Peak Power Reduction and Workload Balancing by Space-Time Multiplexing based Demand-Supply Matching for 3D Thousand-core Microprocessor

Sai Manoj P. D., Kanwen Wang, and Hao Yu
School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore 639798

ABSTRACT

Space-time multiplexing is utilized for demand-supply matching between many-core microprocessors and power converters. Adaptive clustering is developed to classify cores by similar power level in space and similar power behavior in time. In each power management cycle, minimum number of power converters are allocated for space-time multiplexed matching, which is physically enabled by 3D through-silicon-vias. Moreover, demand-response based task adjustment is applied to reduce peak power and to balance workload. The proposed power management system is verified by system models with physical design parameters and benched power traces, which show 38.10% peak power reduction and 2.60x balanced workload.

Categories and Subject Descriptors: B.7.2 [Design Aids]

Keywords: Demand-supply matching, Peak power reduction, Workload balancing, 3D thousand-core

1. INTRODUCTION

Exa-scale cloud computing for big-data applications requires integration of many-core microprocessors on a single chip [1, 2] at thousand-core scale. Though 3D integration is one promising solution [3] to increase integration density and communication bandwidth, the provision of many-core power supply voltages with maintenance of low power density has become an unresolved issue to address [4, 5, 6, 7, 8]. Supplying same voltage-level to all cores will result in high power density because the demand of each core can be different at different time instant. As such, a demand-supply matched dynamic voltage and frequency scaling (DVFS) scenario needs to be employed during power management for both peak power reduction and workload balancing.

From physical hardware perspective, an optimal demand-supply matching requires on-chip power converters [5, 6, 7, 8], which can provide prompt DVFS management with efficient power delivery. However, one power converter for one core has large area overhead in presence of non-scalable buck inductor. The design of single-inductor-multiple-output (SIMO) power converters [6] utilizes one common single buck inductor to provide different voltage-levels at different time slots in a time-multiplexed manner. The capability of SIMO is, however, still limited for many-core microprocessors at thousand-core scale. Moreover, considering hundreds of cores to be integrated on one chip, the remaining area is quite limited to consider on-chip power converter with buck inductor. The 3D integration introduces additional room for on-chip power converters. The recent work in [8] has demonstrated the possibility to design power converter on one die and 64-tile network-on-chip on the other die, which are integrated by through-silicon-via (TSV).

From cyber management perspective, the power management for many-core power-supply system will no longer be the same as the one for the traditional single-core. For big-data applications, there may exist various power patterns deployed on many-cores with multi-time-scale demands for power supply. Moreover, there are many microprocessor cores but limited power converters. A number of power management works for many-core microprocessor system have been explored before [5, 6, 7, 8] but with not fully resolved challenge that requires to not only match various demands from microprocessors with limited number of power converters, but also to reduce peak power and to balance workload on a power converter. As such, the smart power management of many-core microprocessor has similarity as smart-grid though at different time-scale with different workload behaviors. Thereby, the study of workload behavior with classification and also the demand-response can be leveraged from smart-grid management [9] to deal with the on-chip demand-supply matching problem.

In this paper, a space-time multiplexing (STM) based DVFS power management is utilized for demand-supply matching between many-core microprocessors and power converters. In each power management cycle an adaptive clustering is developed such that the minimum number of power converters are allocated for different groups classified by power-magnitudes, called *space multiplexing*. In one group, power converters are further reused in different time slots for different subgroups classified by power-phases, called *time multiplexing*. Such a space-time multiplexed matching is physically enabled by designing a reconfigurable power switch network with the use of 3D through-silicon-vias (TSVs). Moreover, demand-response based task adjustment is applied to reduce peak power and to balance workload. The proposed power management system is verified by system models in SystemC-AMS. The physical design parameters are based on 130nm CMOS process with TSV models. Experiment results show that the proposed power management can achieve 38.10% peak power reduction and 2.60x balanced workload.

The rest of this paper is organized as follows. In Section 2, we present the 3D many-core microprocessor system architecture with space-time multiplexing (STM) problem formulation towards demand-supply matching. In Section 3, we show the solution by STM-based resource allocation of power converters with use of adaptive clustering, which is based on singular-value-decomposition (SVD) analysis of workload correlation. We further show the demand-response based task scheduling to utilize demand slacks and to adjust tasks for both peak power reduction and workload balancing. The experiment results are included in Section 5 with conclusion in Section 6.

DAC '13, May 29-June 07 2013, Austin, TX, USA.
Copyright 2013 ACM 978-1-4503-2071-9/13/05 ...$15.00.

