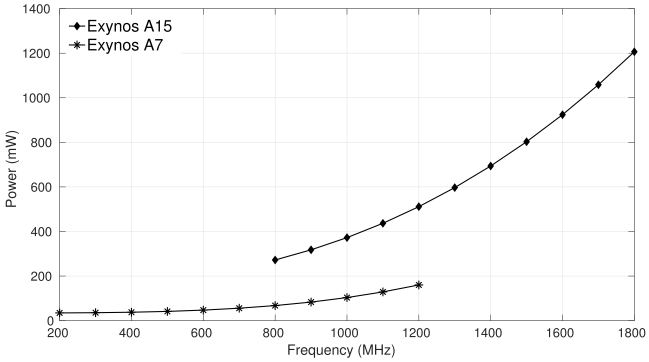


Fig. 6. Power consumption of the Exynos big.LITTLE cores with frequency based on the model given by (4). The parameters applied to the power model are validated in [17].

Real-time periodic tasks with implicit deadlines are considered to analyze the proposed task scheduler. The task scheduling is performed for one hyperperiod $T_{\text{epoch}}$ of the taskset, which is the least common multiple of the implicit deadlines of all tasks $\tau_i \in \mathcal{T}$. The tasks are generated with the parameters listed in Table VI. The computational capacity $u_{ij}$ of the tasks is selected as a random variable with a uniform distribution between 0.1 times and 0.9 times the maximum supported operating frequency of the cores in the CMP (maximum frequency $f_m$ of 1800 MHz). The total computational time requested by the taskset in a hyperperiod is less than the available time on the processing elements to prevent system overload. This ensures that the taskset utilization factor or the

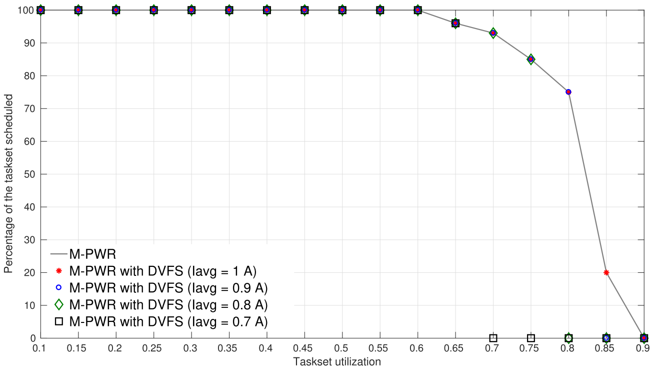
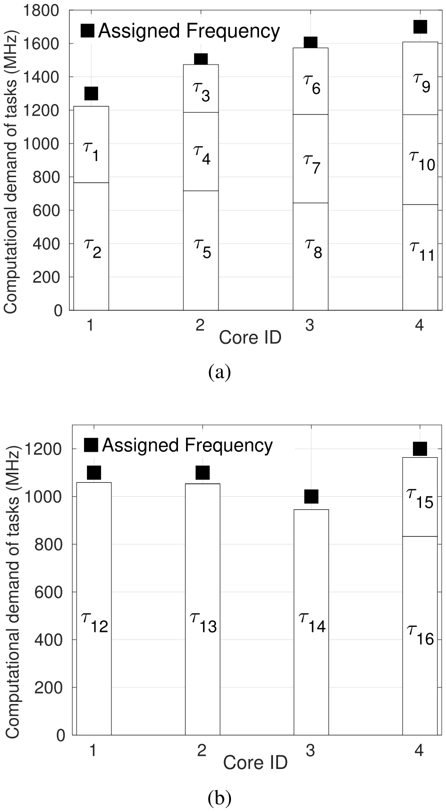


Fig. 7. Percentage of tasks successfully partitioned by the M-PWR heuristic [17] and successfully scheduled by Algorithm 2. Task scheduling is shown for four maximum output current ratings of the VR in a 16 core homogeneous CMP (modeled as an Exynos A15 processor).

system load is less than 1 ($\rho < 1$).

The resulting task schedule, from execution of Algorithm 2 on a homogeneous CMP platform with 16 cores configured as Exynos 5410 A15s, is shown in Fig. 7. The task scheduling is constrained due to the limited power budget of the underprovisioned VRs. For a maximum output current $I_{\text{avg}}$ of 1 A, the percentage of tasks scheduled by Algorithm 2 is identical to the M-PWR heuristic in [17]. The execution of Algorithm 2 is further characterized on a homogeneous platform with VRs of varying maximum output current $I_{\text{avg}}$ rating. For a VR designed with a maximum output current $I_{\text{avg}}$ of 0.7 A, the percentage of tasks scheduled matches the M-PWR heuristic [17] upto a taskset utilization factor $\rho$ of 0.65.

The workload scheduler is also evaluated on a heterogeneous CMP platform with four Exynos A15 (big) and four Exynos A7 (LITTLE) cores. For a randomly chosen taskset hyperperiod, the task distribution and corresponding computational demand ($u_{i,j}$ of each task) is shown in Fig. 8. There are 11 tasks assigned to the big core cluster and five to the LITTLE core cluster. The maximum output current of the VRs serving each of the big cores is set to 800 mA and the LITTLE cores to 110 mA. The frequency assigned to each core to meet the constraint given by (10) is determined and shown in Fig. 8. Depending on the total computational demand of the tasks assigned to each core, the frequency is lowered from the maximum supported frequency of 1800 MHz for the big cores and 1200 MHz for the LITTLE cores. The task partitioning performed by the *PARTITION* procedure further improves power efficiency by preferentially assigning tasks to the little cores, which meet the task utilization constraint given by (6). Consequently, for an identical scaling factor of the peak output current of the OCVRs ($I_{\text{avg}}/I_{\text{peak}}$) serving the LITTLE core and the big core, the percentage of tasks scheduled through the DVFS procedure is lower for the LITTLE cores as compared to the big cores. As the LITTLE core cluster has a load current range of 100 mA, the scaling factor of the maximum output current of the VRs serving the LITTLE cores is set to a larger value than that for the big cores to achieve a high task scheduling rate on the heterogeneous platform. The task scheduling results on the hardware platform demonstrate that the proposed workload scheduler, in conjunction with the execution of Algorithm 1 on the PMU, offer an efficient and robust cross-layer energy optimization mechanism for CMPs with underprovisioned on-chip VRs.



