Measurement-Driven Methodology for Evaluating Processor Heterogeneity Options for Power-Performance Efficiency

William J. Song^{*†}, Alper Buyuktosunoglu[‡], Chen-Yong Cher[‡], and Pradip Bose[‡]

[†]School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332

[‡]IBM T.J. Watson Research Center, Yorktown Heights, NY 10598

wjhsong@gatech.edu, {alperb, chenyon, pbose}@us.ibm.com

ABSTRACT

It is generally perceived that heterogeneous multicore processors will provide better performance and power efficiency over conventional homogeneous cores. However, heterogeneity can also be achieved within a homogeneous core design, instantiated under different voltage-frequency settings or per-core simultaneous multi-threading (SMT) modes. In this paper, we pursue an architectural study motivated by the question, “*Can we get by with a single, complex SMT-equipped core design that can operate at different voltage-frequency points? Or, is it mandatory to invest into two different core types, one complex and the other simple?*” We propose a systematic, measurement-driven methodology to evaluate processor heterogeneity options. Our analysis particularly focuses on the domain of real-time constrained embedded processors. The study is based on a direct measurement of two real processors; one that uses simple in-order cores, and another that uses complex out-of-order cores. The effect of heterogeneous core composition (consisting of complex and simple cores in the same chip) is analytically projected from measurements gleaned from the two different systems. Our analysis yields new interesting insights. When dealing with two core types without SMT enabled, true core heterogeneity does not necessarily provide better performance or power efficiency under area and power constraints. If the complex-core homogeneous processor invokes SMT, it outperforms true heterogeneity by offering 28% better power efficiency, assuming that simple cores in the heterogeneous system operate only in single-threaded mode without SMT capability. If the small cores employ SMT, true heterogeneity yields 32% better power efficiency than the homogeneous processor with SMT.

1. INTRODUCTION

Power-performance efficiency is a key metric in evaluating processor chips in the power-constrained design era. Heterogeneous multicore processors have been suggested as a microarchitectural

choice to boost efficiency measured in throughput-per-watt. By dynamically matching the computational requirements of workloads to different core types, heterogeneous multicore processors could potentially increase power efficiency and performance over homogeneous alternatives [5, 6, 7, 10, 11, 16]. Specifically in the domain of *real-time constrained embedded processing* [2], both execution time latency and throughput-per-watt are important metrics for optimization. In such a scenario, it is easy to understand the rationale behind a heterogeneous multicore design, where complex cores (CC) provide sequential heavy-lifting to improve net execution latency while simple cores (SC) provide power-efficient parallel throughput performance.

However, a key question facing future design teams is: “*Can we get by with a single complex core design and provide the desired heterogeneity by invoking currently available knobs such as voltage-frequency scaling and simultaneous multi-threading (SMT)?*” The obvious limitation of the complex homogeneous solution is that the area of a wide-issue, out-of-order core is typically 3-5 times larger than a narrow-issue, in-order design. With the area constraint, the homogeneous design includes fewer cores. This limits parallel execution speed-up. However, SMT is a commonly found feature in server-class cores and known to boost throughput efficiency. For instance, IBM POWER7 and POWER7+ processors support 4-way SMT per core [8, 17]. By utilizing 4-way SMT, the complex cores in homogeneous design could ostensibly provide as much parallel speed-up as the simple cores of heterogeneous processor without the SMT feature.

Therefore, based on the logic above, it appears possible for the homogeneous processor (supported by aggressive voltage scaling and wide SMT) to provide comparable (or even better) power efficiency and performance relative to the truly heterogeneous core solution. However, without an in-depth analysis of performance and power efficiency as a function of voltage-frequency operational points and degree of multi-threading, such an expectation cannot be validated. In this paper, motivated by the above research question, we make the following contributions:

- We develop an accurate, measurement-driven characterization and modeling methodology to investigate various tradeoffs in examining available processor heterogeneity options. Measurement is conducted on commercial homogeneous core platforms, and heterogeneous options are analytically derived. Arguably, this novel analysis methodology and derived conclusions are more credible than simulation-based analysis in prior work. Related work did not consider voltage-frequency scaling, core composition, and SMT altogether especially in the context of real-time constrained embedded computing.

^{*}This work was done while William J. Song was an intern at IBM T.J. Watson Research Center.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](http://permissions.acm.org).