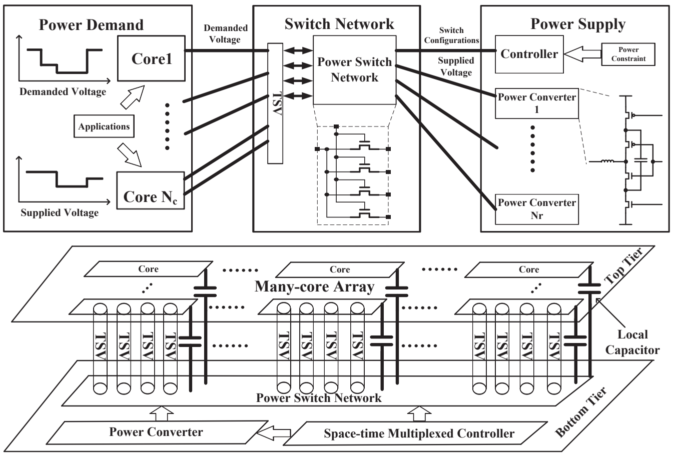

Figure 1: 3D reconfigurable power switch network for demand-supply matching between on-chip multi-output power converters and many-core microprocessors

Table 1: Notations and definitions

Notations	Definitions
$V = \{v_1, \ldots, v_{N_v}\}$	Set of voltage levels
$I = \{i_1, \ldots, i_{N_v}\}$	Set of core current loads
$R = \{r_1, \ldots, r_{N_r}\}$	Set of power converters
$C = \{c_1, \ldots, c_{N_c}\}$	Set of cores
$P = \{p_1, \ldots, p_{N_c}\}$	Set of power trace patterns
$S = \{s_1, \ldots, s_{N_s}\}$	Set of switch boxes
$G = \{g_1, \ldots, g_{N_v}\}$	Set of groups
$K = \{k_1, \ldots, k_{N_k}\}$	Set of subgroups
$A = \{a_1, \ldots, a_{N_k}\}$	Set of slacks
$L = \{l_1, \ldots, l_w\}$	Set of workloads
$B = \{b_1, \ldots, b_r\}$	Set of priorities
$v_d(c_i) \in V$	Demanded voltage-level of core c_i
$v_a(c_i) \in V$	Supplied voltage-level to core c_i
$v(r_i) \in V$	Output voltage-level of converter r_i
$d(r_i) \in V$	Output driving ability of converter r_i
I_L	Maximum converter inductance current
ΔV	Maximum core supply-voltage drop
H	Time slot for time-multiplexing
P_{th}	Peak power threshold

2. 3D SYSTEM ARCHITECTURE WITH SPACE-TIME MULTIPLEXING

In this section, 3D many-core microprocessor system architecture with a reconfigurable power switch network is reviewed with a space-time multiplexing (STM) problem formulated for power management. Table. 1 summarizes necessary notations used in this paper.

2.1 3D System Architecture

As shown in Fig. 1, the 3D many-core microprocessor system architecture is basically composed of two tiers. The bottom tier is for power management, including arrays of power converters and power switches. Each power converter is SIMO type, capable of supplying multiple voltage-levels by one buck inductor. The top tier includes array of many-core microprocessors. In between these two array-structured tiers, there are through-silicon-vias (TSVs), controlled by power switches, to connect power converters and cores. Moreover, there is one local super-capacitor for each core, working as local power storage to supply voltage during the multiplexing when power converter is not available.

The proposed 3D system architecture can be described by a demand-supply system model composed of the following three components:

- *Power Demand*: a set of cores C with demanded voltage-levels with set-size N_c. Each core c_i has a demanded voltage-level $v_d(c_i)$ to meet the deadline of its running workload. In addition, $v_a(c_i)$ is the allocated voltage-level to c_i after power management.
- *Power Supply*: a set of power converters R with set-size N_r. Each power converter outputs the voltage-level $v(r_i) \in V$ to supply the cores, where V is the set of available voltage-levels before power management;
- *Power Switch Network*: a set of reconfigurable switch-boxes S with set-size N_s to connect between R and C for demand-supply matching.

2.2 Space-Time Multiplexing Problem

As aforementioned in the introduction, the primary challenge in 3D thousand-core system to support exa-scale computing is to solve a large-scale demand-supply matching problem. Though there are various big-data applications with different power patterns, most of their power profiles can be still classified by magnitudes and phases. As such, if one can perform a detailed power profile characterization by clustering cores with similar power behaviors, the complexity in matching may be accordingly reduced. With the further consideration for implementation with the minimum cost of power converters, it is still feasible to formulate a resource (power converter) allocation problem with constraints of demand and supply matching. As such, one can formulate the first subproblem as follows.

