Energy Efficient On-Chip Power Delivery with Run-Time Voltage Regulator Clustering

Divya Pathak*, Mohammad Hossein Hajkazemi†, Mohammad Khavari Tavana†, Houman Homayoun† and Ioannis Savidis*

*Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA
(divya.pathak, ioannis.savidis)@drexel.edu
†Department of Electrical and Computer Engineering, George Mason University, Fairfax, Virginia, VA
(mhajkaze,mkhavari,hhomayoun)@gmu.edu

***Abstract*—In this paper, a power delivery system for homogeneous chip multi-processor (CMP) systems is proposed. The power delivery system is modified at run time by clustering multiple on-chip voltage regulators (OCVR) depending on the power demand of the workload. The OCVRs are designed to deliver up to the average current requirement of the typical workloads executed on the CMP platform. When the current demand of a core cluster exceeds the average value, the output of multiple OCVRs is combined through a high-speed switch network to provide the necessary current. Two OCVR topologies (Buck and LDO) are analyzed to characterize the impact on the characteristics of the voltage regulator as the peak load current is reduced. Simulation results for run-time OCVR clustering indicate a 36% reduction in the energy consumption of the system at an average load current with improvement in the load regulation. In addition, the area occupied by the OCVRs is reduced by at least 70%.**

Keywords*—power management, on-chip voltage regulation, run-time voltage regulator clustering.

I. Introduction

Modern ICs operate under tight power and thermal budgets. Power efficiency has therefore emerged as a critical design parameter in CMP systems. The peak power variation of the IC constrains the design of the power delivery system and consequently the thermal dissipation system. The power rating and the design topology of the voltage regulator (VR) is selected based on the maximum possible power consumption of the load circuit served by the VR. Conventionally, for a VR serving a single high-performance processor, the thermal design power (TDP) of the processor is used to set the power rating of the VR. If the VR serves a cluster of identical cores, then the combined TDP of the cores determines the power rating of the VR. Determining the peak power consumption for a single core is a challenging task and is achieved by running carefully written code called a power virus [1]–[3] which emulates all possible execution behaviors of a workload while stressing each component of the core.

Alternatively, simulation tools like McPAT [2] provide cycle accurate estimates of the peak power consumption of a multicore system. However, the precision of the peak power reported by McPAT depends on the granularity at which the simulator provides information on the activity factor of each circuit block of the core [2]. In the absence of information on block activity, a higher estimate of the activity factor is made.

As neither a power virus or power estimation tools like McPAT provide an accurate calculation of the peak power consumption, the power delivery system and consequently the cooling system for a CMP is over-provisioned. In this paper, a CMP work load aware power delivery system is proposed. The proposed power delivery system uses OCVRs, as OCVRs offer the opportunity to apply dynamic voltage and frequency scaling (DVFS) at a finer granularity and provide fast response to load transients. Each OCVR is capable of supporting a peak current rating equal to the average load current I_{avg} consumed across all workloads. For a typical core configuration, I_{avg} is an order of magnitude less than the peak current I_{peak} determined from the peak power P_{peak} reported by McPAT. Reducing the size of the OCVRs to support I_{avg} improves the power and energy efficiency of the circuit along with reducing the occupied on-chip area. A circuit technique to cluster the OCVRs to support load currents in excess of I_{avg} is proposed. The technique does not degrade any of the figures of merit of the OCVRs. A detailed analysis is done for two popular OCVR topologies, the low drop-out (LDO) regulator and the DC-DC switching buck converter, to determine the effect on load transient response and power conversion efficiency (PCE) when designing the power delivery system to support a maximum current of I_{avg}.