ISLPED '16, August 08-10, 2016 San Francisco Airport, CA, USA

© 2016 ACM. ISBN 978-1-4503-4185-1/16/08...\$15.00

DOI: <http://dx.doi.org/10.1145/2934583.2934637>

- We establish power versus performance characteristics as a function of core complexity, core count, voltage-frequency operational points and SMT-mode functionality. Two real cores are measured in this study, i) the in-order A2 core in Blue Gene/Q system [9] that serves as a simple core (SC) and ii) the out-of-order core in POWER7+ system [17] that serves as a complex core (CC). The measurement results of these cores are projected towards embedded systems.
- We consider a suite of applications particularly relevant to embedded processing [2]. We analytically project comparative performance and efficiency metrics to understand conditions under which an investment into just a single, complex homogeneous core may be sufficient and where it is probably not sufficient. In particular, we arrive at the following conclusions:
 - Without the benefit of SMT, true core heterogeneity under both power and area constraints does not necessarily provide better performance or power efficiency.
 - When only the CC is equipped with 4-way SMT, the homogeneous solution wins over the true heterogeneity by 28% when it comes to power efficiency.
 - If the SC is also equipped with SMT, the heterogeneous solution beats the homogeneous design in both power efficiency and performance aspects, regardless of whether the CC operates in SMT mode or not.

2. RELATED WORK

Heterogeneous processors have been suggested and studied to tackle challenges in performance growth bounded by power limitations. Kumar et al. [13] showed that matching the computing requirement of applications to core types could enhance power efficiency. Hill and Marty [10] presented the theoretical performance models of heterogeneous multicores based on Amdahl's Law. Following this work, Woo and Lee [16] showed the power scaling models of heterogeneous processors and evaluated energy efficiency metrics. Chung et al. [6] explored heterogeneous designs comprised of unconventional cores such as custom logics, FPGA, and GPGPUs. Esmailzadeh et al. [7] presented the projection of power-constrained multicore processors at future device technologies in the view of dark silicon. Koufaty et al. [12] and Li et al. [14] studied required OS-level support for heterogeneous processors since conventional OS does not discern core heterogeneity. Joao et al. [11] presented methods to identify critical threads in parallel executions and execute them on complex cores to speed up the overall execution.

Prior work studied heterogeneous multicores based on analytical projections of performance and power models. In contrast, our study is based on the accurate hardware measurement of real processors. Cao et al. [5] conducted similar measurement-based analysis to support virtual machine services in heterogeneous processors. In particular, their target model was comprised of one complex core and two simple cores. This one case study is insufficient to answer the question whether heterogeneous processors will guarantee the superiority over conventional homogeneous processors, particularly in the domain of embedded systems. In our analysis, we measure the Blue Gene/Q processor that includes 16 user cores (and 2 additional cores for system operations and spare, respectively) [9] and POWER7+ processor comprised of 8 complex cores with 4-way SMT per core [17]. Such hardware availability enables us to explore more diverse composition of heterogeneous processors and compare them with the existing homogeneous platforms.

3. EVALUATION METHODOLOGY

Our evaluation methodology is driven by direct hardware measurement of A2 in-order cores of the Blue Gene/Q processor [9] and out-of-order cores of the POWER7+ processor [17]. They serve as simple core (SC) and complex core (CC) exemplary embodiments in this study towards configuring a real-time constrained embedded processor.

Selected DARPA PERFECT [2] and PARSEC [4] benchmarks serve as exemplary embedded applications and are measured on the processors at multiple voltage-frequency operating points with a varying number of activated cores and SMT threads. Heterogeneous processor options are constructed by varying the number of CC and SC-type cores within the same area as a homogeneous multi-CC processor [17]. The results are compared in performance and power efficiency aspects measured by execution time and Giga operations per second/Watt (GOPS/W), respectively. We also develop and use an analytical model to project power-performance efficiency of viable heterogeneous core solutions (see Section 4).