Subproblem 1: Resource Allocation Problem is to decide the minimum number of power converters such that demand-supply matching can be satisfied.

What is more, due to spatial and temporal variation of power profiles, there may exist lots of power slacks to be utilized for a demand-response based workload scheduling. Without violating the workload execution priority or deadline, one can delay over-loaded workloads in one time-slot to the other time-slot with under-loaded workloads. As such, the peak power can be reduced as well as the workload can be balanced at power converters, which can be formulated as the second subproblem below.

Subproblem 2: Workload Scheduling Problem is to delay over-loaded workloads to under-loaded time-slots based on availability of slack and without violation of priority.

In this paper, we show that based on the aforementioned 3D system architecture, a space-time multiplexing (STM) based power management can be developed to solve the two subproblems in sections 3 and 4, respectively.

3. ADAPTIVE CLUSTERING BASED RESOURCE ALLOCATION

This section deals with resource allocation by adaptive clustering, resulting in the use of the minimum number of power converters for matched demand-supply. To deal with a large-scale demand-supply matching problem, we start with classification of cores into clusters by studying their power profile characteristic within one power management control-cycle T_c.

3.1 Grouping by Power Magnitude for Space Multiplexing

Grouping is the process of clustering different cores, which have similar power magnitudes and hence will demand the similar voltage-level.

Note that z-th group g_z, $g_z \in G$, can be formed by the following criteria

$$g_z = \{c_i; v_d(c_i) = v_d(c_j) = v_z, \forall i, j = 1, ...N_c, z \leq N_v\}. \quad (1)$$

Here, v_z, $v_z \in V$ represents the z-th voltage-level and c_i, $c_i \in C$ and $v_d(c_i) \in V$.

Based on the power magnitude levels, different groups are formed. Each group may contain different number of cores, which can have similar power magnitudes but maybe different power phases. The group formulation can change at different control-cycle. Based on the partitioned groups, power converters can be also partitioned in space to provide the specified voltage-levels for groups. This grouping process has less complexity because it involves just numerical comparisons.

3.2 Subgrouping by Power Phase for Time Multiplexing

Subgrouping is the process of clustering different cores, which have similar power phase (or pattern) and are within the same group.

Subgroup k_s, $k_s \in K$, can be formed by the following criteria

$$k_s = \{c_i; (v_d(c_i) = v_d(c_j) = v_z) \& (p_i \sim p_j), \forall i, j = 1, ...N_c\}. \quad (2)$$

Here, p_i, $p_i \in P$, represents the phase or pattern of one power trace of the core c_i, $c_i \in C$. $v_d(c_j)$ represents the demanded voltage-level of core c_i and v_z represents the z-th voltage-level, v_z, $v_d(c_j) \in V$. However, the subgrouping by phase is more difficult than grouping by magnitude and may consume bit more time in clustering. In the next subsection, we show a solution by means of spectral clustering to perform subgrouping of power profiles, which can be easily deployed to make power management faster compared to the one without subgrouping. Moreover, all the computations can be pre-stored in a look-up-table for implementing a real-time control.

3.3 Spectral Clustering for Subgrouping

Spectral clustering algorithm is discussed below. To find similarity between two power profiles p_i and p_j, $p_i, p_j \in P$, with N samples in one control-cycle, correlation in term of covariance matrix can be evaluated by

$$X = \frac{1}{N} \sum_{i,j=1}^{N} (p_i - \overline{P})(p_j - \overline{P})^T \quad (3)$$

where $\overline{P}$ is the mean of all power profiles $(\frac{1}{N}\sum_{i=1}^{N}(p_i))$.

Based on the order of covariance matrix, number of clusters, K can be analyzed by the singular-value-decomposition (SVD) of covariance matrix

$$X = U \times S \times V^{-1}. \quad (4)$$

Matrices U and V are orthogonal matrices with S as the diagonal matrix. Based on the rank analysis of S, the number of clusters K can be decided. A new matrix can be formed with K independent vectors, extracted from either of orthogonal matrices. Let the newly formed matrix be V_K, assuming it is extracted from V. The product of V_K with the covariance matrix X

$$X_K = X \times V_K \quad (5)$$

will result in a reduced matrix X_K, which becomes the basis of spectral clustering for subgrouping. For example, one core will be allocated to i-th subgroup if the value of $X_K(j, i)$ is the maximum in jth-row. The procedure for subgrouping is described in Algorithm 1.