(a)

(b)

Fig. 8. Snapshot of the task assignment on a heterogeneous CMP platform with (a) big cores modeled on A15 parameters, and (b) little cores modeled on A7 parameters. The maximum output currents of the VRs serving each of the big cores and little cores are, respectively, 800 and 110 mA.

## VI. Improvement in Figures of Merit of OCVRs and Other System Parameters With the Proposed Reconfigurable PDN

The increasing power density of CMPs requires efficient OCVRs with a small area and a fast slew rate. The figures of merit of the OCVR depend on the circuit topology and the designed operating point (regulated output voltage at the maximum output current). The conventional choice of dc–dc voltage converter topologies includes buck converters, LDO VRs, or SCVRs. The LDO offers a high PCE with the smallest foot print. However, the PCE of the LDO varies linearly with the $V_{\text{out}}/V_{\text{in}}$ ratio [19], [20]. Considerable power is, therefore, lost in the LDO at lower operating voltages when applying DVFS. The SCVRs are capable of stepping the input voltage up or down based on the configuration of switches and capacitors. The power loss in the resistive switches due to conduction and frequent switching limit the use of SCVRs in on-chip voltage conversion. The output voltage regulation of the SCVR topology is inferior to other VR topologies due to: 1) the strong output voltage dependence on the current demand of the load circuit and 2) the feedback for output voltage regulation is difficult to implement. The output voltage regulation is improved by increasing the switching frequency of the MOS

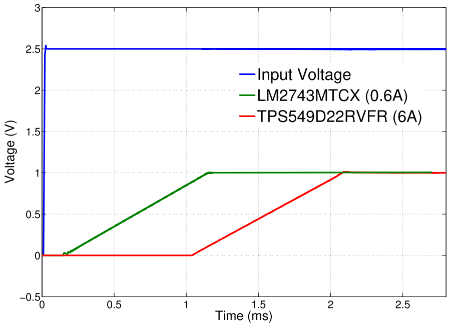


Fig. 9. Startup time of the VRs designed with a peak load current rating of 6 A ($I_{\text{peak}}$) and 0.6 A ($I_{\text{avg}}$).

switches, which requires a wider width and, therefore, results in higher dynamic power consumption (producing a lower PCE). Due to an inferior voltage regulation and PCE, switched capacitor voltage converters are not considered in this paper. The switching dc–dc buck converter offers superior power supply voltage regulation and wider output voltage range with a comparable PCE to LDOs. A buck converter consists of a switching network and a passive low pass filter. The inductor in the low pass filter acts as a low-loss energy transfer device that improves the PCE. Buck converters are, therefore, an optimum choice to power processing elements that implement DVFS. After analyzing each OCVR topology and accounting for the proposed RPDN, the buck converter is selected for integration with the CMP system considered in this paper.

### A. Improvement in Figures of Merit of a Buck Converter

The figures of merit of a buck converter are analyzed for changes to the peak load current rating of the regulator. The goal is to characterize the impact on the OCVR response time and energy efficiency when designing the OCVRs to support only the average load current demand of the cores. The switching dc–dc buck converters with an input voltage of 2.5 V and an output voltage of 1 V are simulated using TI Webench [21]. Different buck converter configurations are simulated with varying maximum output current or peak current rating.

The improvement in the figures of merit of a buck converter with peak current rating of $I_{\text{avg}}$ as compared to $I_{\text{peak}}$ are summarized in Table VII. The capability to maintain a constant output voltage with changes in the load current is described by $\beta_{\text{load}}$ and with changes in the input voltage by $\beta_{\text{line}}$ [19]. The time taken to stabilize the output voltage to the desired regulated value is given by $T_{\text{startup}}$. The startup times of the buck converter for a peak rating of $I_{\text{avg}}$ and $I_{\text{peak}}$ [21] are shown in Fig. 9.

### B. Improvement in Power Conversion Efficiency of a Buck Converter

The power consumed by the buck converter $P_{\text{buck}}$ is given by [19]

$$P_{\text{buck}} = P_{\text{mos}} + P_{\text{ind}} + P_{\text{cap}} + P_{\text{pwm}}. \tag{11}$$

The $P_{\text{mos}}$, $P_{\text{ind}}$, $P_{\text{cap}}$, and $P_{\text{pwm}}$ are the power loss in, respectively, the MOS power transistors and the cascaded buffers driving them, the inductor and capacitor of the filter circuit, and the pulsewidth modulator circuit. The detailed mathematical formulae of each of the components that contribute to $P_{\text{buck}}$ are given in [19] and [22].

TABLE VII

IMPROVEMENT IN THE FIGURES OF MERIT OF THE OCVR SUPPORTING $I_{\text{avg}}$ INSTEAD OF $I_{\text{peak}}$ [19]

| | |
|---|---|
| Area reduction | 0.16 x $Area$ |
| PCE | 1.3 x $PCE$ |
| Output voltage ripple reduction | 0.1 x $V_{ripple}$ |
| Improvement in load regulation | 10 x $\beta_{load}$ |
| Line regulation | 1 x $\beta_{line}$ |
| Reduction in startup time | 0.5 x $T_{startup}$ |

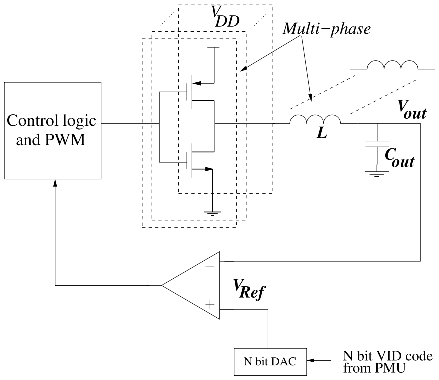


Fig. 10. Schematic of a multiphase dc–dc switching buck converter.