3.1 Test Hardware: CC versus SC Cores

A Blue Gene/Q chip is comprised of 18 A2 in-order cores [9] that serve as representative SC of choice in this study. They are connected via a crossbar and operate at 1.6GHz clock frequency. User applications can execute on 16 cores, and the remaining two cores are reserved for system operations and spare, respectively [9]. We used libBGQT library [3] built in the Blue Gene/Q processor to probe core-logic power and collect performance counters. A POWER7+ processor chip has 8 out-of-order cores that are used as our exemplary CC, and it supports the dynamic voltage and frequency scaling (DVFS) range of 2.3-4.3GHz [17]. We used Cornflake tool [15] to probe the processor and collect power, voltage, and performance counters.

When measuring the power of reference SC or CC-based multi-core processors, we first monitored idle state power. This was subtracted from active state power to calculate dynamic power for any given application run. We delimited the applications by multiple phases, based on their execution behaviors (e.g., serial or parallel) and algorithm types (e.g., row or column FFT). Single core power was estimated by the sum of i) idle state power divided by the number of cores in the processor and ii) dynamic power divided by the number of activated cores, as expressed in Eq. (1):

$$P_{core} = \frac{P_{idle}}{\text{Total \# of cores}} + \frac{P_{active} - P_{idle}}{\text{\# of activated cores}} \quad (1)$$

3.2 Technology Scaling Factors

Our chosen SC and CC components belong to processors fabricated in different technology nodes, 45nm and 32nm CMOS SOI, respectively. We empirically obtained power and area scaling factors based on the measurement of a real 45nm processor chip [8] and its 32nm re-mapped product [17]. The obtained power and area scaling factors were applied to the measured data of our chosen SC to normalize the results at the same technology node.

In addition, the SC and CC multicore processors used in the experiment have disjoint operating voltage and frequency ranges. To account for this difference, we measured the real CC processor at multiple DVFS points and extrapolated the results at lower frequency points, as plotted in Figure 1. This example is shown with the Inner-Product benchmark from DARPA PERFECT suite [2]. This extrapolation process was individually applied to all measured benchmarks. It is assumed that out-of-order cores could be designed to operate at the same voltage and frequency as SC-class cores. In fact, the CC-type processors are optimized for high-performance environment and generally shipped with higher V_{min}

Figure 1. Execution time and power of the exemplary CC are measured at 2.4-4.0GHz DVFS points. Execution time and power in low frequency ranges are extrapolated from the measured data at higher frequency points. Power with V_{min} denotes the extrapolation with a voltage value at 2.4GHz.

than the SC-class cores that are optimized for low-power environments. If the CC-type processor cannot be designed to operate at low voltages (e.g., due to SRAM delay limitation), its power consumption in low frequency ranges is bounded by the fixed voltage that corresponds to 2.4GHz clock frequency, as shown in Figure 1. Our analysis in subsequent sections is based on the comparison between the technologically-scaled results of SC and analytically-extended data of CC-based processors.

3.3 Multi-Threaded Test Benchmarks

Test benchmarks listed in Table 1 were executed in parallel mode using pthread. Each benchmark was measured separately on reference SC and CC-class processors by varying the number of activated cores, voltage and frequency levels, and SMT modes. Since the measurement was conducted on the real hardware, parallelization overhead (e.g., barriers, locks) was already included in the measured data. At the boundaries of execution phases (e.g., applications switching from serial to parallel operations or vice versa), we observed that the transition was too quick to produce meaningful data samples even at finer sampling rates. Hence, we disregarded the phase transition overhead in the analysis. However, we could observe unbalanced execution time between parallel threads (e.g., due to barriers), and the slowest thread was used to represent the execution time of a given execution phase.

Each benchmark was partitioned into multiple phases depending on operation types and algorithms, since the workload showed distinct performance and power behaviors across the phases. The initial and final phases of benchmarks were mostly file handling operations (i.e., reading input files and writing results to disk). These phases were highly dependent on off-chip configurations rather than capturing the computational characteristics of measured cores. Therefore, we excluded these I/O phases in our study and used the measurement data of computational phases of test benchmarks.

