A Survey of CPU-GPU Heterogeneous Computing Techniques

SPARSH MITTAL, Oak Ridge National Laboratory
JEFFREY S. VETTER, Oak Ridge National Laboratory and Georgia Tech

As both CPUs and GPUs become employed in a wide range of applications, it has been acknowledged that both of these Processing Units (PUs) have their unique features and strengths and hence, CPU-GPU collaboration is inevitable to achieve high-performance computing. This has motivated a significant amount of research on heterogeneous computing techniques, along with the design of CPU-GPU fused chips and petascale heterogeneous supercomputers. In this article, we survey Heterogeneous Computing Techniques (HCTs) such as workload partitioning that enable utilizing both CPUs and GPUs to improve performance and/or energy efficiency. We review heterogeneous computing approaches at runtime, algorithm, programming, compiler, and application levels. Further, we review both discrete and fused CPU-GPU systems and discuss benchmark suites designed for evaluating Heterogeneous Computing Systems (HCSs). We believe that this article will provide insights into the workings and scope of applications of HCTs to researchers and motivate them to further harness the computational powers of CPUs and GPUs to achieve the goal of exascale performance.

Categories and Subject Descriptors: A.1 [**General Literature**]: Introductory and Survey; I.3.1 [**Computer Graphics**]: Graphics Processor; C.1.3 [**Other Architecture Styles**]: Heterogeneous (hybrid) Systems; H.3.4 [**Systems and Software**]: Performance Evaluation (Efficiency and Effectiveness); C.0 [**Computer Systems Organization**]: System Architectures

General Terms: Experimentation, Management, Measurement, Performance, Analysis

Additional Key Words and Phrases: CPU-GPU heterogeneous/hybrid/collaborative computing, workload division/partitioning, dynamic/static load balancing, pipelining, programming frameworks, fused CPU-GPU chip

ACM Reference Format:
Sparsh Mittal and Jeffrey S. Vetter. 2015. A survey of CPU-GPU heterogeneous computing techniques. ACM Comput. Surv. 47, 4, Article 69 (July 2015), 35 pages.
DOI: http://dx.doi.org/10.1145/2788396

1. INTRODUCTION

Computer architects, programmers, and researchers are now moving away from the CPU *versus* GPU debate [Gregg and Hazelwood 2011; Lee et al. 2010; Mittal and Vetter 2015; Vuduc et al. 2010] toward a CPU *and* GPU paradigm where the best features of both can be intelligently combined to achieve even further computational gains. This paradigm, known as Heterogeneous Computing (HC), aims to match the requirements of each application to the strengths of CPU/GPU architectures and also achieve load balancing by avoiding idle time for both the Processing Units (PUs). The importance of heterogeneous computing can be seen from the fact that an astonishingly large fraction

Support for this work was provided by U.S. Department of Energy, Office of Science, Advanced Scientific Computing Research.
Authors' address: S. Mittal and J. S. Vetter, 1 Bethel Valley Road, Oak Ridge National Laboratory, Building 5100, MS-6173, Tennessee, USA, 37830; emails: {mittals, vetter}@ornl.gov.

DOI: http://dx.doi.org/10.1145/2788396

of TOP500 and Green500 supercomputers now use both CPUs and GPUs [Green500 2014; Top500 2014]. An even closer level of integration between CPUs and GPUs can be achieved by fabricating them on the same chip, and many such processors have already been produced, such as AMD Llano [Branover et al. 2012], Intel Sandy Bridge [Yuffe et al. 2011], Ivy Bridge [Damaraju et al. 2012], etc.

The vastly different architectures and programming models of CPUs and GPUs, however, also present several challenges in achieving such collaborative computing. Due to the interaction between them in a heterogeneous system, optimizing performance and energy efficiency requires taking into account the characteristics of both the PUs. For this reason, conventional CPU-only or GPU-only optimization techniques may not work well in a heterogeneous system and hence, novel techniques are required to realize the potential and promise of heterogeneous computing and also move toward the goals of exascale performance.

Contributions. In this article, we provide an extensive survey of techniques and architectures proposed for heterogeneous computing. We first discuss the motivation for and challenges involved in heterogeneous computing and also clarify the terminology used in this research field (Section 2). In Section 2, we also discuss how the hardware architecture of both CPUs and GPUs have evolved over the years to identify important trends. We then analyze the research works from several perspectives to highlight their similarities and differences. From the perspective of researchers and algorithm designers, we summarize the works related to CPU-GPU workload partitioning and classify them based on their key idea, such as nature of scheduling (dynamic/static), basis of division of work among PUs, etc. (Section 3 and Tables I and II). From the point of view of programmers and application developers, we classify the works based on the languages used for programming CPUs and/or GPUs in those works and also review the techniques proposed that are related to compiler and programming framework/library (Section 4 and Tables III and IV).

We then discuss the techniques for saving energy in HCSs and classify them based on their essential approach (Section 5 and Table V). Further, we review research works dealing with fused (integrated) CPU-GPU chips to show their relative features/limitations compared to the discrete GPUs (Section 6 and Table VI). Furthermore, we classify the research works based on their application (or workload) domain and discuss benchmark suites designed for evaluating HCSs (Section 7 and Tables VII and VIII). We conclude this article with a discussion of the future trends (Section 8). We hope that this article will be useful for a wide range of readers, including computer architects, developers, researchers, and technical marketing professionals.