The power consumed by the filter circuit, power transistors, and the buffers driving the power transistors increases with the maximum supported output current of the buck converter. Alternatively, multiple phases are used to drive higher output currents. The circuit schematic of a buck converter with multiple phases of the filter circuit, MOS power transistors, and cascaded buffers is shown in Fig. 10.

Two custom buck converters with maximum output current ratings of 6 A and 0.6 A are implemented [21]. The two converters represent VRs that support the $I_{\text{peak}}$ and $I_{\text{avg}}$ currents of a CMP with core parameters listed in Table I. The size of the filter inductor is chosen such that the percentage of peak current ripple $I_{\text{pp},L}$ remains the same even with an order of magnitude reduction in the peak current drive. The theoretical calculations of an approximate one-third reduction in $P_{\text{buck}}$ when using the smaller converter are verified through simulation. The power consumption of the various components of the buck converter along with the occupied area are listed in Table VIII. The on-chip implementation of the two buck converters yields similar ratios between the power consumed by each component, although at a higher switching frequency for the smaller regulator to reduce the size of the filter inductor and capacitor.

The large reduction in the power dissipation of the buck converter due to a reduction in the peak load current rating results in an improvement in the PCE. Although the overprovisioned buck converter offers the same peak PCE at an output current of 6 A as the underprovisioned buck converter at a current of 0.6 A, the reduction in the PCE with

TABLE VIII
POWER CONSUMPTION OF DC–DC SWITCHING BUCK CONVERTERS DESIGNED FOR 0.6 A AND 6 A PEAK LOAD CURRENTS

| Maximum load current | $I_{pp,L}$ | Switching frequency | PWM duty cycle | $P_{mos}$ | $P_{ind}$ | $P_{cap}$+$P_{pwm}$ | $P_{buck}$ | Foot-print |
|---|---|---|---|---|---|---|---|---|
| 6 A | 2.4 A | 695 KHz | 20.90% | 319.97 mW | 315 mW | 242.13 mW | 889.71 mW | 273 $mm^2$ |
| 0.6 A | 0.24 A | 3 MHz | 26.84% | 197.91 mW | 43.56 mW | 58.97 mW | 300.44 mW | 63 $mm^2$ |

decreasing output current for the overprovisioned converter is significant (see Fig. 11). The typical workloads executed on the CMP, with the core configuration listed in Table I, consume currents less than $I_{\text{avg}}$ for 70% of the execution time. The buck converter with an output rating of 0.6 A, therefore, offers a higher average PCE for a majority of the run-time of the workloads. Circuit techniques like automatic mode switching between pulse frequency modulation (PFM) and pulsewidth modulation (PWM) are used to boost the light load efficiency of buck converters. However, PFM mode in buck converters suffers from inferior transient response to changes in load current as compared to PWM. In addition, the output voltage ripple is greater in PFM mode as compared to PWM mode. The limitation in transient response and output voltage ripple deteriorate the line and load regulation offered by the OCVR. Despite the disadvantages of enhancing the light load efficiency of buck converters, in Section VII, buck converters with enhanced light load efficiency are simulated to analyze the impact on the energy efficiency of a CMP system.

## VII. ENERGY EFFICIENCY OF THE CMP SYSTEM WITH RPDN AND UNDERPROVISIONED VOLTAGE REGULATORS

The total energy consumption of a CMP implemented with a conventional PDN for a given execution time $T_{\text{epoch}}$ with $N$ cores and $N$ OCVRs is given by

$$E_{\text{CMP,conventional}} = \left\{ \sum_{i=1}^{N} \frac{(P_{\text{dynamic\_}i} + P_{\text{static\_}i})}{\text{PCE}_1} \right\} \cdot T_{\text{epoch}}. \tag{12}$$

The cores are served by overprovisioned OCVRs identical to the buck converter with a maximum output current of 6 A. The dynamic and static power consumed by the cores in the presence of DVFS are given by $P_{\text{dynamic}}$ and $P_{\text{static}}$, respectively. $\text{PCE}_1$ represents the PCE of the overprovisioned OCVR. At low load currents close to $I_{\text{avg}}$, the $\text{PCE}_1$ offered by the overprovisioned buck converter is 80%. Alternatively, if the power delivery system is designed with each core supported by a buck converter that supplies a maximum output current of 0.6 A, the achieved $\text{PCE}_2$ at $I_{\text{avg}}$ is 96.36%. In addition, the static power consumed by the cores or core clusters is zero as the idle core(s) are power gated through the HSS fabric. The HSS fabric, however, imposes an additional switching loss $P_{\text{switch}}$, which is the dynamic power consumed by the PMOS transistors while switching, and a conduction loss $P_{\text{conduction}}$ while in the ON state and passing the average current $I_{\text{avg}}$.