SMT is known as a power efficiency enhancer, generally supported in CC-type cores. For instance, IBM POWER7 and POWER7+ processors support 4-way SMT per core [8, 17]. To understand the effect of SMT, thread affinity mode was added to each benchmark to precisely locate hardware threads across cores. Selected PERFECT [2] and PARSEC [4] benchmarks were used in our analysis. DARPA PERFECT suite is comprised of conventional signal processing workloads including FFT, convolution, Cholesky factorization, etc (see Table 1). We selected only a few benchmarks from PARSEC suite since thread counts were not controllable within execution phases in many other benchmarks. Ex-

Table 1. DARPA PERFECT [2] and PARSEC [4] Suites

Suite	Abbreviation	Benchmark Algorithm
PERFECT	2D-Cnv	2D Convolution
	DWT53	Discrete Wavelet Transform
	Hst-Eq	Histogram Equalization
	FFT-R	Row-wise 2D FFT
	FFT-C	Column-wise 2D FFT
	In-Prd	Inner Product
	Choles	Cholesky Factorization
	FB-Sub	Forward/Back Substitution
	Bck-Pr	Back Projection
	Intrp1	Interpolation Algorithm #1
	Intrp2	Interpolation Algorithm #2
PARSEC	Black	Blackscholes
	Stream	Streamcluster
	Swap	Swaptions

Table 2. Processor Configurations Under Area Constraint

Configuration	# of CC cores	# of SC cores
Heterogeneous 1+14	1	14
Heterogeneous 2+9	2	9
Heterogeneous 3+5	3	5
Homogeneous 4+0	4	N/A

cluded benchmarks created the designated number of threads (e.g., $n_threads=4$) per algorithm, but algorithms often overlapped over time. As a result, these benchmarks generated varying number of threads, and the locations of hardware threads were not manageable via pthread affinity mode to accurately measure the effect of SMT.

4. ILLUSTRATIVE ANALYSIS RESULTS

Heterogeneous and homogeneous processors contrast in two primary aspects; efficiency and complexity. The heterogeneous processors are known to provide better power efficiency and performance than conventional homogeneous processors. However, the improvements are traded with greater complexity in the processor design and scheduling method, which may increase development costs. In this section, we analyze the effects of i) voltage scaling (lower-voltage operations) and ii) SMT to answer the question addressed in the introduction of this paper. The homogeneous CC-based processor with these knobs is compared to the hypothetical heterogeneous processor comprised of CC and SC-type cores.

Prior work suggested various thread scheduling methods in heterogeneous processors. For instance, CC-class cores can be utilized along with SC-class cores during parallel executions to maximize throughput [11]. Alternatively, only SC-type cores can be used to run parallel threads to simplify thread scheduling and limit the total power [7]. In this paper, we assume the first case that maximizes throughput but constrain the total power such that heterogeneous processors dissipate similar power as the homogeneous processor for given applications. Thus, it may require the heterogeneous processors to turn off a few cores to meet the power limitation depending on workload power behaviors.

4.1 Multicore Processor Configurations

We used a 4-core CC-based processor as a reference homogeneous multicore model. Considering $4.7\times$ area difference between our chosen CC and SC-class cores (normalized at 32nm technology node) [8, 9, 17], utilizing all 8 cores in the CC-based reference

Figure 2. Normalized (a) execution time and (b) GOPS/W comparison between homogeneous processor and heterogeneous processor options at 1.6GHz with single thread per core. The results contrast maximum difference and do not include sequential operations that are identical in both homogeneous and heterogeneous processors. The result of each benchmark is normalized to the 4-core execution of CC-based reference processor at 4.0GHz.

processor requires up to 33 SCs in a heterogeneous processor (i.e., 1 CC and 33 SCs). Since our SC-based reference processor chip [9] only accommodates 16 user cores, we limited the size of homogeneous CC-based processor to a 4-core model. Then, it required only up to 14 SC in the heterogeneous processor as listed in Table 2, which were measurable from the SC-based reference processor [9]. If the heterogeneous processor incorporates increasing number of CCs, there are less number of SC-class cores on the die under the area constraint. When executing only one application at a time, one CC is sufficient to handle the sequential fraction of the application. However, in a generalized situation such as multiplexing applications and system operations (i.e., virtual environment), there can be a need for including multiple CCs to handle multiple, concurrent sequential operations. Hence, we consider the case in which the heterogeneous processor deploys multiple CC-type cores, and we analyze how much degradation in performance and power efficiency will occur in the heterogeneous processor with the area and power constraints. A power constraint is applied such that the heterogeneous processor dissipates similar power as (but not greater than) the homogeneous processor for each test benchmark. For instance, a heterogeneous option with 14 SC and 1 CC dissipates larger power when all cores are active than the 4-core CC-based homogeneous processor. Depending on workload power behaviors, the heterogeneous processor could use only 10-12 SC to meet the power limitation.