Algorithm 1 Subgrouping by correlation extraction and spectral clustering

```
INPUT: Power trace matrix $P$ with $p_i$ power trace vectors after grouping
1. Compute covariance matrix $R \in R^{p_i \times p_i}$
2. Perform SVD: $R = U \times S \times V^{-1}$
3. Determine number of clusters: $K = rank(S)$
4. Compute the first K singular-value vectors $v_1, ....v_K$ of $V$
5. Let $V_K = [v_1, ..., v_K] \in R^{N \times K}$ and $R_K = R \times V_K$
6. Add $ith$ core to $jth$ cluster if $R_K(i, j)$ is maximum in the $ith$ row
7. Form $P_K$ matrices by finding corresponding indices in power trace matrix $P$
OUTPUT: New clustered subgroup matrices $P_K, (k = 1, ..., K)$
```

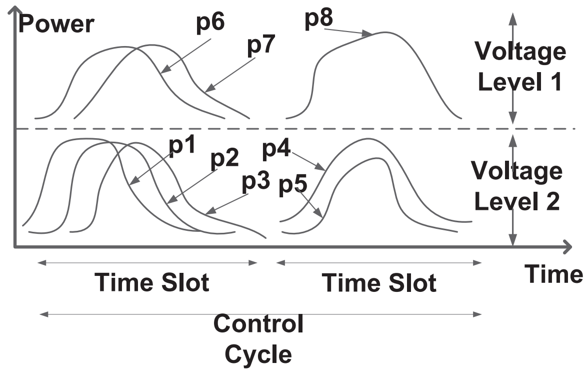

Figure 2: Grouping and subgrouping based on power levels and power phases

The formulation of groups and subgroups are illustrated in Fig. 2. At one control-cycle, power traces p_1, p_2, p_3, p_4 and p_5 are operating at one power magnitude level and other cores are working at a different power magnitude level. As such, one can form two groups with two voltage-levels v1 and v2. Inside the group supplied by voltage-level v1, one can observe that p_1, p_2, p_3, have a similar power phase compared to p_4 and p_5; so p_1, p_2 and p_3 further form a subgroup and p_4 along with p_5 forms another subgroup. The formed groups and subgroups can change at the next control-cycle.

In the following, we show that with the help of adaptive clustering, one can find the minimum number of power converters to satisfy the demand-supply matching. Moreover, by clustering, the complexity from the demand (power profiles) can be significantly reduced. As such, the large-scale demand-supply matching can be efficiently solved by the proposed two-step clustering in every control-cycle.

3.4 Solution to Subproblem 1

Once subgroups are formed, the maximum workloads of one subgroup can be determined. As such, the minimum number of power converters can be also determined to supply that subgroup. This results in one feasible solution to solve the Subproblem 1 in Section 2 as rephrased below.

$$\begin{aligned} \text{min:} \quad & \sum_{i=1}^{N_v} r_i \\ \text{s.t.:} \quad & \text{(i) } v_a(c_j) \geq v_d(c_j), \forall c_j \in C. \\ & \text{(ii) } d(r_i) \leq N_{max}, \forall r_i \in R. \end{aligned} \quad (6)$$

If one can determine the minimum number of power converters r_i for each group, the total number of power converters can be correspondingly minimized. Note that constraint (i) guarantees that the supplied voltage-level $v_a(c_j)$ from power converter will

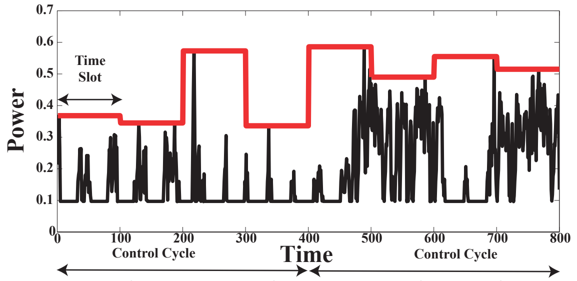

Figure 3: Peak power envelope extracted in each time-slot in one control cycle

satisfy the demanded voltage-level $v_d(c_j)$ from core c_j. Moreover, constraint (ii) imposes the driving ability $d(r_i)$ of each power converter is N_{max}, i.e., the maximum number of cores to drive. The driving ability can vary with the voltage-level: the higher the voltage-level is, the lower the number of cores that one power converter could drive.