The total energy consumed by the CMP with $N$ OCVRs, where each OCVR is designed for an $I_{\text{avg}}$ rating, and $N$x($N$-$1$) PMOS switches is given by

$$\begin{aligned} E_{\text{CMP,proposed}} &= \left\{ \sum_{i=1}^{j} \frac{(P_{\text{dynamic\_i}} + P_{\text{static}})}{\text{PCE}_2} \right. \\ &\quad + \sum_{i=1}^{k} \frac{(P_{\text{dynamic\_}i} + P_{\text{static}} + P_{\text{switch}} + P_{\text{conduction}})}{\text{PCE}_2} \\ &\quad \left. + \sum_{i=1}^{l} I_{\text{quiscent}} \cdot V_{\text{out}} \right\} \cdot T_{\text{epoch}}, \quad j + k + l = N. \end{aligned} \tag{13}$$

The parameters $j$, $k$, and $l$ are, respectively, the number of active core(s) consuming current below $I_{\text{avg}}$, the number of active core(s) consuming current above $I_{\text{avg}}$, and the number of idle core(s) power gated through the HSS network. In the case of idle cores, the power consumed by the OCVRs ($I_{\text{quiscent}} \cdot V_{\text{out}}$) is the only component contributing to the system energy. As described in Section III, applications consume current less than $I_{\text{avg}}$ for approximately 78% of the time. The $P_{\text{switch}}$ loss is incurred for 22% of the execution time of the workloads when the load current exceeds $I_{\text{avg}}$.

TABLE IX
PARAMETERS OF THE PDN DETERMINED THROUGH SPICE SIMULATION

| Parameter | Value |
|---|---|
| PMOS Width | 800 μm |
| PMOS switching time | 160 ps |
| Area occupied by 16x15 switching network | 9600 $\mu m^2$ |
| Per core maximum switching capacitance | 2.4 nF |
| Piecewise constant load current variation | Stochastic model |
| $P_{switch}$ | 0.1 W |
| $P_{conduction}$ | 0.06 W |
| $P_{static}$ (obtained through McPAT) | 0.175 W |

Circuit simulations of a PDN designed to support $I_{\text{avg}}$ for each of the 16 cores are performed to determine the energy consumption as given by (13). The 16 cores are simulated as piecewise constant current sinks. The current variation for the 16 current sinks is shown in Fig. 4 for 1000 consecutive clock cycles. A 16 × 15 pMOS switching network is implemented in a 45-nm technology. The gates of the PMOS switches are controlled through time varying voltage signals. The $P_{\text{switch}}$ and $P_{\text{conduction}}$ for a PMOS switch with an output capacitance provided by a single core is determined through SPICE simulations. The $P_{\text{static}}$ and $P_{\text{dynamic}}$ of a core is measured through McPAT. The $P_{\text{dynamic}}$ for each core is overestimated as the power consumption per clock cycle is characterized at the highest supported $V_{\text{DD}}$ level of 1 V. The parameter values of the switching network and the PDN are summarized in Table IX.

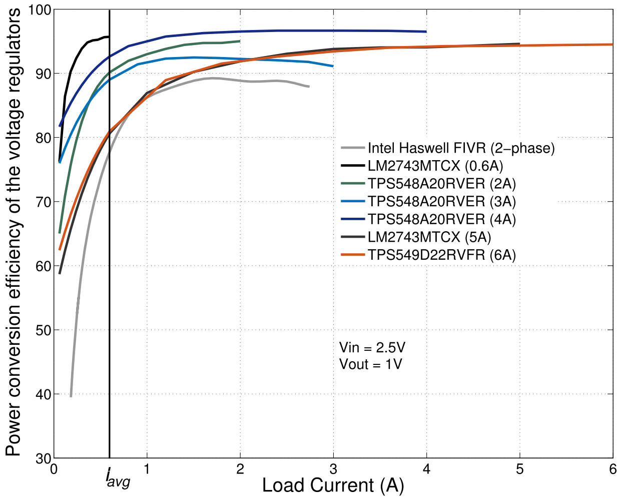


Fig. 11. Power conversion efficiency of the dc–dc switching buck converters simulated with parameters listed in Table X. The model numbers of the devices used for simulation are specified in the figure [21]. The PCE of the two-phase FIVR [6] is provided for comparison with the simulated buck converters with enhanced light load efficiency.

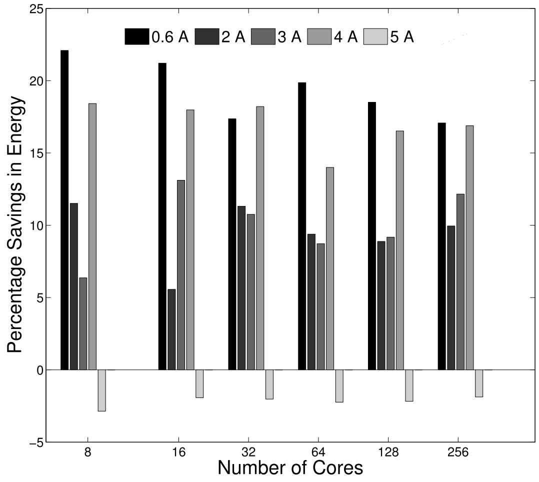


Fig. 12. Percentage energy saved with the proposed RPDN as compared with a baseline OCVR configuration of 6 A with enhanced light load efficiency. The energy savings are calculated for the number of cores in the CMP and varying VR peak current rating for the dc–dc switching buck converters summarized in Table X.

The additional switching and conduction loss due to the PMOS switches is an insignificant fraction of the total power consumed by the CMP as the losses only occur when the PDN is reconfigured to combine the outputs of the OCVRs. The energy consumption of the proposed power delivery system is up to 44% less than the energy consumed by the CMP with overprovisioned OCVRs. The reduction in energy is primarily due to the optimal PCE offered by the buck converter at the maximum output current supported.

The energy savings with the proposed PDN when compared to a conventional PDN with overprovisioned buck converters designed for light load efficiency is also evaluated. Buck converters are designed placing the highest priority to the PCE and improved light load efficiency. The footprint and other figures of merit of the buck converters are, therefore, suboptimal. The simulated designs are summarized in Table X and the PCE variation with load current is shown in Fig. 11.