Scope of the Article. Since it is practically infeasible to review the broad spectrum of research works, we take the following approach to limit the scope of this survey. We discuss only those heterogeneous computing systems/techniques that use both CPUs and GPUs, although it is also possible to build heterogeneous systems using Field Programmable Gate Arrays (FPGAs), etc. We discuss heterogeneous computing techniques proposed for both discrete and fused CPU/GPU architectures. We do not include research works that obtain performance improvements from using GPUs alone or that discuss load balancing within a single GPU or within multiple GPUs only. In other words, we include works where a CPU is used as a processing unit also, in addition to as a host running operating system. We do not include circuit/device/microarchitectural level research works; rather, we include research works dealing with runtime and algorithm/application execution, compiler, programming framework/language/library and system-level techniques, etc. We mainly focus on the key research idea of the papers to gain insight, but we also mention the CPU/GPU used in their experiments to get an idea of their evaluation platform.

2. PROMISES AND CHALLENGES OF HETEROGENEOUS COMPUTING

2.1. Clarifying the Terminology

Since different research works use different terminologies, we first clarify them and also mention the nomenclature that we use in this article. In literature, CPU-GPU heterogeneous computing approaches have also been referred to as collaborative, hybrid, co-operative or synergistic execution, coprocessing, divide and conquer approach, etc. In accordance with its purpose of use, a GPU is conventionally referred to as an *accelerator*. However, given that CPU-GPU workload division based techniques on heterogeneous systems use CPUs also to get performance gain (and not just as a host), sometimes a CPU is also referred to as an accelerator [Sun et al. 2012]. Similarly, while a CPU is generally termed as *host* and GPU as a *device*, in the context of HCSs, some researchers use the term "device" to refer to both CPUs and GPUs [Liu and Luk 2012; Veldema et al. 2011]. To maintain clarity, we do not use the term accelerator to refer to GPU or device as a generic term for CPU or GPU. Instead, in this article we use *PU* (processing unit) as the generic term for either CPU or GPU, following other researchers (e.g., Binotto et al. [2010]). A few other equivalent terms used in literature are Computing Unit (CU) [Verner et al. 2011], Computing Element (CE) [Li et al. 2012], and Processing Element (PE) [Tsoi and Luk 2010].

At the architecture level, a chip that has both CPU and GPU integrated on the same chip is referred to as Accelerated Processing Unit (*APU*) or Single-Chip Heterogeneous Processor (*SCHP*). This is also referred to as a *fused* or *integrated* system, in contrast to a conventional *discrete* system, which has CPU and GPU on different chips, connected through a Peripheral Component Interconnect Express (PCIe) bus.

2.2. Evolution of Hardware Architecture of PUs

Before delving into *heterogeneous computing*, it is interesting to note how the hardware architecture and performance of each PU *individually* have evolved over time. For brevity, we only discuss some key parameters and focus on the last seven to eight years to see the trends by sampling a few products.

2.2.1. Transistor Count. A few years ago, CPUs had nearly 1B transistors [Wendel et al. 2010], while the recently announced Oracle SPARC M7 CPU will have more than 10B transistors on chip [Morgan 2014]. Similarly, the GT200 GPU (2008) had 1.4B transistors, while the recent Geforce GTX TITAN X GPU has 8B transistors [Pirzada 2015].

2.2.2. Core Count. Earlier CPUs had only one or two cores, while recent CPUs have 8 to 32 cores [Fluhr et al. 2014; Morgan 2014], and CPUs with more than 60 cores are likely to be available in the near future [Gardner 2014]. Along similar lines, the number of CUDA cores in GeForce GTX 280, GTX 480, GTX 680, and GTX TITAN X GPU is 240, 448, 1536, and 3072, respectively [NVIDIA 2015].

2.2.3. Cache Size. With increasing number of cores, the size of Last Level Cache (LLC) on a CPU is also increasing [Mittal 2014b]. The 45nm POWER7 processor had 32MB Embedded DRAM (eDRAM) LLC [Wendel et al. 2010], the 32nm POWER7+ processor had an 80MB eDRAM LLC [Zyuban et al. 2013], while the 22nm POWER8 processor has a 96MB eDRAM LLC [Fluhr et al. 2014]. Earlier GPUs only provided software-managed caches and not hardware-managed caches; however, as the GPUs move toward general-purpose computing, they are now featuring increasingly larger-sized hardware-managed caches. For example, GT200 architecture GPU did not have an L2 cache, the Fermi GPU has 768KB LLC, the Kepler GPU has 1536KB LLC, and the Maxwell GPU has 2048KB LLC [Mittal 2014a].

2.2.4. 3D Stacking. 3D stacking facilitates high bandwidth and memory capacity [Poremba et al. 2015] and hence, both CPUs and GPUs are increasingly moving toward use of 3D stacking; for example, Intel's Knights Landing [Gardner 2014] and NVIDIA's Pascal GPU [Gupta 2014] will feature 3D stacked memory.

2.2.5. Interconnect Bandwidth. The limited bandwidth of PCIe has conventionally remained a bottleneck in GPU performance, especially for applications that transfer large amounts of data between CPUs and GPUs [Gregg and Hazelwood 2011; Lee et al. 2010]. However, a recently proposed interconnect, called NVLink, promises to offer 5 to 12× bandwidth compared to PCIe Gen3 interconnect [Gupta 2014] and this is likely to offset the bandwidth limitation of conventional interconnects.

These trends make it evident that the hardware architecture of both CPUs and GPUs has undergone and is still undergoing a process of never-ending evolution. This motivates the research on CPU-GPU heterogeneous computing, which is exactly what we review in this article.

2.3. Motivation for Heterogeneous Computing

While the use of a GPU as a stand-alone device seems a promising idea at first, there are several compelling reasons for moving toward a heterogeneous CPU/GPU computing approach:

2.3.1. Acknowledging and Leveraging Unique Architectural Strengths of PUs. Both CPUs and GPUs possess distinct architectural features. Modern multicore CPUs use up to a few tens of cores, which are typically out-of-order, multi-instruction issue cores. Also, CPU cores run at high frequency and use large-sized caches to minimize the latency of a single thread. Clearly, CPUs are suited for latency-critical applications. In contrast, GPUs use a much larger number of cores, which are in-order cores that share their control unit. Also, GPU cores use lower frequency, and smaller-sized caches [Mittal 2014a]. Thus, GPUs are suited for throughput-critical applications. Thus, a heterogeneous system can provide high performance for a much wider variety of applications and usage scenarios than using either a CPU or a GPU alone [Vetter and Mittal 2015].