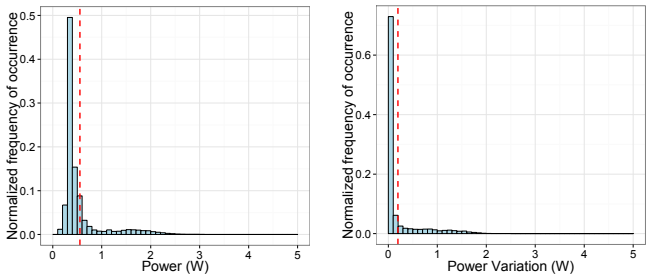
Recent work [4], [5], [6] has attempted to address the over-provisioning of the power distribution network (PDN). Techniques have been proposed to reconfigure the PDN according to the power requirements of the work load. The analysis, however is limited to one VR topology. In this work a technique for run time reconfiguration of the PDN is developed irrespective of the OCVR topology.

The rest of the paper is organized as follows: The system level simulation and results for the power consumption profile of multi-application workloads is described in Section II. The proposed power delivery circuit is discussed in Section III. The simulated results are analyzed in Section IV. Concluding remarks are provided in Section V.

II. Power dissipation in CMPs

Applications show different power dissipation behavior in different execution phases. The power dissipation pattern of the 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. In this work, a 16-core CMP in a 45 nm technology is modeled using a processor architectural simulator [7]. McPAT is integrated with the simulator to analyze the power consumption of the core. Each core has a 2-way issue and out-of-order execution unit. A set of 38 benchmarks from the SPEC2000 and SPEC2006 suites are studied to determine the power dissipation behavior of the core. Each benchmark is simulated for four different timing intervals to cover multiple execution phases with different power consumption profiles. The

978-1-4799-5341-7/16/$31.00 ©2016 IEEE



(a) (b)

Fig. 1. Histogram of (a) power dissipation, and (b) variation in power for consecutive cycles when executing SPEC2000 and SPEC2006 benchmarks.

simulations run for 10K cycles per time interval and sampling is performed cycle by cycle.

The consolidated power consumption histogram and the power variation histogram of the power consumed by the studied benchmarks is shown in Fig. 1. The dashed line in Fig. 1(a) indicates the average value of the power consumption. As is depicted in Fig. 1(a), the average power dissipation is around 0.55 W. For approximately 65% of the execution time, the power dissipation of the applications is in the range of 0.3 W to 0.5 W. The average variation in power dissipation between cycles is 0.2 W, as is shown in Fig. 1(b). The power variation is less than 0.1 W for about 90% of the time. Furthermore, the studied benchmarks spend more than 78% of the run-time consuming less than the average power. The peak power P_{peak} of 5.73 W reported by McPAT is never consumed. The maximum power consumption of 4.75 W across all workloads is consumed for a very small percentage of the run-time (7.5×10^{-5}%).

III. Proposed Power Delivery Methodology

The statistical analysis of the power consumption shown in Fig. 1 indicates that the VR rating is over-provisioned for the majority of the run-time of the workloads. As a result, a reconfiguration of the OCVRs in a CMP system provides opportunity for increased energy and area savings. The block representation of the proposed power delivery topology is shown in Fig. 2. For a multi-core system consisting of N cores or core clusters, N OCVRs provide the regulated power. The output of each OCVR is connected to the inputs of an NxN crossbar switch. The N outputs of the crossbar switch are connected to the PDN of the N clusters. The high-speed switching (HSS) fabric is controlled by the power management unit (PMU). 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_{sense}) reaches a threshhold $\triangle I$ below I_{avg}, the PMU configures the HSS to source additional current from the OCVRs which are operating at the same power supply voltage (V_{dd}) level when DVFS is implemented.

The logic controlling the HSS fabric within the PMU operates on two system parameters, the V_{dd} levels and the total current load sensed from each core cluster. The analysis of the power consumption described in Section II indicates that the probability of the load current demand exceeding I_{avg} is 22%. As a result, there are always more than one core clusters operating at or below I_{avg}. The PMU is provisioned to add at least one additional OCVR to serve a cluster requiring current higher than I_{avg}. When DVFS is implemented, the challenge becomes in finding an additional OCVR operating at the same V_{dd} level. A