4.2 Benchmark Measurement Results

Our assumption in both homogeneous and heterogeneous processors is that a CC-class core executes the sequential part of an application. Increasing the fraction of sequential operations diminish differences between the two processors, which can be expressed by Amdahl's Law [1, 10, 16]. The key difference between two processor types originates from handling the parallelizable portion of workloads. Therefore, in the remaining part of the paper, we focus on the maximum performance and power efficiency difference between the homogeneous and heterogeneous processors. We as-

sume that unused cores such as those during sequential executions are ideally power-gated and do not contribute to the total power. Figure 2 shows the performance (in execution time) and power efficiency (in GOPS/W) of individual benchmarks at 1.6GHz clock frequency with single thread per core. On average, the best case of heterogeneous processor without power limitation (i.e., activating 14 SC and 1 CC on the die) has 32% shorter execution time (not shown in the figure) but is 16% slower when the power constraint is applied, compared to the homogeneous processor. In the former case, it is assumed that the processor can ideally offload the tasks of SC-type cores to CC to maximize the performance for each benchmark. The latter case requires the heterogeneous processor to turn off a few cores to meet the power limitation, which lowers the overall performance. After the power constraint is applied, the heterogeneous processor comprised of 5 SC and 3 CC performs the best. The power-limited heterogeneous processors achieve 1-20% better power efficiency than the homogeneous processor. When the effect of V_{min} (e.g. 1.6GHz operation of CC-class core at the voltage corresponding to 2.4GHz) is considered, the heterogeneous processor options provided 29-53% better power efficiency than the homogeneous alternative.

4.3 Lower Voltage Operations

Figure 3 shows an execution time versus GOPS/W chart. An individual point is the average of benchmarks for each configuration with single thread per core at 1.6GHz clock frequency. The dashed line connects the results of homogeneous processor between 0.8-4.0GHz. The graph shows that performance and power efficiency are inversely proportional when the DVFS is employed. With aggressive scaling to low voltage points, the homogeneous processor achieves better power efficiency, but the power efficiency enhancement is traded with performance degradation. Therefore, the use of voltage scaling alone still leaves the gap between the homogeneous and heterogeneous processors. When V_{min} effect is included, the improvement of power efficiency becomes bounded, leaving wider gap between the two processor types. The V_{min} affects only the

Figure 3. Execution time vs GOPS/W of homogeneous and heterogeneous processors at 1.6GHz. Individual point is the average of measured benchmarks for each configuration. The dashed line connects the points of homogeneous processor between 0.8-4.0GHz with single thread per core, and the branch line is showing the V_{min} effect when operating under 2.4GHz. The results (each axis of the graph) are normalized to the 4-core CC-based reference processor at 4.0GHz.

Figure 4. Execution time vs GOPS/W between the homogeneous processor with 4-way SMT per core and heterogeneous processors without SMT support in the SC cores. The dashed line connects the points of homogeneous processor between 0.8-4.0GHz with 4-way SMT per core, and the branch line shows the V_{min} effect under 2.4GHz operations.

power efficiency not performance, so the data points of branch line in Figure 3 have the same execution time as those corresponding to the non- V_{min} line.

4.4 Simultaneous Multi-Threading

SMT is known as a power-efficiency enhancer. By scheduling multiple threads onto the same core and sharing resources, increased pipeline utilization leads to power efficiency improvement. SMT is a commonly found feature in CC-type cores. A POWER7+ processor, which serves as our CC model, supports 4-way SMT per core. We analyze how activating SMT in the CC affects net power efficiency and performance. Figure 4 shows execution time versus GOPS/W chart when 4-way SMT is enabled in the CC-based homogeneous processor. By activating SMT, 4 times more threads are used for parallel executions. The data points of heterogeneous processors are identical to those in Figure 3 without SMT in the SC. We note that it is uncommon for SC-type cores to support SMT (especially in real-time domain since it breaks the determinism in execution time), although they can be seen to support SMT for power-efficient, throughput-oriented environments [9].