Next, we show that the minimization of total number of power converters can be solved by grouping and subgrouping. By performing grouping, power converters are shared in space among N_v number of groups and subgrouping makes sharing of power converters inside a group in time. Based on the driving capability d_i^j of i-th power converter in group g_j, $g_j \in G$, having k subgroups, and the maximum number of cores among different subgroups, $max(c_i)$, $c_i \in C$, the minimum number of power converters for group g_j can be determined as

$$r_{g_j} = max(c_i)/d_i^j.$$

As such, for the whole system, the total number of power converters needed will be $\sum(r_{g_j})$, which is the minimum number to satisfy the demand-supply matching.

4. DEMAND RESPONSE BASED WORKLOAD SCHEDULING

This section deals with peak reduction and load balancing after the minimum number of power converters are allocated. A demand-response based workload scheduling will be developed towards uniform distribution of workload with reduction in peaks at one power converter.

4.1 Peak Power Envelope Extraction

To deal with peak reduction and load balancing, we first discuss the extraction of *peak power envelope* in one control-cycle, because it is impractical to perform power management in continuous form. Based on the extracted peak power envelope, one can build workload behavior model for each subgroup to be used in scheduling.

Assume that in one control-cycle T^i for the ith-group, g_i, $g_i \in G$ with N_k number of subgroups, each core is assigned with one workload. One can have *time slot* T_j^i, which is is the amount of time to finish all workloads in a subgroup, k_j, $k_j \in K$. Relation between T^i and T_j^i is

$$T^i = \sum_{j=1}^{N_j} T_j^i. \tag{7}$$

As such, in one time-slot T_j^i, peak power envelope Pe is extracted for workloads $p(t)$ of one subgroup by

$$Pe(T_j^i) = max(p(t)). \tag{8}$$

This is repeated for whole control cycle T^i. Thus peaks are extracted and one envelope is formed. Peak extraction by forming one envelope is shown in Fig. 3. The control-cycle T^i is 400ns with time-slot T_j^i of 100ns. At each time slot, the power envelope is formed on the peak value.

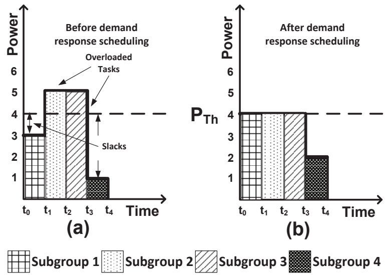

Figure 4: (a) Load before demand response scheduling (b) Peak reduction by demand response scheduling

4.2 Peak Reduction and Load Balancing

When the peak envelope of subgroup k_i, $k_i \in K$ is compared with one threshold power P_{th}^i of group g_i, the *slack* can be calculated by

$$a_j^i = Pe(T_j^i) - P_{th}^i. \tag{9}$$

If the value of slack is positive, then the allocated power converter, r_j, $r_j \in R$, is overloaded and not capable of handling extra load at that time-slot; otherwise, the power converter r_j is underloaded and can be allocated with additional workloads. After calculating the amount of slack, the workload of the power converter r_j can be rescheduled such that priority is not violated.

We call such a scheduling as *demand-response* based workload scheduling. The procedure for scheduling is described in Algorithm 2. It is deployed after clustering to decide the time slot. The first step in scheduling is to calculate the threshold and slack. Line 2-4 of Algorithm 2 explains the scheduling of task from a power converter that is overloaded and reduction of corresponding load. Similarly Line 6-8 describes adding of workloads on an underloaded power converter. In short, it can be viewed as re-clustering or refinement. The overhead includes the time to perform the calculation and movement, which is negligible in the whole control cycle.

Algorithm 2 Demand-response based workload scheduling

```
 1: INPUT: Initial set Workload $L$, Slack $A$
 2: if $a_j^i > 0$ then
 3:     Decrease workload on $r_j$
 4:     $l(r_j) - -$;
 5: else
 6:     while $a_j^i < 0$ do
 7:         Increase workload on $r_j$
 8:         $l(r_j) + +$;
 9:         $a_j^i + +$;
10:     end while
11: end if
```

Example in Fig. 4 shows the peaks of four subgroups. Before performing demand-response based workload scheduling, subgroup 2 and subgroup 3 are overloaded and subgroups 1 and 4 have slacks for scheduling. The peak value in subgroup 2 and 3 is 5, which means there are 5 peaks in those two subgroups. The peak power reduction is then achieved with the comparison of the highest value in subgroups before and after the demand-response scheduling. After the demand-response scheduling, the peak value will be reduced to 4. So, a 20% peak power reduction will be achieved.