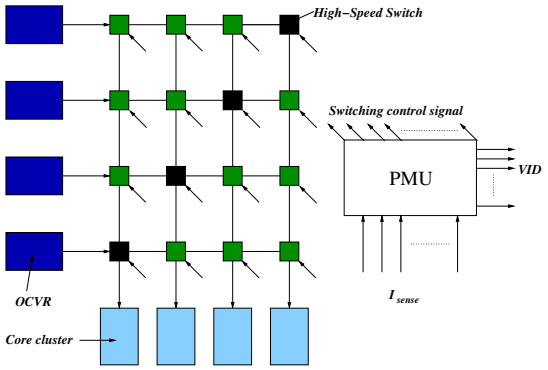


Fig. 2. Run-time voltage regulator clustering performed by the PMU through a cross bar switching fabric.

free OCVR is ensured if the number of DVFS levels is less than the number of clusters in the multi-core system. If a spatially proximal OCVR is found that operates at a different voltage level, the PMU aligns the voltage levels of the two OCVRs that are to be combined. The sum of the decision time of the PMU and the HSS switching time 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_{peak}, to ensure an uninterrupted power supply to the core cluster. The control of the HSS fabric is described by Algorithm 1.

Algorithm 1 Run-time OCVR clustering to support higher than I_{avg} current consumption by core clusters.

n = number of core clusters, m = number of DVFS levels
Core/Core cluster id: i, x, and y ∈ [1,...,n]
▷ **Default HSS configuration**:
if x == y **then**
  $HSS_{xy} \leftarrow 1$
else
  $HSS_{xy} \leftarrow 0$
end if
▷ **Inputs**
Sensed current from each core: I_{sense_x}, threshhold ($\triangle I$),
Voltage level applied to all core clusters: $V_x \in [V_{dd\_1}, V_{dd\_2}, ..., V_{dd\_m}]$
▷ **Constraints**
m < n; $t_{switch} + t_{PMU} < t_{core}$; $\sum_{i=1}^{n} V_x \cdot I_{sense\_x} <$ n$\cdot V_{dd_m} \cdot I_{avg}$
▷ **OCVR Clustering: Positive load transient**
if ($I_{sense\_x} \geq I_{avg}$ - $\triangle$I) **then**
  search for smallest i such that y = x ± i
  if ($I_{sense\_x} + I_{sense\_y} < 2 \cdot I_{avg}$ && $V_x == V_y$) **then**
    $HSS_{xy} \leftarrow 1$
  else
    if (x+i == n || x-i == 0) **then**
      search for smallest i such that y = x ± i && $I_{sense\_y} < 0.5 \cdot I_{avg}$
      $V_y \leftarrow V_x$
      $HSS_{xy} \leftarrow 1$
    end if
  end if
end if
▷ **De-cluster: Load release**
if ($HSS_{xy} == 1$) && (I_{sense_x}+I_{sense_y}<2$\cdot$(I_{avg} -$\triangle$I)) **then**
  $HSS_{xy} \leftarrow 0$
end if

IV. Simulation Results and Analysis

Two popular OCVR topologies, the LDO and buck converter, are examined and an analysis of the changes in the figures of merit of each when reducing the load current rating is analyzed. In general, the OCVR response times vary based on the value of the output capacitance, the associated effective series resistance (ESR) and effective series inductance (ESL), and the magnitude of the load current transient. SPICE simulations are performed for a baseline case of one OCVR per core and the results are analyzed to determine the optimum power delivery configuration.

TABLE I. IMPROVEMENT IN THE FIGURES OF MERIT OF THE OCVR SUPPORTING I_{avg} INSTEAD OF I_{peak}.

OCVR Topology	Area reduction	PCE	Output voltage ripple reduction	Load regulation	Line regulation
LDO	0.1 x *Area*	1 x *PCE*	0.1 x V_{ripple}	10 x β_{load}	1 x β_{line}
Buck	0.16 x *Area*	1.3 x *PCE*	0.1 x V_{ripple}	10 x β_{load}	1 x β_{line}

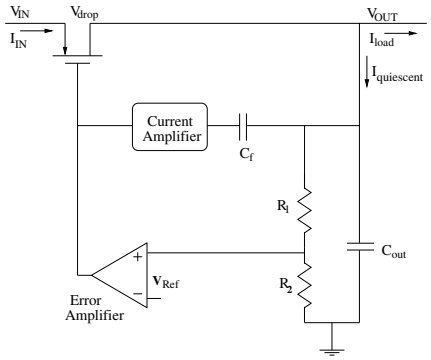


Fig. 3. Circuit schematic of an LDO with only a primary feedback loop.