2.3.2. Matching Algorithmic Requirements to Features of PUs. For applications where data transfers dominate execution time, or branch divergence does not allow for the uninterrupted execution on all GPU cores, CPUs can provide better performance than GPUs. Not only different applications, but even different phases of a single application may exhibit properties that make it more suitable for execution on a particular PU [He and Hong 2010; Nere et al. 2012; Shen et al. 2013]. For example, Ding et al. [2009] note that for a query processing application, CPUs are more efficient for queries involving short lists, while GPUs are more efficient for those involving long lists.

2.3.3. Improving Resource Utilization. To meet the worst-case performance requirements, the CPU-only or GPU-only systems are usually overprovisioned; however, their average utilization remains low [Mittal 2012; Mittal and Vetter 2015]. Further, after allocating the task to the GPU (i.e., starting the kernel), the CPU stays idle, which leads to wastage of energy. Similarly, for applications where the GPU memory bandwidth acts as a bottleneck [Daga et al. 2011; Spafford et al. 2012], the computational resources of GPUs remain underutilized. HCTs can address these inefficiencies by intelligently managing the resources of both PUs [Gelado et al. 2010; Hu et al. 2011]. As a definite case in point, the June 2014 Green500 list of most energy efficient supercomputers shows that *all* the top 15 systems in the list are CPU-GPU heterogeneous systems [Green500 2014].

2.3.4. Reaping the Fruits of Advancements in CPU Design. In systems with GPUs, CPUs have been conventionally used as host for GPU to manage I/O and scheduling; however, as continuing innovations improve CPU performance even further (refer to Section 2.2), using their computation capabilities also has become more attractive. Further, while several initial works report that GPUs provide up to 100× to 1000× speedup, other researchers claim that on applying careful optimizations on *both* CPUs and GPUs, CPUs may equal or even outperform the performance of GPUs [Gummaraju et al. 2010; Lee et al. 2010]. Due to this, different amounts of work divisions to CPUs and GPUs can lead to vastly different performance [Grewe and O'Boyle 2011]. These points highlight the importance of using the computational capabilities of CPUs also.

2.4. Factors and Challenges Involved in Heterogeneous Computing

The vastly different architecture, programming model, and performance (for a given program) of CPUs and GPUs present unique challenges in heterogeneous computing. Several factors relating to both the PU and the application itself and spanning from microarchitecture level to system level need to be taken into account for fully leveraging their potential in an HCS. In what follows, we briefly mention these factors/issues.

2.4.1. PU Specific. (1) Architecture of HCS (discrete or fused). (2) Computation power of the PUs. (3) Current load on PUs and achieving load balancing between them. (4) Memory bandwidth and CPU-GPU data transfer overhead; avoiding and/or amortizing overhead of launching GPU kernel and data transfer. (5) Pipelining for overlapping data transfer with computation or CPU computation with GPU computation. (6) Taking into account limitations of PUs, for example, CPU-GPU memory bandwidth, size of GPU and CPU memory, number of GPU threads and CPU cores, reduced performance of GPU for double-precision computations, etc. [Vetter and Mittal 2015]. Also, aggressively using CPU for computation affects its ability to act as a host, which harms the performance [Endo et al. 2010; Gregg et al. 2010; Sun et al. 2012].

2.4.2. Application/Problem Specific. (1) Nature of algorithms, for example, amount of parallelism, presence of branch divergence. (2) Subdividing the workload and selecting suitable work sizes to be allocated to PUs. (3) Accounting for data dependencies, for example, if a task has data dependencies on a previous task, where was the previous task executed?

2.4.3. Objective Specific. Achieving higher level optimization targets such as energy saving, performance and fairness, etc.

The HCTs discussed in the next several sections account for these factors to avoid their impact on performance.

3. ALGORITHM AND PROGRAM-EXECUTION LEVEL WORKLOAD PARTITIONING TECHNIQUES

Tables I and II categorize the techniques proposed for intelligently partitioning the workload between a CPU and a GPU at the level of algorithm or during program execution. We classify the works based on two main criteria, which are as follows.

(1) *Dynamic or static scheduling*. This answers *when* the scheduling is done (Table I). In dynamic division, the decision about running the subtasks or program phases or code portions on a particular PU is taken at runtime. In static division, the subtasks that are executed on a particular PU are already decided before program execution; in other words, the mapping of subtasks to PUs is fixed.
(2) *Basis of workload partitioning*. This answers *why* a particular scheduling of tasks to PUs is done (Table II). This can be motivated by characteristic/capability of the PU itself and/or the subtasks themselves. For example, assuming that the

Table I. A Classification of Algorithm/Runtime Level HCTs Based on Nature of Scheduling/Mapping