4.3 Solution to Subproblem 2

The aforementioned demand-response based workload scheduling can be deployed to solve the Subproblem 2 addressed in Section 2, which is reformulated as

$$\begin{aligned} \text{min:} \quad & \sum_{j=1} |\sum_{i=1} s_j^i| \\ \text{s.t.:} \quad & Pe(T_j^i) < P_{th}^i \end{aligned} \tag{10}$$

Table 3: Clustering result for 64 cores

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Group 1	31, 37, 52 58, 59, 63	12, 49, 54	33, 43	7, 8, 14
Group 2	27, 29	17, 40, 41, 50 51, 56, 62	22, 42	N/A
Group 3	6, 21, 32 36, 39, 46 47, 64	9, 15, 16 20, 26 ,28 35, 53, 55	1, 5 11, 18 19, 38	N/A
Group 4	2, 3, 23, 25 34, 44, 45, 48 57, 60, 61	10, 13	N/A	4, 24, 30

Table 4: Comparison of number of allocated power converters under different PM schemes

		STM	SM	TM	STM/SM	STM/TM
32-core	Group 1	1	2	3	-50.00%	-66.67%
	Group 2	1	2	2	-50.00%	-50.00%
	Group 3	3	7	5	-57.14%	-40.00%
	Group 4	4	9	4	-55.56%	0.00%
	Total	9	20	14	-55.00%	-35.71%
64-core	Group 1	2	4	6	-50.00%	-66.67%
	Group 2	3	4	7	-25.00%	-57.14%
	Group 3	5	12	9	-58.33%	-44.44%
	Group 4	11	16	11	-31.25%	0.00%
	Total	21	36	33	-41.67%	-36.36%

Table 5: Comparison of peak power reduction and workload balancing by demand-response scheduling

	Peak Reduction	Balance before	Balance after
Group 1	33.33%	1.00	0.58 (1.72X)
Group 2	42.86%	1.00	0.50 (2X)
Group 3	33.33%	0.91	0.50 (1.82X)
Group 4	42.86%	0.63	0.13 (4.85X)
Average	38.10%	0.89	0.43 (2.60X)

Solution to this problem is to minimize the overall sum of slacks. This can be achieved by rescheduling workloads that demand power more than the threshold. So, initially peak reduction has to be performed followed by load balancing. Based on the value of slack for a subgroup k_j, if the slack is positive, then the workload on that subgroup needs to be delayed or advanced to other time-slot. As such, the workloads are allocated to subgroups with highly negative slack, and the differences in slack is reduced. As a result, peak reduction and load balancing can be achieved eventually.

5. SIMULATION RESULTS

5.1 System Modeling and Settings

The proposed system is validated by Matlab and system-level models built from SystemC-AMS. Table 2 summarizes the system design specifications. All units are scaled or modeled at CMOS 130nm CMOS process. The specification of low-power MIPS microprocessor core [10] is taken as the core model. Each core has the nominal frequency of 250MHz with the maximal power consumption of 0.4W. Benchmarks from SPEC2000 [11] are simulated by Wattch [12] to generate power profiles. The extracted power profiles are used as workload models, which are distributed to different cores randomly. The typical control cycles for power management is 400ns.

A 2-phase multi-output power converter [13] is designed to generate 4 different voltage-levels. As driving ability of power converter depends on supply voltage-level, driving abilities are set as 4, 3, 2, 1 for voltage-levels of 0.6V, 0.8V, 1.0V and 1.2V respectively. Moreover, the inductance value in power converter is set as $1nH$ per phase to support the maximum current on the buck inductor. Such an inductor requires an area of $0.25mm^2$, occupying 30% area of the power converter. The local super-capacitor for each core is set as $1\mu F$ to support time-multiplexing scheme between clusters. The design of on-chip power converter thereby needs to consider the limitation of inductor and capacitor area, which are both placed in 3D fashion and hence has the minimum area overhead to cores all on the other tier.

In addition, the vertical TSV [14] with size of $500\mu m^2$ works as connections between cores and power converters. According to the model in [15], it has a dc-resistance of $20m\Omega$. Considering the maximum current of $330mA$, the IR-drop of is around $7mV$, which is quite small. Note that the capacitor of TSV is in fF-scale and hence does not influence the load capacitance. What is more, for each TSV channel, one switch box is assigned with Nr power switches to support the core-converter connection. The switch box offers a compact reconfigurable unit driven by the controller. The power switch inside each switch box occupies $520\mu m^2$ and is able to deliver the maximum core current with switching time of 300ns. As such, the TSV coupling is also quite small to consider under such a slow power switching.