The energy consumption is characterized for 8, 16, 32, 64, 128, and 256 core CMP systems and OCVR peak current ratings of 0.6, 2, 3, 4, 5, and 6 A. The core configuration is listed in Table I. The static $P_{\text{static}}$, switching $P_{\text{switch}}$, and conduction $P_{\text{conduction}}$ power losses are listed in Table IX. The dynamic power consumption $P_{\text{dynamic}}$ per core is statistically modeled as a Type IV Pearson distribution given by (1) [15]. The energy consumption of the CMP is computed with (13) for a given number of cores and OCVR peak current rating. The percentage savings in energy in comparison to a conventional overprovisioned CMP for different core counts and OCVR ratings is shown in Fig. 12. The larger energy consumption of the CMP system with a peak OCVR rating of 5 A as compared to the 6 A rated OCVR is due to higher reconfiguration of the RPDN, which leads to increased switching and conduction losses in the HSS network. The energy savings for a given OCVR rating do not change significantly with the scaling of the number of cores in the CMP system since the dynamic power consumption of the cores is stochastic and the ratio of OCVRs clustered or declustered at run time with respect to the total number of OCVRs in the system remains constant. The ratio of the number of cores consuming current below $I_{\text{avg}}$, above $I_{\text{avg}}$, and idle (the j:k:l ratio) does not change significantly.

Simulations with a stochastic model of the dynamic current indicate that the proposed RPDN offers an average reduction of 15% in the energy consumption as compared to a CMP with overprovisioned OCVRs designed with enhanced light load efficiency. With DVFS, the reduction in consumed energy will increase proportionally with the scaled voltage and frequency.

## VIII. Transient Analysis of the Proposed RPDN

An RPDN for integrated circuits has been proposed in [23] using microelectromechanical (MEM) switches. The topology proposed in [23] is suitable for power gating certain sections of the PDN but does not provide a fast reconfiguration of the PDN as the switching speed of the MEM switches is three orders of magnitude slower than high performance CMOS switches. Nanoelectromechanical switches currently offer a switching speed in the range of 10 to 20 ns [24]. However, for the HSS network proposed for the load balanced RPDN, NEM switches are still an order of magnitude slower than MOS switches. MOS switches, however, introduce finite on-state resistance and off-state leakage. As shown in Section VII, despite the static and dynamic power loss of the MOS switches, the energy efficiency of the multi-core system improves by implementing the proposed RPDN with underprovisioned OCVRs. In this section, the impact of the MOS switches on power supply voltage droop when the outputs of two or more OCVRs are combined to provide higher than average current is analyzed.

The PDN selected for characterization of the power supply voltage droop is the two metal layer *ibmpg1t* time domain analysis benchmark released by IBM [26] as part of the 2012 TAU Power Grid Simulation Contest [25]. The configuration of the PDN is summarized in Table XI. The

TABLE X

PROPERTIES OF THE SWITCHING DC–DC BUCK CONVERTERS SIMULATED FOR MAXIMUM POWER CONVERSION EFFICIENCY [21]

| Device | LM2743MTCX | TPS548A20RVER | TPS548A20RVER | TPS548A20RVER | LM2743MTCX | TPS549D22RVFR |
|---|---|---|---|---|---|---|
| Maximum output current (A) | 0.6 | 2 | 3 | 4 | 5 | 6 |
| Efficiency at maximum output current | 96.36 | 94.83 | 91.3 | 96.6 | 94.6 | 94.44 |
| High efficiency at light load feature | No | No | No | Yes | Yes | Yes |
| Peak-to-peak inductor ripple current (A) | 0.18 | 0.55 | 0.79 | 1.26 | 0.83 | 1.22 |
| Switching frequency (kHz) | 100 | 234.2 | 231.1 | 100 | 328.13 | 327.5 |
| Duty cycle (%) | 40.16 | 40.64 | 42.6 | 40.3 | 40.77 | 41 |
| Peak-to-peak output ripple voltage (mV) | 2.28 | 1.83 | 2.69 | 6.3 | 1.15 | 1.69 |
| Total power dissipation (mW) | 22.65 | 109.1 | 285.3 | 141.3 | 285.85 | 353.24 |

TABLE XI

IBM POWER GRID BENCHMARK FOR TRANSIENT ANALYSIS (*ibmpg1t*) [25]. THE BENCHMARK INCLUDES THE NUMBER OF CURRENT SOURCES *I*, NODES *N*, RESISTORS *R*, INDUCTORS *L*, CAPACITORS *C*, SHORTS *S* (ZERO VALUE RESISTORS AND VOLTAGE SOURCES), AND CONNECTIONS *V* TO THE 1.8-V OFF-CHIP VOLTAGE SOURCE THAT ARE REPLACED BY A VERILOG-A MODEL OF THE OCVR WITH *V* CONNECTIONS TO THE ON-CHIP PDN

| #I | #N | #R | #L | #C | #S | #V |
|---|---|---|---|---|---|---|
| 10774 | 39681 | 40801 | 277 | 10774 | 14208 | 100 |

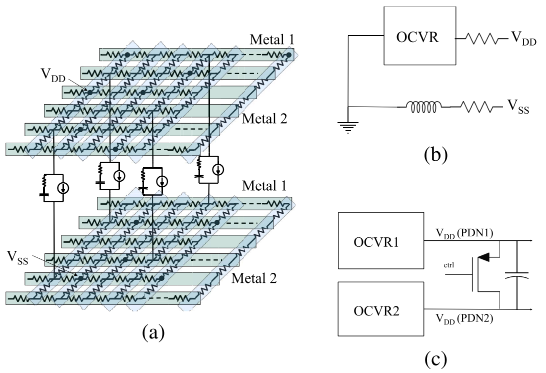


Fig. 13. Structure of the RPDN for transient analysis including (a) schematic of a section of the PDN based on the IBM Power Grid Benchmark *ibmpg1t* with interdigitated $V_{\mathrm{DD}}$ and $V_{SS}$ nets shown on different planes for the same metal layers, (b) schematic representation of the Verilog-A model of the current limited OCVR connected to the $V_{\mathrm{DD}}$ nets of the modified *ibmpg1t* PDN, and (c) outputs of two Verilog-A models of the current limited OCVR connected through a PMOS device from an IBM 180 nm technology.

metal lines of the power grid are modeled as resistive elements. The C4 bumps are modeled as inductances of 1 nH in series with a resistance of 0.25 Ω. A spatially distributed current load is modeled through SPICE current pulse sources. Decoupling capacitors along with the effective series resistance of the capacitors are modeled proportional to the pulsed current load [26]. The typical period of the pulsed current sources is 3 ns, and the peak current drawn is 25 A. The passive elements (*R*, *L*, and *C*) and the structure of the *ibmpg1t* benchmark PDN is not modified for the proposed RPDN with run-time OCVR clustering.