A 16 core CMP is considered with one OCVR per core. The peak current rating of each OCVR is set to 0.6 A, which corresponds to the average current consumed across all workloads. The cores are connected through a 16x16 switching network, which is implemented with PMOS switches.

A. Analysis of OCVR figures of merit

The figures of merit of an OCVR [8] designed to support I_{avg} improve as compared to an OCVR that supports I_{peak}. The improvement for some of the figures of merit of an OCVR supporting I_{avg} are summarized in Table I. The capability to maintain a constant output voltage with changes in the load current is described by β_{load} and with changes in the input voltage by β_{line}.

1) Power conversion efficiency of an LDO: The PCE of an LDO is given by (1).

$$\begin{aligned} P_{out} &= V_{OUT} I_{load} = (V_{IN} - V_{drop}) I_{load} \\ P_{in} &= V_{IN}(I_{load} + I_{quiescent}) \\ \eta &= P_{out}/P_{in} \end{aligned} \tag{1}$$

The power conversion efficiency (PCE) of an LDO does not significantly improve with a reduction in I_{load} even if $I_{quiescent}$ is proportionately lowered by reducing the drive strength of the error amplifier circuit. The $I_{quiescent}$ becomes a critical parameter for the LDO when the output current rating is low. Lowering $I_{quiescent}$ degrades the transient response of an LDO, which does not have a secondary feedback loop to drive the pass element directly [9], as shown in Fig. 3.

2) Power conversion efficiency of a buck converter: The power consumed by the buck converter P_{buck} is given by (2) [8]. The P_{mos}, P_{ind}, P_{cap}, and P_{pwm} are the power loss in, respectively, the MOS power transistors and the cascaded buffers driving them, the inductor of the filter circuit, the capacitor of the filter circuit, and the pulse width modulator circuit. The detailed mathematical formulae of each of the components which contribute towards P_{buck} are given in [8], [10].

$$P_{buck} = P_{mos} + P_{ind} + P_{cap} + P_{pwm} \tag{2}$$

The size of the filter inductor is chosen such that the percentage of peak current ripple (L_{pp}) remains the same even with

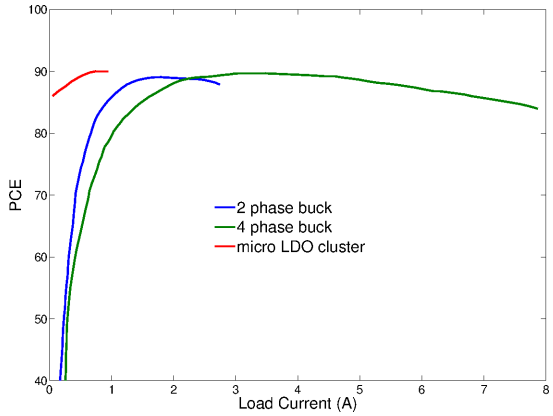


Fig. 4. Variation in power conversion efficiency with load current for a multi-phase buck converter OCVR (V_{IN}/V_{OUT}=1.7V/1.0V, $F_{switching}$ = 140 MHz) and clustered micro LDOs (V_{IN}/V_{OUT}=1.1V/1.0V). The micro LDOs offer higher PCE at low load current.

an order of magnitude reduction in the peak current drive. The result is an approximate one-third reduction in P_{buck} when using the smaller converter. The theoretical calculations are verified through simulation of two custom buck converters with peak load current ratings of 6 A and 0.6 A [11]. The power consumption of the various components of the buck converter along with the occupied area are listed in Table II. The on-chip implementation of the two buck converters yields similar ratios between the power consumed by each component.