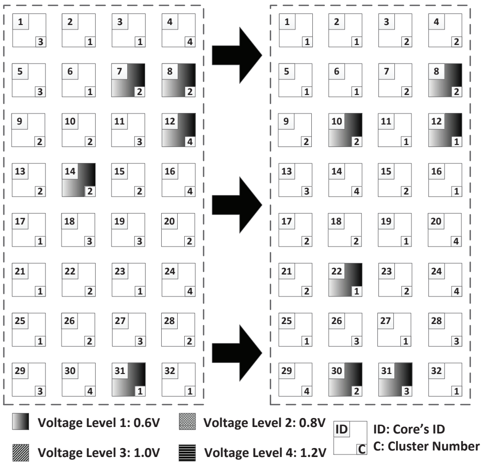

Figure 5: Results of adaptive clustering at two continuous control-cycles

5.2 Results and Comparisons

Firstly, we take 32-core and 64-core microprocessors as two examples to show results under adaptive clustering. The input power traces are first grouped into 4 based on the power magnitudes, then in each group subgroups are formed based on their power phases.

Fig. 5 illustrates the adaptive clustering result of 32-core between two consecutive control cycles. Different filling-shapes represent different groups or voltage-levels. Different clustering numbers on the downright-corner of cores represent different subgroups. For example, in the first control cycle, the 30th core will be assigned to subgroup 4 with voltage-level 4 (group 4). And in the next control cycle, it will be assigned to subgroup 2 with voltage-level 1 (group 1). For 64-core case, Table. 3 summarized the clustering results with the value in the table to represent the core ID. One can also observe that the runtime of clustering is small at the scale of 200ms.

Next, we use the space-time multiplexing (STM) scheme to perform the demand-supply matching. The first step is for resource allocation and adaptive clustering is deployed. After clustering, we extract simplified workload models to represent the peak power in one control cycle; and also determine the minimum number of power converters for each group. When comparing to two schemes, namely space-multiplexing (SM) and time-multiplexing (TM) with the same driving ability and time slot, the STM-based approach takes the advantage of both space and time to minimize the number of power converters. Table. 4 shows the comparison for 32-core and 64-core cases with the three schemes. One can observe that 55.00% (SM) and 35.71% (TM) number of power converters can be reduced for the case of 32-core, while 41.67% (SM) and 36.36% (TM) number of power converters can be reduced for the case of 64-core. Therefore, STM based adaptive clustering can satisfy the demand-supply matching with the minimum number of power converters to reduce the area overhead and also on-chip implementation cost.

Lastly, we perform demand-response based workload scheduling for time-multiplexing of power converters inside one

Table 2: System settings of 3D many-core microprocessors, on-chip power converters, TSVs and power switches

Item	Description	Symbol	Value	Size
Microprocessor	Performance	N.A.	410 DMIPS	$1.5mm^2$
	Frequency	f_c	250MHz	
	Power Consumption	P_c	0.4W	
Power Converter	Input Voltage	V_{in}	2.4V	$1.6mm^2$
	Output Voltage	V_{out}	0.6V, 0.8V, 1.0V, 1.2V	
	Load Current	I_L	120mA, 150mA, 220mA, 350mA	
	Number of Phases	N.A.	2	
	Inductor per Phase	L	1nH	
	Switching Frequency	f_s	50-200MHz	
	Peak Efficiency	N.A.	77%	
TSV	Length	l	$25\mu m$	$500\mu m^2$
	Diameter	W	$5\mu m$	
	Isolation Film	r	$120nm$	
	Resistance	R_{TSV}	$20m\Omega$	
	Capacitance	C_{TSV}	37 fF	
Power Switch	Width	w_s	4mm	$520\mu m^2$
	Length	l_s	$130nm$	
	Switching Time	N.A.	300ns	