Classification	References
Dynamic	Acosta et al. [2010], Agulleiro et al. [2012], Agullo et al. [2011], Albayrak et al. [2012], Álvarez-Melcón et al. [2013], Becchi et al. [2010], Belviranli et al. [2013], Bernabé et al. [2013], Bhaskaran-Nair et al. [2013], Binotto et al. [2011], Boyer et al. [2013], Breß et al. [2013], Chen et al. [2012], Choi et al. [2013], Clarke et al. [2012], Delorme [2013], Deshpande et al. [2011], Diamos and Yalamanchili [2008], Gao et al. [2012], Garba and González-vélez [2012], Gregg et al. [2010], Gregg et al. [2011], Hamano et al. [2009], Hartley et al. [2010], Hawick and Playne [2013], He and Hong [2010], Hermann et al. [2010], Horton et al. [2011], Humphrey et al. [2012], Huo et al. [2011], Jiang and Agrawal [2012], Jiménez et al. [2009], Joselli et al. [2008], Kofler et al. [2013], Kothapalli et al. [2013], Lang and Rünger [2013], Lecron et al. [2011], Lee et al. [2012], Li et al. [2011], Li et al. [2012], Liu et al. [2012], Ma et al. [2012], Ma et al. [2013], Mariano et al. [2012], Munguia et al. [2012], Muramatsu et al. [2011], Murarașu et al. [2012], Odajima et al. [2012], Papadrakakis et al. [2011], Pienaar et al. [2011], Ravi and Agrawal [2011], Ravi et al. [2012], Scogland et al. [2012], Shen et al. [2010], Shirahata et al. [2010], Siegel et al. [2010], Silberstein and Maruyama [2011], Stefanski [2013], Su et al. [2013], Tan et al. [2012], Teodoro et al. [2009], Teodoro et al. [2012], Teodoro et al. [2013], Udupa et al. [2009], Vömel et al. [2012], Wang et al. [2013b], Wang et al. [2014], Wu et al. [2012], Yang et al. [2010], and Yao et al. [2010]
Static	Agullo et al. [2011], Balevic and Kienhuis [2011], Banerjee and Kothapalli [2011], Benner et al. [2010], Benner et al. [2011], Binotto et al. [2010], Boratto et al. [2012], Boyer et al. [2013], Chai et al. [2013], Chen et al. [2010], Chen et al. [2012], Choi et al. [2013], da S Junior et al. [2010], Delorme [2013], Dziekonski et al. [2011], Endo et al. [2010], Gao et al. [2012], Gregg et al. [2011], Grewe and O'Boyle [2011], Hamano et al. [2009], Hampton et al. [2010], Hardy et al. [2009], Hu et al. [2011], Jetley et al. [2010], Korwar et al. [2013], Kothapalli et al. [2013], Liu and Luk [2012], Liu et al. [2009], Liu et al. [2011], Liu et al. [2012], Lu et al. [2012a], Lu et al. [2012b], Luo et al. [2011], Mariano et al. [2012], Matam et al. [2012], Munguia et al. [2012], Murarașu et al. [2012], Nakasato et al. [2012], Nigam and Narayanam [2012], Ogata et al. [2008], Ohshima et al. [2007], Pajot et al. [2011], Panetta et al. [2009], Park et al. [2011], Phothilimthana et al. [2013], Pienaar et al. [2011], Pirk et al. [2012], Rahimian et al. [2010], Ravi and Agrawal [2011], Scogland et al. [2012], Shen et al. [2013], Shukla and Bhuyan [2013], Singh and Aruni [2011], So et al. [2011], Sottile et al. [2013], Stpiczynski and Potiopa [2010], Sun et al. [2012], Takizawa et al. [2008], Toharia et al. [2012], Tsoi and Luk [2010], Tsuda and Nakamura [2011], Venkatasubramanian and Vuduc [2009], Verner et al. [2011], Wang and Song [2011], Wang et al. [2013a], Xiao et al. [2011], Xu et al. [2012], Yang et al. [2013], and Zhong et al. [2012]

different subtasks are of similar nature and can run on any of the PUs, the decision about where a subtask should be mapped depends on which PU provides higher performance and which mapping helps in achieving load balancing. However, if subtasks differ, it may be more suitable to map a particular subtask to a particular PU; for example, highly parallel subtasks can be mapped to the GPU, while the sequential subtasks can be mapped to the CPU. In some cases, a subtask cannot be mapped on a particular PU; for example, when the memory footprint of a subtask exceeds the memory size of the GPU.

In Table II, we also identify the techniques that use pipelining between the CPU and GPU, such that different phases or dependent subtasks are handled by the CPU and GPU in an overlapped manner, or transfer of data between them is overlapped with computation. We now discuss a few of these techniques.

3.1. Scheduling Based on Relative Performance of PUs

Ogata et al. [2008] present a library for 2D Fast Fourier Transform (FFT) that automatically uses both PUs to achieve optimal performance. Using a performance model, it evaluates the respective contributions of each PU and then makes an estimation of the total execution time of the FFT problem for arbitrary work distribution and

Table II. A Classification of Algorithm/Runtime Level HCTs Based on Criterion used for Workload Division (and Other Aspects)