A Verilog-A model of the current limited VR is developed with a maximum current rating of 15 A. The VR model includes load current sense and voltage droop control, which are features of commercial buck converters [27]. The output node of the Verilog-A model is connected to 100 voltage source nets across the *ibmpg1t* grid. The connections to C4 bumps are removed as the Verilog-A model emulates an on-chip VR. The schematic representation of the SPICE model of the *ibmpg1t* PDN and the Verilog-A model of the OCVR connected to the PDN are shown, respectively, in Fig. 13(a) and (b).

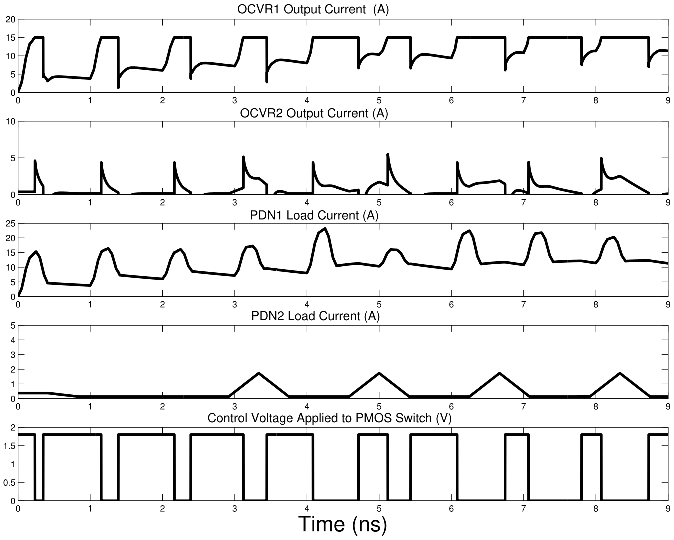


Fig. 14. Transient analysis of two power distribution networks (modified *ibmpg1t* power grid benchmark) with a peak load current consumption of 25 A and 2 A. Each PDN is served by a dedicated OCVR with a maximum load current rating of 15 A.

Two PDNs that include OCVR models are connected through a PMOS switch using an IBM 180 nm technology, as shown in Fig. 13(c). The 180 nm technology node was chosen to match the 1.8 V power supply voltage ($V_{\mathrm{DD}}$) of the *ibmpg1t* benchmark PDN. A decoupling capacitor is added in parallel to the PMOS switch to reduce the impact of the current transient generated when the switch is turned on. The total decoupling capacitance for the original *ibmpg1t* benchmark PDN is 1.63 μF and is 1.79 μF for the proposed PDN with PMOS switches, an approximate 10% increase. Instead of a single PMOS transistor per OCVR, multiple PMOS switches are connected to voltage source nets on the *ibmpg1t* grid along with a corresponding decoupling capacitor to suppress the inrush current. The peak load current demand on PDN2 is set to 2 A to provide time intervals when the output currents of the two OCVRs are combined through the PMOS switch, which occurs when OCVR1 is not able to supply the total current demand on PDN1.

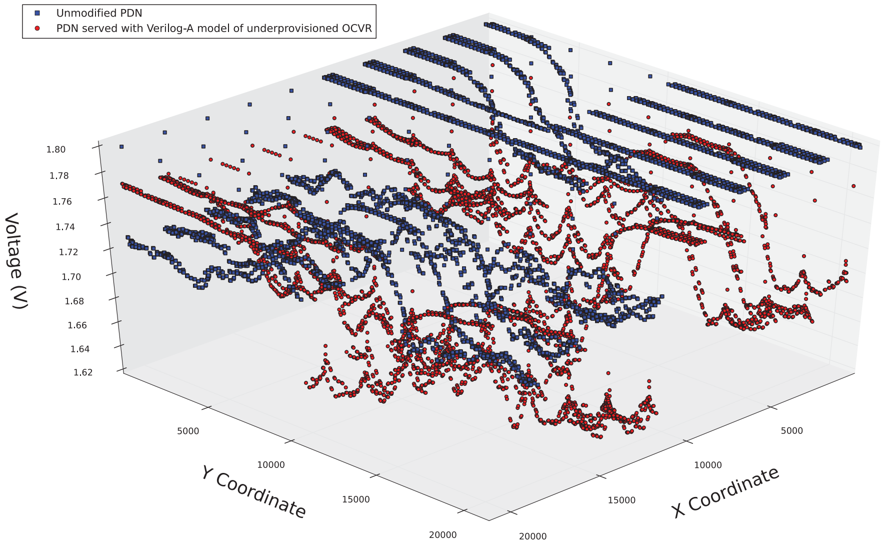


Fig. 15. Comparison of the worst case voltage droop across the PDN for an unmodified *ibmpg1t* PDN benchmark and a PDN served with the proposed interconnected current limited OCVRs combined at run time.