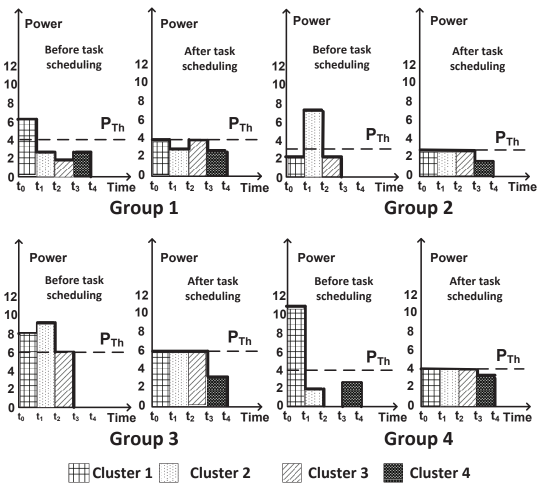

Figure 6: Peak power reductions for 4 subgroups of 64-core case

group. The peak power reduction is defined as the difference of peak power value before and after the scheduling. The workload balancing is defined as the number of cores which one power converter drives over control cycles. We compare the peak power reduction by averaging the reduction in each group; and compare workload balancing by averaging the standard-deviation (SD) of workload on each power converter. For a 64-core microprocessor results shown in Fig. 6, in Group 3, the peak power value has been reduced from 9 to 6 with 33.33% peak power reduction. The average standard deviation of workload on each power converter before and after scheduling are 0.91 and 0.50 respectively, with a standard deviation improvement by 1.82x. Table. 5 shows the summarized results for peak reduction and workload balancing by demand-response scheduling. One can observe an average of 38.10% peak power reduction and 2.60x workload balancing.

6. CONCLUSION

A space-time multiplexed power management is developed for large-scale demand-supply matching between on-chip power converters and many-core microprocessors. The power switch network is configured to perform space-time multiplexing between power converters and cores by vertical TSVs in 3D. Based on adaptive clustering of cores classified by both power magnitudes and power phases, the minimum number of power converters are allocated to supply the demanded voltage-levels from cores. What is more, demand-response based workload scheduling is deployed by utilizing the power slacks, such that peak power can be reduced as well as workload can be balanced. As verified by system-level behavior models implemented in SystemC and SystemAMS, and also physical-level models with design parameters, experiment results show that the space-time multiplexing can reduce peak power by 38.10% and improve load balancing by 2.60x improvement on average with the minimum number of allocated power converters.

7. ACKNOWLEDGMENTS

This work is sponsored by Singapore MOE TIER-2 fund MOE2010-T2-2-037 (ARC 5/11) and A*STAR SERC-PSF fund 11201202015. Please address comments to haoyu@ntu.edu.sg.

8. REFERENCES

[1] S. Vangal and et.al., “An 80-Tile 1.28TFLOPS network-on-chip in 65 nm CMOS,” in *IEEE ISSCC*, 2007.

[2] S. Bell and et.al., “TILE64^{TM} processor: a 64-core SoC with mesh interconnect,” in *IEEE ISSCC*, 2008.

[3] M. Healy and et.al., “Design and analysis of 3D-MAPS: a many-core 3D processor with stacked memory,” in *IEEE CICC*, 2010.

[4] H. Yu, J. Ho, and L. He, “Allocating power ground vias in 3d ics for simultaneous power and thermal integrity,” *ACM TODAES*, vol. 14, no. 3, 2011.

[5] W. Kim and et.al., “System level analysis of fast, per-core DVFS using on-chip switching regulators,” in *IEEE HPCA*, 2008.

[6] R. Bondade and D. Ma, “Hardware-software codesign of an embedded multiple-supply power management unit for multicore SoCs using an adaptive global/local power allocation and processing scheme,” *ACM TODAES*, vol. 16, no. 3, 2011.

[7] J. Howard and et. al, “A 48-core ia-32 processor in 45 nm cmos using on-die message-passing and dvfs for performance and power scaling,” *IEEE JSSC*, vol. 46, pp. 173–183, January 2011.

[8] N. Sturcken and et.al., “A 2.5D integrated voltage regulator using coupled-magnetic-core inductors on silicon interposer delivering 10.8A/mm^2,” in *IEEE ISSCC*, 2012.

[9] R. H. Katz and et. al, “An information-centric energy infrastructure: The berkley view,” *Sustainable Computing: Informatics and Systems*, no. 1, pp. 7–22, March 2011.

[10] “MIPS processor cores,” http://www.mips.com/products/processor-cores/.

[11] “SPEC 2000 CPU benchmark suits,” http://www.spec.org/cpu/.

[12] “Wattch version 1.02,” http://www.eecs.harvard.edu/˜dbrooks/wattch-form.html.

[13] W. Kim and et.al., “A fully-integrated 3-level DC/DC converter for nanosecond-scale DVS with fast shunt regulation,” in *IEEE ISSCC*, 2011.

[14] V. der Plas and et.al., “Design issues and considerations for low-cost 3D TSV IC technology,” in *IEEE ISSCC*, 2010.

[15] G. Katti and et.al., “Electrical modeling and characterization of through silicon via for three-dimensional ICs,” *IEEE Trans. on Electron Devices*, vol. 57, no. 1, pp. 256–262, 2010.