Classification	References
Based on relative performance of PUs for load balancing	Acosta et al. [2010], Agulleiro et al. [2012], Agullo et al. [2011], Albayrak et al. [2012], Becchi et al. [2010], Belviranli et al. [2013], Bernabé et al. [2013], Bhaskaran-Nair et al. [2013], Binotto et al. [2011], Boratto et al. [2012], Boyer et al. [2013], Chai et al. [2013], Chen et al. [2012], Choi et al. [2013], Clarke et al. [2012], Endo et al. [2010], Gao et al. [2012], Garba and González-vélez [2012], Gharaibeh et al. [2012], Gregg et al. [2010], Gregg et al. [2011], Grewe and O'Boyle [2011], Hardy et al. [2009], Hartley et al. [2008, 2010], Hawick and Playne [2013], He and Hong [2010], Hermann et al. [2010], Hu et al. [2011], Humphrey et al. [2012], Huo et al. [2011], Jiang and Agrawal [2012], Jiménez et al. [2009], Joselli et al. [2008], Kofler et al. [2013], Kothapalli et al. [2013], Lang and Rünger [2013], Lecron et al. [2011], Lee et al. [2012], Li et al. [2011], Li et al. [2012], Liu et al. [2012], Lu et al. [2012a], Ma et al. [2012], Ma et al. [2013], Matam et al. [2012], Murarașu et al. [2012], Nakasato et al. [2012], Nigam and Narayanam [2012], Odajima et al. [2012], Ogata et al. [2008], Ohshima et al. [2007], Papadrakakis et al. [2011], Phothilimthana et al. [2013], Pienaar et al. [2011], Pienaar et al. [2012], Rahimian et al. [2010], Ravi and Agrawal [2011], Ravi et al. [2012], Rofouei et al. [2008], Scogland et al. [2012], Shen et al. [2010], Shimokawabe et al. [2011], Shirahata et al. [2010], Shukla and Bhuyan [2013], Siegel et al. [2010], Singh and Aruni [2011], Sottile et al. [2013], Su et al. [2013], Sun et al. [2012], Teodoro et al. [2012], Teodoro et al. [2013], Udupa et al. [2009], Venkatasubramanian and Vuduc [2009], Verner et al. [2011], Vömel et al. [2012], Wang et al. [2013b], Wang et al. [2014], Wen et al. [2012], Wu et al. [2012], Wu et al. [2013], Yang et al. [2010], Yang et al. [2013], Yao et al. [2010], and Zhong et al. [2012]
Based on low/high parallelism or other characteristics of subtasks/phases	Álvarez-Melcón et al. [2013], Balevic and Kienhuis [2011], Banerjee and Kothapalli [2011], Banerjee et al. [2012], Benner et al. [2010], Benner et al. [2011], Binotto et al. [2010], Breß et al. [2013], Chen et al. [2010], Choudhary et al. [2012], Deshpande et al. [2011], Diamos and Yalamanchili [2008], Dziekonski et al. [2011], Garba and González-vélez [2012], Hampton et al. [2010], He and Hong [2010], Hong et al. [2011], Horton et al. [2011], Hu et al. [2011], Jetley et al. [2010], Joselli et al. [2008], Kothapalli et al. [2013], Liu et al. [2009], Liu et al. [2011], Ltaief et al. [2011], Lu et al. [2012b], Luo et al. [2011], Mariano et al. [2012], Munguia et al. [2012], Muramatsu et al. [2011], Pajot et al. [2011], Panetta et al. [2009], Pirk et al. [2012], Shen et al. [2013], Shimokawabe et al. [2011], So et al. [2011], Stefanski [2013], Stpiczynski [2011], Stpiczynski and Potiopa [2010], Su et al. [2013], Teodoro et al. [2009], Toharia et al. [2012], Tomov et al. [2010], Tsoi and Luk [2010], Tsuda and Nakamura [2011], Wang and Song [2011], Wang et al. [2013a], Xu et al. [2012], and Yang et al. [2013]
Based on floating-point precision of PUs	Benner et al. [2011], Gao et al. [2012], and Stpiczynski and Potiopa [2010]
Based on limitation in size of GPU memory	Huo et al. [2011], Ogata et al. [2008], Teodoro et al. [2013], and Yao et al. [2010]
Other aspects	
Pipelining	Balevic and Kienhuis [2011], Banerjee and Kothapalli [2011], Banerjee et al. [2012], Benner et al. [2011], Chen et al. [2012], Choudhary et al. [2012], Conti et al. [2012], Deshpande et al. [2011], Hampton et al. [2010], Huo et al. [2011], Jetley et al. [2010], Korwar et al. [2013], Li et al. [2012], Liu et al. [2011], Ltaief et al. [2011], Pajot et al. [2011], Panetta et al. [2009], Park et al. [2011], Pienaar et al. [2012], Shimokawabe et al. [2011], Su et al. [2013], Tan et al. [2012], Udupa et al. [2009], Wen et al. [2012], Wu et al. [2012], Xiao et al. [2011], Yang et al. [2010], and Yang et al. [2013]
Use of MapReduce framework	Chen et al. [2012], Hong et al. [2010], Jiang and Agrawal [2012], Ravi et al. [2012], Shirahata et al. [2010], Tan et al. [2012], Tsoi and Luk [2010], and Wu et al. [2013]

problem sizes. It decomposes the computation of FFT into multiple substeps, and uses profiling along with data transfer time to estimate the execution time of each step. Using these estimates, the optimal workload division between PUs is found for achieving load balancing. Their experimental system uses Core 2 Duo E6400 CPU and GeForce 8800 GTX GPU.

Ding et al. [2009] propose an approach for accelerating query processing. Based on the length of the query, their technique decides whether a query can be more efficiently analyzed on a CPU or GPU. Based on this, the tasks are put in one of the two task queues. A third task queue is made with the queries that may be efficiently analyzed on both CPUs and GPUs. A PU first processes queries from its own task queue and then from the third task queue. If it is still idle, it can steal a task from the task queue of another PU, following the idea of work-stealing scheduling. Their evaluation platform uses Core 2 Duo CPU and GeForce 8800 GTS GPU.

Gregg et al. [2010] propose a workload division technique that schedules works on devices based on several factors such as the contention of devices, historical performance data, number of cores, processor speed, problem size, and device status; for example, busy or free. If the CPU is busy with running many threads, the process is run on the GPU; and if the GPU is significantly faster than the CPU for a process, their technique waits for the GPU even if there is contention on the GPU. Moreover, if the problem size is such that the data associated with it cannot fit into the GPU memory, the process is run on the CPU. They perform experiments using Core 2 Duo CPU and Radeon HD 4350 GPU.

Becchi et al. [2010] propose a technique for automatically scheduling computation tasks over an HCS and managing data placement. Their technique intercepts function calls to kernels and schedules them on a PU based on their argument size, historical profile, and location of data. Their technique accounts for both computation time and data transfer time. They observe that if the data are already available on a GPU (say), scheduling a task on a GPU may provide an advantage even if the GPU provides less performance than the CPU. Thus, their technique defers all data transfers between PUs until necessary. Their experiments are conducted using a quad-core Xeon CPU and a Tesla C870 GPU.