Figure 4 shows that activating SMT significantly improves both power efficiency and performance of the conventional homogeneous core solution. Power efficiency in GOPS/W increases by 48-55%, and performance improves by 42-47%. As a result, the homogeneous processor achieves 28% better power efficiency or 51% shorter execution time than the best cases of the heterogeneous pro-

Figure 5. Execution time vs GOPS/W between homogeneous and heterogeneous processors both with 4-way SMT per core. The dashed line connects the points of homogeneous processor between 0.8-4.0GHz with 4-way SMT/core, and the branch line shows V_{min} effect. The results are normalized to the 4-core CC-based reference processor at 4.0GHz.

cessor for each metric at the same 1.6GHz clock frequency. When V_{min} is considered (the branch curve of 4-SMT dashed line in Figure 4), it still achieves better performance as well as power efficiency than the heterogeneous options.

In our case, we are fortunate to have access to an SC-class hardware that also supports 4-way SMT. We activated 4-way SMT in the same way as the homogeneous CC-based processor to understand how the heterogeneous processor would perform. By executing 4 times more threads on the same number of cores, SC-class cores could achieve more performance speed-up than the CC-type cores. We observed that the out-of-order pipeline of a CC-class core already well-utilized instruction-level parallelism with single-thread executions, so adding more threads via SMT moderately increased the total execution time. On the other hand, an SC-class core suffers from frequent pipeline stalls due to in-order execution, so scheduling multiple threads via SMT effectively enhances core utilization and has relatively smaller increases in total execution time. Figure 5 shows execution time versus GOPS/W graph when both homogeneous and heterogeneous processors support 4-way SMT. The data point and trend line of the homogeneous processor are identical to those in Figure 4. This graph shows that the heterogeneous options generally offer better throughput-per-watt than the CC-based homogeneous baseline when both processors are equipped with SMT. Overall, if one considers both execution time and throughput-per-watt, the Heterogeneous 3+5 configuration (in Table 2) performs the best. In particular, it has 12% shorter execution time than the CC-based homogeneous multicore operating at the same frequency, and it is also 32% better when it comes to throughput-per-watt.

5. CONCLUSION

In this paper, we presented a measurement-driven modeling methodology to investigate heterogeneous core options in optimizing power-performance efficiency. We directly measured performance and power characteristics (for target workloads) on two different homogeneous core platforms, one with complex (wide-issue, out-of-order) cores and the other with simple (narrow-issue, in-order) cores. DVFS and SMT knobs were both varied to derive parametric sensitivities. Using these experimental characterization data, we applied a customized analytical modeling methodology to project the relative benefits of different core heterogeneity options. This approach enables system architecture teams to devise the appropriate level of core composition with DVFS and SMT in defining the next generation heterogeneous core machines. The best achievable performance and power efficiency levels can thus be investi-

gated before the pre-RTL heterogeneous chip microarchitecture is frozen. Our novel modeling methodology is more credible than pre-RTL heterogeneous core power-performance simulators since the core axioms of the model are directly measured on real commercial systems (i.e., IBM POWER7+ and Blue Gene/Q). Also, full application workloads can be measured for CC and SC, as opposed to drastically sampled ones on slow many-core simulators.

The particular study reported in this paper yields a set of conclusions that would help future embedded system designers make better decisions depending on their available core library and optimization priorities. The observations based on our CC and SC choices are summarized as follows:

- True heterogeneity with power and area constraints does not necessarily produce better performance or power efficiency relative to a complex-core homogeneous solution.
- SMT functionality in a complex core provides a significant lever when it comes to throughput and efficiency boost. If simple cores in the heterogeneous processor lack SMT capability, the homogeneous solution with 4-way SMT per core and lower operating voltages would provide a superior option over the true heterogeneity.
- When simple cores within the heterogeneous multicore are also equipped with SMT, the heterogeneous solution offers 32% better power efficiency.

While the specifics within the conclusion may vary with particular CC or SC choices, the methodology is robust to deduce key pre-RTL parameters with the goal of maximizing power-performance efficiency of the end products.