Although a 33% reduction in the total power consumption of the buck converter is achieved by reducing the peak load current rating, the reduction in the PCE with decreasing load current remains significant. The variation of the PCE with load current for a 2 and 4 phase fully integrated voltage regulator (FIVR) [12] used in the Intel Haswell processor is shown in Fig. 4. Each phase of the FIVR is a buck converter that supports a peak load current of 1.75 A. The combined PCE of the on-chip micro LDOs based on the model described in [13], [14] is also shown in Fig. 4. The micro LDOs support a maximum load current of 50 mA to 100 mA and offer a peak PCE of 90% when the outputs are combined to support the I_{avg} requirement of the core or the core clusters.

B. Energy efficiency of the CMP system

The total energy consumption of the CMP system for a given execution time T_{epoch} with N cores and N OCVRs is given by (3). The cores are served by over-provisioned OCVRs (load current rating of I_{peak}) identical to a 4-phase FIVR. The dynamic and static power consumed by the cores in the presence of DVFS are given by, respectively, $P_{dynamic\_i}$ and P_{static}. PCE_1 represents the power conversion efficiency of the OCVR. For low load currents close to I_{avg}, the PCE_1 offered by the OCVR is 50%. Alternatively, if the power delivery system is designed with each core supported by a cluster of micro LDOs with a combined load current rating equal to I_{avg}, the achieved PCE_2 is 90%. In addition, the static power consumed by the cores or core clusters is almost zero as the idle core(s) are power

TABLE II. POWER CONSUMPTION ANALYSIS OF A DC-DC SWITCHING BUCK CONVERTER WITH VARYING PEAK LOAD CURRENT.

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

gated through the high speed switching (HSS) fabric. The HSS fabric imposes an additional switching loss P_{switch}, which is the dynamic power consumed by the PMOS transistors while switching, and a conduction loss $P_{conduction}$ while in the ON state and passing the average current I_{avg}.

$$E_{CMP,conventional} = \{\sum_{i=1}^{N} \frac{(P_{dynamic\_i} + P_{static})}{PCE_1}\} \cdot T_{epoch} \quad (3)$$

The total energy consumed by the CMP with *N* OCVRs and *N*x*N* PMOS switches (each OCVR is designed for an I_{avg} rating) is given by (4). The parameters *j*, *k*, and *l* are, respectively, the number of active cores consuming current below I_{avg}, the number of active core(s) consuming current above I_{avg}, and the number of idle core(s) power gated through the high speed switching network. In the case of idle cores, the power consumed by the OCVRs ($I_{quiscent}$·V_{out}) is the only component contributing to the system energy. As described in Section II, the benchmark applications consume current less than I_{avg} for about 78% of the execution time. The P_{switch} loss is incurred for 22% of the execution time of the workloads when the load current demand exceeds I_{avg}.

$$\begin{aligned} E_{CMP,proposed} = \sum_{t=1}^{T_{epoch}} \{ & \sum_{i=1}^{j} \frac{(P_{dynamic\_i} + P_{static})}{PCE_2} \\ & + \sum_{i=1}^{k} \frac{(P_{dynamic\_i} + P_{static} + P_{switch} + P_{conduction})}{PCE_2} \\ & + \sum_{i=1}^{l} I_{quiscent} \cdot V_{out}\}; \\ & j + k + l = N \end{aligned} \quad (4)$$

The P_{static} of a single core is measured through McPAT. The P_{switch} and $P_{conduction}$ for the PMOS switch with an output capacitance provided by a single core is determined through SPICE simulations. Despite the additional switching and conduction loss due to the PMOS switches, the percentage reduction in energy consumption for a core consuming less than I_{avg} is 36%. The energy efficiency of the CMP therefore improves when designing the power delivery system with micro LDOs and an HSS switching fabric. In addition, if a workload aware thread to core mapping algorithm accounts for the PCE variation of the LDO, serving the core clusters with LDOs further reduces energy consumption by 38% as compared to PCE agnostic algorithms [15]. The multi-phase buck converter is a suitable candidate for serving core(s) when I_{avg} is above 2 A, as the buck offers a higher PCE for larger load currents, as shown in Fig. 4.