Agullo et al. [2011] propose a method for accelerating QR factorization using a CPU/GPU heterogeneous system. Their method works in three steps. In the first step, the QR factorization problem is expressed in terms of sequence of tasks of desired granularity such that they can be executed on a suitable PU. In the second step, CPU or GPU functions (or kernels) are designed for executing those tasks. Finally, in the third step, static or dynamic scheduling is used for scheduling these tasks on a CPU or GPU. Static scheduling utilizes a priori knowledge of the schedule and provides high performance but does not offer portability. Dynamic scheduling uses a StarPU runtime system [Augonnet et al. 2011] and thus provides high productivity and the ability to schedule complex algorithms on heterogeneous systems. StarPU is a tasking Application Programming Interface (API) that facilitates the execution of parallel tasks on heterogeneous computing platforms, and incorporates multiple scheduling policies. They perform experiments using two HCSs: one with Xeon X5550 CPU and Quadro FX5800 GPU and another with Opteron 8358 SE CPU and Tesla S1070 GPU.

Lecron et al. [2011] present an HCT for detecting and segmenting vertebra in X-ray images. The most computation intensive part of vertebra segmentation is edge detection. After initial loading of images, their technique uses the StarPU system to map edge detection function on GPUs and CPUs to effectively utilize both PUs. StarPU also provides functionality to transfer the GPU results to the CPU at the end of computation. Using their approach, multiple CPU cores and GPUs can be simultaneously

used for achieving further speedup. Their experiments use Core 2 Duo 6600 CPU along with Tesla C1060 GPU.

Li et al. [2011] propose an HCT for Cryo-EM 3D reconstruction, where tasks (in this case, individual images) are assigned to a CPU and a GPU based on their relative performance. The estimates of performance of PUs is updated during each iteration of execution of algorithm. They also propose techniques to exploit hardware parallelism provided by each PU by leveraging thread-level and data-level parallelism. Their experiments are performed on a supercomputer whose nodes are composed of two six-core Xeon X5650 CPUs and a Tesla C2050 GPU.

Deshpande et al. [2011] present two techniques for accelerating image dithering operation. In image dithering, the amount of parallelism available changes for different regions of the image. The parallelism is low toward the beginning and end, and is high in the middle of dithering. Their first technique, called "CPU-GPU handover" uses a CPU for executing the initial part of the algorithm (step 1), then transfers the data to a GPU for doing the middle part of the algorithm (step 2), and finally hands over the data back to the GPU for doing the last part of the algorithm (step 3). In this technique, although both a CPU and a GPU are used for the task for which they are most efficient, they are not used together at the same time. Their second technique called "CPU-GPU hybrid" changes the previously mentioned step 2, such that the work of step 2 is divided between the CPU and GPU. Thus, the CPU performs steps 1 and 3 alone, and also shares the work in step 2. They have shown that the second technique provides higher performance than the first technique. Their experiments use a Core 2 Duo P8600 CPU along with a GeForce 8600M GT.

Verner et al. [2011] propose an algorithm for implementing hard real-time stream scheduling in heterogeneous systems. Their algorithm partitions the incoming streams into two subsets: one for processing by the GPU and the other for the CPU. Using a schedulability criteria, it is ensured that both subsets are schedulable. The algorithm works to find an assignment that satisfies both the deadline constraint of each stream alone and the aggregate throughput requirements of all the streams. A Core 2 Quad CPU and a GeForce GTX 285 GPU were used for their experiments.

Grewe and O'Boyle [2011] propose a static partitioning technique for OpenCL programs on HCSs. Their technique conducts static analysis on OpenCL programs to extract code features. Using this information, their technique first determines the best work-division ratio across the PUs in a system. Afterward, it divides the workload into suitable sized chunks for each PU using machine learning approach. Use of machine learning makes their technique portable across different implementation platforms and different implementations of OpenCL. They evaluate their technique using a system with a Xeon E5530 CPU and a Radeon HD 5970 GPU.

Agulleiro et al. [2012] present an HCT for 3D tomographic reconstruction. In electron tomography, single-tilt axis geometry is used to decompose a 3D reconstruction problem into several independent 2D reconstruction tasks. These 2D slices of the volume can be computed in parallel using a reconstruction method, on either a CPU or a GPU. Their technique starts a CPU thread and one thread for each GPU, which is responsible for sending data to the GPU and receiving the slices after reconstruction. When the threads become idle, more work (slices) is assigned to them and parallel execution takes place on both PUs. The amount of slices allocated to a PU depends on its performance, specifically the GPU is allocated alarger number of slices. Thus, their technique uses on-demand allocation for achieving load balancing. They use two HCSs for evaluation purposes: one with a quad-core Xeon E5640 CPU and Tesla C2050 GPU and another with a Quad-core Xeon W3520 CPU and GeForce GTX 285 GPU.

Teodoro et al. [2012] use CPU-GPU on-demand allocation on an HCS for accelerating an image processing application. To amortize the overhead of a GPU kernel call and

data transfer, they group large pieces of data to be processed by each kernel and ensure that each kernel performs a maximum amount of possible computations (i.e., steps of the image processing algorithm) with the data. Their experiments use a six-core Xeon X5660 CPU along with a Tesla M2070 GPU.

Boratto et al. [2012] use static scheduling to divide the workload of matrix computation on a CPU and a GPU for solving the problem of landform attributes representation. Their evaluation platform uses six-core Xeon X5680 CPUs and Tesla C2070 GPUs.

Matam et al. [2012] present a workload-division technique for generalized sparse matrix-matrix multiplication (SPGEMM). Due to the irregular computations in SPGEMM, a GPU does not provide large speedup for this problem. Their technique observes the speedup of a GPU over a CPU and accordingly assigns a fraction of work to PUs such that their finish times become equal. They perform experiments using a system with a Core i7 920 CPU and a Tesla C2050 GPU.