6. ACKNOWLEDGEMENT

This work was sponsored by Defense Advanced Research Projects Agency (DARPA) under contract #HR0011-13-C-0022. The view expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

7. REFERENCES

- [1] G. Amdahl, "Validity of the single processor approach to achieving large scale computing capability," *AFIPS Spring Joint Computer Conference*, pp. 483-485, 1967.
- [2] K. Barker, T. Benson, D. Campbell, D. Ediger, R. Gioiosa, A. Hoisie, D. Kerbyson, J. Manzano, A. Marquez, L. Song, N. Tallent, and A. Tumeo, "Power efficiency revolution for embedded computing technologies (PERFECT) benchmark suite manual," *Pacific Northwest National Labs (PNNL) and Georgia Tech Research Institute (GTRI)*, Dec. 2013. <http://hpc.pnnl.gov/projects/PERFECT/>
- [3] R. Bertran, Y. Sugawara, H. Jacobson, A. Buyuktosunoglu, and P. Bose, "Application-level power and performance characterization and optimization on IBM Blue Gene/Q systems," *IBM Journal of Research and Development*, vol. 57, no. 1, pp. 4:1-4:17, Mar. 2013.
- [4] C. Bienia, S. Kumar, J. Singh, and K. Li, "The PARSEC benchmark suite: characterization and architectural implications," *International Conference on Parallel Architectures and Compilation Techniques*, pp. 72-81, Oct. 2008.
- [5] T. Cao, S. Blackburn, T. Gao, and K. McKinley, "The yin and yang of power and performance for asymmetric hardware and managed software," *ACM/IEEE International Symposium on Computer Architecture*, pp. 225-236, Jun. 2012.
- [6] E. Chung, P. Milder, J. Hoe, and K. Mai, "Single-chip heterogeneous computing: does the future include custom logic, FPGAs, and GPGUs?," *IEEE/ACM International Symposium on Microarchitecture*, pp. 225-236, Dec. 2010.
- [7] H. Esmaeilzadeh, E. Blem, R. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," *ACM/IEEE International Symposium on Computer Architecture*, pp. 365-376, Jun. 2011.
- [8] R. Kalla, B. Sinharoy, W. Starke, and M. Floyd, "POWER7: IBM's next-generation server processor," *IEEE Micro*, vol. 30, no. 2, pp. 7-15, Apr. 2010.
- [9] R. Haring, M. Ohmacht, T. Fox, M. Gschwind, D. Satterfield, K. Sugavanam, P. Coteus, P. Heidelberger, M. Blumrich, R. Wisniewski, A. Gara, G. Chiu, P. Boyle, N. Christ, and C. Kim, "The IBM Blue Gene/Q compute chip," *IEEE Micro*, vol. 32, no. 2, pp. 48-60, Dec. 2011.
- [10] M. Hill and M. Marty, "Amdahl's Law in the multicore era," *Computer Journal*, vol. 41, no. 7, pp. 33-38, Jul. 2008.
- [11] J. Joao, M. Suleman, O. Mutlu, and Y. Patt, "Utility-based acceleration of multithreaded applications on asymmetric CMPs," *ACM/IEEE International Symposium on Computer Architecture*, pp. 154-165, Jun. 2013.
- [12] D. Koufaty, D. Reddy, and S. Hahn, "Bias scheduling in heterogeneous multi-core architectures," *EuroSys Conference*, pp. 125-138, Apr. 2010.
- [13] R. Kumar, D. Tullsen, N. Jouppi, and P. Ranganathan, "Heterogeneous chip multiprocessors," *IEEE Micro*, vol. 38, no. 11, pp. 32-38, Nov. 2005.
- [14] T. Li, P. Brett, R. Knauerhase, D. Koufaty, D. Reddy, and S. Hahn, "Operating system support for overlapping-ISA heterogeneous multi-core architectures," *IEEE Symposium on High Performance Computer Architecture*, pp. 1-12, Feb. 2010.
- [15] A. Vega, A. Buyuktosunoglu, H. Hanson, P. Bose, and S. Ramani, "Crank it up or dial it down: coordinated multiprocessor frequency and folding control," *IEEE/ACM International Symposium on Microarchitecture*, pp. 210-221, Dec. 2013.
- [16] D. Woo and H. Lee, "Extending Amdahl's Law for energy-efficient computing in the many-core era," *Computer Journal*, vol. 41, no. 12, pp. 24-31, Dec. 2008.
- [17] V. Zyuban, S. Taylor, B. Christensen, A. Hall, C. Gonzalez, J. Friedrich, F. Clougherty, J. Tetzloff, and R. Rao, "IBM POWER7+ design for higher frequency at fixed power," *IBM Journal of Research and Development*, vol. 57, no. 6, pp. 1:1-1:18, Dec. 2013.