The current demand of the two PDNs and the current supplied by the respective OCVRs connected to the PDNs is shown in Fig. 14 for a time interval of 9 ns. The voltage signal controlling the gate of the PMOS switch connecting the output of the two OCVRs is also shown in Fig. 14. When the current consumption of PDN1 exceeds 13 A (the value of $\triangle I$ using Algorithm 1), the PMOS switch is turned ON, and the output current of the two OCVRs is combined. The combined outputs of the two OCVRs provide uninterrupted current flow to the two PDNs. The impact on the power supply voltage droop on PDN1 while combining the outputs of the OCVRs is analyzed. The time stamp corresponding to the maximum *IR* drop on any node in PDN1 is identified, and the power supply voltage profile across all nodes for that time is compared with an unmodified *ibmpg1t* power grid with off-chip voltage regulation. The comparison of power supply voltage noise for both the proposed and unmodified PDNs is shown in Fig. 15. Despite the activation and deactivation of the PMOS switches, the maximum *IR* drop across PDN1 does not exceed 10% of the power supply voltage of 1.8 V and is 0.18% greater than the worst case *IR* drop measured on the unmodified *ibmpg1t* benchmark PDN. The transient analysis of the *ibmpg1t* PDN with a Verilog-A model of the OCVR is performed to demonstrate the feasibility of the proposed RPDN for a CMP or many-core system. The analysis is performed on two PDNs for two cores. Therefore, the maximum current capability of the OCVR, the width of the PMOS switch, and the additional decoupling capacitance are much larger than that needed when the RPDN with OCVRs is designed for a system with larger core count.

## IX. Conclusion

A load-balanced circuit and system technique to deliver average power through on-chip VRs (OCVRs) is developed. The current rating of each OCVR is reduced to support only the average current demands of typical workloads executed on the CMP system. The reduction in the maximum output current of the OCVRs improves the PCE of the OCVRs, reduces the footprint of the PDN by at least 23%, and improves the energy efficiency of the CMP system by up to 44%. The proposed interconnected PDN is applicable to any OCVR circuit topology and offers higher reliability through a run-time OCVR clustering technique that prevents system failure when the current demand of the core exceeds the maximum output current supported by the OCVRs. A real-time workload mapping heuristic is developed to minimize the reconfiguration of the PDN, which schedules the tasks and assigns optimum DVFS levels for each core. The workload mapping heuristic in conjunction with the run-time reconfiguration of the PDN ensure reliable and energy-efficient operation of the CMP. The cross-layer techniques developed for power delivery introduce intelligence in the on-chip PDN analogous to a smart grid.

## Acknowledgment

The authors would like to thank Md. Hossein Hajkazemi and Md. Khavari Tavana from Northeastern University, Boston, MA, USA, for their contribution in performing the power trace analysis on SPEC CPU benchmarks.

## References

[1] D. Pathak, M. H. Hajkazemi, M. K. Tavana, H. Homayoun, and I. Savidis, "Energy efficient on-chip power delivery with run-time voltage regulator clustering," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2016, pp. 1210–1213.

[2] D. Pathak, M. H. Hajkazemi, M. K. Tavana, H. Homayoun, and I. Savidis, "Load balanced on-chip power delivery for average current demand," in *Proc. Great Lakes Symp. VLSI*, May 2016, pp. 439–444.

[3] I. H. R. Jiang, G. J. Nam, H. Y. Chang, S. R. Nassif, and J. Hayes, "Smart grid load balancing techniques via simultaneous switch/tie-line/wire configurations," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des. (ICCAD)*, Nov. 2014, pp. 382–388.

[4] M. Kar, S. Carlo, H. K. Krishnamurthy, and S. Mukhopadhyay, "Impact of process variation in inductive integrated voltage regulator on delay and power of digital circuits," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design*, Aug. 2014, pp. 227–232.

[5] W. Godycki, C. Torng, I. Bukreyev, A. Apsel, and C. Batten, "Enabling realistic fine-grain voltage scaling with reconfigurable power distribution networks," in *Proc. 47th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Dec. 2014, pp. 381–393.

[6] E. A. Burton *et al.*, "FIVR—Fully integrated voltage regulators on 4th generation Intel core SoCs," in *Proc. IEEE Appl. Power Electron. Conf. Expo.*, Mar. 2014, pp. 432–439.

[7] W. Lee, Y. Wang, and M. Pedram, "Optimizing a reconfigurable power distribution network in a multicore platform," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, no. 7, pp. 1110–1123, Jul. 2015.

[8] W. Kim, M. S. Gupta, G.-Y. Wei, and D. Brooks, "System level analysis of fast, per-core DVFS using on-chip switching regulators," in *Proc. IEEE 14th Int. Symp. High Perform. Comput. Archit.*, Feb. 2008, pp. 123–134.

[9] S. Manoj, H. Yu, and K. Wang, "3D many-core microprocessor power management by space-time multiplexing based demand-supply matching," *IEEE Trans. Comput.*, vol. 64, no. 11, pp. 3022–3036, Nov. 2015.

[10] I. Vaisband and E. G. Friedman, "Heterogeneous methodology for energy efficient distribution of on-chip power supplies," *IEEE Trans. Power Electron.*, vol. 28, no. 9, pp. 4267–4280, Sep. 2013.

[11] I. Vaisband and E. G. Friedman, "Energy efficient adaptive clustering of on-chip power delivery systems," *Integr., VLSI J.*, vol. 48, pp. 1–9, Jan. 2015.

[12] M. K. Tavana, D. Pathak, M. H. Hajkazemi, I. Savidis, and H. Homayoun, "Realizing complexity-effective on-chip power delivery for many-core platforms by exploiting optimized mapping," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des.*, Oct. 2015, pp. 581–588.

[13] D. M. Tullsen, "Simulation and modeling of a simultaneous multithreading processor," in *Proc. Int. Conf. Resource Manage. Perform. Eval. Enterprise Comput. Syst.*, 1996, pp. 819–828.

[14] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *Proc. 42nd Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2009, pp. 469–480.