Lu et al. [2012a] present an HCT for accelerating a Rapid Radiation Transfer Model (RRTM) application used in radiation physics. RRTM involves a loop over a two-dimensional grid and thus, presents an opportunity for workload division by splitting along the x or y direction. Their technique divides the workload between a CPU and a GPU by taking into account their processing power. They use a Message-Passing Interface (MPI) + OpenMP/CUDA programming model, where communication between different nodes takes place using MPI; CPU parallelization is implemented using OpenMP and the GPU is programmed using CUDA. They perform experiments using a Tianhe-1A supercomputer whose compute node has two six-core Xeon X5670 CPUs and a Tesla M2050 GPU.

Hawick and Playne [2013] present an HCT for cluster component-labeling analysis in critical phase modeling. Their technique performs a simulation of the Potts model on a GPU, since the Potts system has regular data structure and memory access locality. It also uses random number generation, which can be efficiently implemented on a GPU. For connected component labeling, their technique uses a CPU if at least one CPU core is available, otherwise, this labeling is performed on the GPU itself. This approach indirectly takes the different performance values of the GPU/CPU into account and avoids CPU idling. They perform experiments using multiple CPU-GPU HCSs, viz., an i7-2700K with a GTX590 (utilizing only one of its GPUs), an i7-970 with a GTX580, two Xeon E5-2640's with an M2090, two Xeon X5675's with an M2075, and finally two Opteron 6274's with a GTX680. They note that in all cases, the hybrid algorithm provided higher performance than using just the GPU. They also observed that based on decreasing order of performance, GPUs can be ranked as GTX680, GTX580, M2090, GTX590, and M2075.

Choi et al. [2013] present a workload division based scheduling technique for HCS. They also discuss simple scheduling policies such as the first-free scheduling policy, which schedules an incoming task on the first available PU; an alternate-assignment scheduling policy, which schedules the tasks alternately on a CPU and a GPU, and a performance-history scheduling policy, which computes the ratio of performance of a GPU over a CPU for a task based on performance history and schedules the task on a GPU if the ratio is more than a threshold. Their technique provides enhancements to performance-history scheduling policy by also taking into account the information about the time when a PU will become available. Using this, along with actual estimated execution time of the task, the completion time of a task on both a CPU and a GPU can be estimated and the best PU can be chosen to minimize the makespan. Their evaluation platform is a system with a Core 2 Quad Q9400 CPU and a Geforce 8500GT GPU.

Bernabé et al. [2013] present an HCT for accelerating 3D-Fast Wavelet Transform. For a GPU, they use kernels implemented in CUDA or OpenCL and for a CPU, they use

kernels implemented in Pthread. Using these kernels, they first profile the performance of the GPU and CPU and then allocate the workload to the PUs in proportion to their performance. They experiment using two systems. The first system has a Xeon E5620 CPU, a Tesla C2050 GPU, and an ATI FirePro V5800 GPU. The second system has a Core 2 Quad Q6700 CPU and a Tesla C870 GPU. Of the two, the first system provides higher performance.

Belviranli et al. [2013] present a dynamic load-balancing technique for loop iterations on HCS. Their technique works in two phases. In the first phase, the relative performance of PUs is estimated by experimenting with different task size allocations to the PU. Small task sizes provide better load balancing, while large task sizes provide better resource utilization and an optimal value of task sizes achieves a trade-off between these two factors. In the second phase, the remainder and majority of computations are performed based on the relative performance values obtained in the first phase. The second phase utilizes a modified version of a self-scheduling algorithm [Belviranli et al. 2013] to achieve load balancing. Their experiments use a system with a 16-core Opteron 6200 CPU and a Tesla C2050 GPU.

3.2. Scheduling Based on Nature of Subtasks

Liu et al. [2009] present a method to parallelize Nonrigid Registration (NRR) of medical images. They first divide the original serial algorithm into regular and irregular parts. Then, the regular part is mapped to a GPU since it offers high parallelism. The irregular part is mapped to a multicore CPU since it requires synchronization and communication and hence, cannot benefit from GPU implementation. A system with a dual-core Opteron 2218 CPU and a GeForce 8800 GT GPU is used for their experiments.

Yao et al. [2010] propose an HCT for HMMER application that is used for biosequence analysis. Their technique spawns one thread each for the CPU and the GPU; and when a sequence is retrieved from the database it can be assigned to either a CPU or a GPU. The HMMER algorithm has a large memory requirement and thus, for large-sized sequences, shared memory of the GPU becomes unsuitable while global memory incurs a large performance penalty. Hence, their technique maps sequences larger than a threshold on the CPU and smaller sequences on the GPU. Their experiments are conducted on a system with a Core 2 Duo E7200 CPU and a GeForce 8800 GTX GPU.

Hermann et al. [2010] propose a workload-division technique for interactive physics simulations. Their technique divides the work of time integration between PUs using a conventional graph partitioner and then uses work-stealing scheduling to achieve load balancing. Their work-stealing algorithm takes spatial and temporal locality into account. For leveraging spatial locality, the tasks using the same data are likely to be mapped to the same PU. For leveraging temporal locality, at the time of starting a new time integration step, preference is given to a PU that executed the previous iteration. Also, during work stealing, their technique takes into account the fact that smaller tasks are more efficient on the CPU and vice versa. Their experiments use a quad-core Nehalem CPU and a GeForce GTX 295 GPU.

Benner et al. [2010] propose an HCT for accelerating computation of matrix sign function using Gauss-Jordan elimination. In each iteration, the factorization of the current column panel is performed on the CPU. It is because this step involves smaller data size, BLAS-1 operations, and pivoting, which are not suitable for parallelization. On the GPU, matrix multiplication and pivoting of the columns outside the current column panel are performed, since they can be easily parallelized. Their evaluations are performed using a system with a Xeon quad-core E5405 CPU and a Tesla C1060 GPU.