V. CONCLUSIONS

A circuit technique to deliver average current through on-chip voltage regulators is described. The current rating of each on-chip voltage regulator is reduced to support only the average current demands of typical workloads executed on the CMP system. The reduction in load current improves the figures of merit of the OCVRs along with a minimum reduction of 70% in the foot print and a maximum improvement of 36% in the system energy efficiency. A run-time OCVR clustering technique is proposed which does not degrade the OCVR load current transient response time in the case that the load current exceeds the average value. The simulated results indicate that the optimum OCVR configuration for a CMP system depends on the average load current requirement per core.

REFERENCES

[1] K. Ganesan, J. Jo, W.L. Bircher, D. Kaseridis, Z. Yu, and L.K. John, "System-level Max Power (SYMPO)-A Systematic Approach for Escalating System-level Power Consumption using Synthetic Benchmarks Categories and Subject Descriptors," *Proceedings of the International Conference on Parallel Architectures and Compilation Techniques*, pp. 19–28, September 2010.

[2] 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," *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*, pp. 469–480, December 2009.

[3] mersenne.org, "Great Internet Mersenne Prime Search," http://www.mersenne.org/download/#stresstest.

[4] W. Lee, Y. Wang, and M. Pedram, "Optimizing a Reconfigurable Power Distribution Network in a Multicore Platform," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 34, No. 7, pp. 1110–1123, July 2015.

[5] W. Godycki, C. Torng, I. Bukreyev, A. Apsel, and C. Batten, "Enabling Realistic Fine-Grain Voltage Scaling with Reconfigurable Power Distribution Networks," *Proceedings of the IEEE/ACM International Symposium on Microarchitecture*, pp. 381–393, December 2014.

[6] M. Sai and H. Yu, "3D Many-core Microprocessor Power Management by Space-Time Multiplexing based Demand-supply Matching," *IEEE Transaction on Computers*, Vol. 64, No. 11, pp. 3022–3036, November 2015.

[7] D.M. Tullsen, "Simulation and Modeling of a Simultaneous Multithreading Processor," *Proceedings of the International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems, CMG. Part 2(of 2)*, pp. 819–828, December 1996.

[8] E. Salman and E.G. Friedman, *High Performance Integrated Circuit Design*, McGraw Hill, 2012.

[9] E. Rogers, "Stability Analysis of Low-Dropout Linear Regulators with a PMOS Pass Element," *Texas Instruments Inc.*, pp. 1–4, August 1999.

[10] Y. Choi, N. Chang, and T. Kim, "DC–DC Converter-Aware Power Management for Low-Power Embedded Systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 26, No. 8, pp. 1367–1381, August 2007.

[11] Texas Instruments, "WEBENCH Design Center," http://webench.ti.com.

[12] E. A. Burton, G. Schrom, F. Paillet, J. Douglas, W. J. Lambert, K. Radhakrishnan, and M. J. Hill, "FIVR–Fully Integrated Voltage Regulators on 4th Generation Intel Core SoCs," *Proceedings of the IEEE Applied Power Electronics Conference and Exposition*, pp. 432–439, March 2014.

[13] J.F. Bulzacchelli, Z. Toprak-Deniz, *et al.*, "Dual-Loop System of Distributed Microregulators With High DC Accuracy, Load Response Time Below 500 ps, and 85-mV Dropout Voltage," *IEEE Journal of Solid-State Circuits*, Vol. 47, No. 4, pp. 863–874, April 2012.

[14] C.J Park, M. Onabajo, and J. Silva-Martinez, "External Capacitor-Less Low Drop-Out Regulator with 25 dB Superior Power Supply Rejection in the 0.4–4 MHz Range," *IEEE Journal of Solid-State Circuits*, Vol. 49, No. 2, pp. 486–501, February 2014.

[15] M. Tavana, D. Pathak, M. Hajkazemi, I. Savidis, and H. Homayoun, "Realizing Complixity-Effective On-Chip Power Delivery for Many-Core Platforms by Exploiting Optimized Mapping," *Proceedings of the International Conference on Computer Design*, pp. 581–588, October 2015.