[15] K. Pearson, "Contributions to the mathematical theory of evolution," *Philos. Trans. Roy. Soc. London*, vol. 185, pp. 71–110, Jan. 1894.

[16] C. W. Gellings and J. H. Chamberlin, *Demand-Side Management*. Boca Raton, FL, USA: CRC Press, 1988.

[17] A. Colin, A. Kandhalu, and R. Rajkumar, "Energy-efficient allocation of real-time applications onto heterogeneous processors," in *Proc. IEEE 20th Int. Conf. Embedded Real-Time Comput. Syst. Appl.*, Aug. 2014, pp. 1–10.

[18] Y. Shin *et al.*, "28 nm high-$K$ metal gate heterogeneous quad-core CPUs for high-performance and energy-efficient mobile application processor," in *Proc. IEEE Int. SoC Design Conf.*, Nov. 2013, pp. 198–201.

[19] E. Salman and E. G. Friedman, *High Performance Integrated Circuit Design*. New York, NY, USA: McGraw-Hill, 2012.

[20] E. J. Fluhr *et al.*, "The 12-core POWER processor with 7.6 Tb/s IO bandwidth, integrated voltage regulation, and resonant clocking," *IEEE J. Solid-State Circuits*, vol. 50, no. 1, pp. 10–23, Jan. 2015.

[21] Texas Instruments. *WEBENCH Design Center*. [Online]. Available: http://webench.ti.com.

[22] Y. Choi, N. Chang, and T. Kim, "DC–DC converter-aware power management for low-power embedded systems," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 26, no. 8, pp. 1367–1381, Aug. 2007.

[23] H. C. Cranford, Jr., L. L. C. Hsu, and J. S. Mason, "Power network reconfiguration using MEM switches," U.S. Patent 7 305 571, Dec. 4, 2007.

[24] O. Y. Loh and H. D. Espinosa, "Nanoelectromechanical contact switches," *Nature Nanotechnol.*, vol. 7, no. 5, pp. 283–295, Apr. 2012.

[25] Z. Li, R. Balasubramanian, F. Liu, and S. Nassif, "2012 TAU power grid simulation contest: Benchmark suite and results," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Des. (ICCAD)*, Nov. 2012, pp. 643–646.

[26] *IBM Power Grid Benchmarks*. [Online]. Available: http://dropzone.tamu.edu/ pli/PGBench/

[27] Texas Instruments. *DC to DC Converter Parallel Operation Using Droop Compensation of TPS5210*. [Online]. Available: http://www.ti.com/lit/an/slva060/slva060.pdf

**Divya Pathak** (S'01) received the M.S. degree in computer engineering from Drexel University, Philadelphia, PA, USA, in 2015, where she is currently pursuing the Ph.D. degree in electrical and computer engineering.

She was a Research Scientist with Indian Space Research Organization, Ahmedabad, India, from 2003 to 2006, where she was involved in analog signal processing modules for communication and remote sensing satellites. She was a Staff Engineer with STMicroelectronics, G.Noida, India, from 2006 to 2013, where she focused on functional and electrical postsilicon validation of set-top box MPSoCs. She was an Intern with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, from 2015 to 2016, where she was involved in power supply noise mitigation techniques for IBM POWER and Z Systems. She is the recipient of the 2017 IEEE Circuits and Systems Pre-Doctoral Scholarship. Her current research interests include circuits, systems, and microarchitecture design for post-Moore computing, including algorithms, modeling, optimization, and VLSI design for computing devices ranging from exascale systems to low-power internet of things.

**Houman Homayoun** (S'07–M'12) received the B.S. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in 2003, the M.S. degree in computer engineering from the University of Victoria, Victoria, BC, Canada, in 2005, and the Ph.D. degree from the Department of Computer Science, University of California, Irvine, CA, USA, in 2010.

He was with the University of California at San Diego, La Jolla, CA, USA, as a National Science Foundation Computing Innovation Fellow awarded by the CRA and CCC. He is currently an Assistant Professor with the Electrical and Computer Engineering Department, George Mason University (GMU), Fairfax, VA, USA, where he currently holds a Joint Appointment with the Computer Science and Information Science and Technology Departments. He is the Director of the Green Computing and Heterogeneous Architectures Laboratory, GMU. He is currently leading a number of research projects, including the design of next generation heterogeneous multicore accelerators for big data processing, nonvolatile STT logic, heterogeneous accelerator platforms for wearable biomedical computing, and logical vanishable design to enhance hardware security, which are all funded by the National Science Foundation, Arlington, VA, USA, General Motors Company, Detroit, MI, USA and Defense Advanced Research Projects Agency, Arlington County, VA, USA.

**Ioannis Savidis** (S'03–M'13) received the B.S.E. degree in electrical and computer engineering and biomedical engineering from Duke University, Durham, NC, USA, in 2005, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Rochester, Rochester, NY, USA, in 2007 and 2013, respectively.

He was with the 3-D Integration Group, Freescale Semiconductor, Inc., Austin, TX, USA, in 2007. He was with the System on Package and 3-D Integration Group, IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, from 2008 to 2011. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, 0050A, USA, where he currently directs the Integrated Circuits and Electronics Design and Analysis Laboratory. His current research interests include analysis, modeling, and design methodologies for high-performance digital and mixed-signal integrated circuits, power management for system-on-chip and microprocessor circuits (including on-chip dc-dc converters), emerging integrated circuit technologies, IC design for trust (hardware security), and interconnect-related issues in 2-D and 3-D ICs, with a specific emphasis on electrical and thermal modeling and characterization, signal and power integrity, and power and clock delivery for heterogeneous 2-D and 3-D circuits.

Dr. Savidis is an Editorial Board Member of the IEEE Transactions on Very Large Scale Integration Systems, the *Microelectronics Journal*, and the *Journal of Circuits, Systems and Computers*.