Hu et al. [2011] present an approach for accelerating Fast Multipole Method (FMM) on an HCS. They study the cost of different phases of FMM and the communication involved. Then, they distribute the work by considering the characteristics of both the CPU and GPU, such that each PU is given the work that it can do in the most efficient manner. This approach also provides load balance and minimizes the data transfer between them. Their evaluations are performed over three CPU-GPU HCSs: one with a Xeon X5560 CPU and a Tesla S1070 GPU, a second with a quad-core Xeon Harpertown 5300 CPU and a Tesla S1070 GPU, and a third with a Xeon E5504 CPU and a Tesla C2050 GPU.

Muramatsu et al. [2011] present an HCT for accelerating Hessenberg reduction for nonsymmetric eigenvalue problems. Basic Linear Algebra Subprograms (BLAS) operations form the majority of operations in this algorithm. Their technique assigns level-1 and level-2 BLAS operations on a CPU, since parallelism present in these operations is limited and hence, the data transfer cost outweighs the performance gain obtained by using a GPU. Further, for level-3 BLAS operations, the data are divided between a CPU and a GPU, and they both perform operations on the data. This enables leveraging the performance of the GPU, while avoiding CPU idling. They perform evaluations using a Core i7 920 CPU and a Tesla C1060 GPU.

Hong et al. [2011] propose a technique for accelerating Breadth-First Search (BFS) application. For each level of BFS algorithm execution, their technique dynamically selects the most suitable implementation from multiple choices, viz., a sequential execution and two possible parallel execution methods on a CPU and a GPU execution. This decision is taken based on the number of nodes to be traversed in each level. For a small number of nodes, the parallelism present is small, due to which GPU resources cannot be fully utilized and overhead of data transfer to a GPU cannot be amortized. By choosing the suitable PU, their approach optimizes the performance achieved for each graph size. The experiments are conducted using a Xeon X5550 CPU and a Tesla C2050 GPU.

Stpiczynski [2011] implements linear recurrence systems involving constant coefficients on an HCS. The algorithm for solution of this problem can be expressed in terms of BLAS-2 and BLAS-3 operations. His heterogeneous implementation uses a CPU for sequential steps and a GPU for parallel steps to leverage the benefits of both. An HCS with a Core i7 950 CPU and a Tesla C2050 GPU is used for the experiments.

Stefanski [2013] proposes a workload-division technique for Discrete Green's Functions (DFGs), where short length DFGs are assigned to a CPU and long length DFGs are assigned to a GPU. They conduct experiments on a system with a quad-core Core i7 and a GeForce GTX 680 GPU.

In the context of electromagnetic scattering, Gao et al. [2012] present a technique for implementing the "Shooting and Bouncing Ray" (SBR) method in conjunction with truncated wedge "incremental length diffraction coefficients" (TW-ILDCs) on HCS. They map SBR on a GPU since numerous independent ray tubes can fully utilize the massively parallel resources on the GPU. TW-ILDCs are mapped to a CPU since they require high-precision, complex numerical calculations to get an accurate result. Evaluations are performed using a system with a Core i5 CPU and a GeForce 470 GTX GPU.

Wang et al. [2013a] present an HCT for accelerating mapping of high-resolution human brain connectomes. Out of different steps of the algorithm, their technique maps computation of Pearson's correlation, modular detection, and all-pairs shortest path to a GPU (based on available parallelism), and the other steps such as computation of clustering coefficients to a CPU. Their experiments are performed on a system with a quad-core Core i7-3770 CPU and a Geforce GTX 580 GPU.

Yang et al. [2013] present an HCT for global atmospheric simulations using a cubed-sphere shallow-water model. This simulation involves processing 2D data arranged

in the form of rectangular patches. Their technique divides the patch into sub-blocks, where the computation of a sub-block also depends on its neighbors. Further, the subblock is divided into an inner part, the computation for which does not depend on neighboring subblocks, and an outer part consisting of four boundaries. Their technique schedules the inner part on a GPU to leverage parallelism and the outer part on a CPU since this part does not present large parallelism. By adjusting the size of the inner part, their technique adjusts the workload division between PUs to achieve the best possible performance. They conduct evaluations on a Tianhe-1A supercomputer (configuration shown previously) and achieve 0.8 PetaFLOP performance.

3.3. Pipelining

Park et al. [2011] present an HCT for ultrawideband signal processing applications. Their technique divides the imaging component of this application, such that the back-projection step is carried out on a GPU and the data interpolation step is carried out on a CPU. Moreover, using asynchronous kernel launch, the CPU starts processing the next frame while the GPU is completing the current frame. Using this approach, the latency of the GPU backprojection processing time is effectively hidden. They use three CPU-GPU HCSs, which, in decreasing order of performance are (1) Xeon 5570 + Tesla 1060, (2) Xeon 5160 + GeForce 8800 GTX, and (3) Core 2 Duo T9500 + Quadro 3600M.

Xiao et al. [2011] propose an HCT for accelerating protein sequence search using Basic Local Alignment Search Tool for Protein (BLASTP) sequence search. BLASTP has four steps (viz., hit detection, ungapped extension, gapped alignment, gapped alignment with traceback), of which the first three stages consume 99% of the execution time. Also, execution of the first two stages is performed together, while that of the third stage is done independently. In their technique, the first two stages are executed on a GPU and the third stage is executed on a CPU. Also, the database is separated into several chunks. The GPU executes the first stages for a given chunk, transfers the output to the CPU and then the CPU starts stage three. Meanwhile, the GPU starts processing the next chunk. Thus, they improve resource utilization using pipelining. Their CPU-GPU hybrid experimental system uses a Core 2 Duo CPU and one of the two following GPUs: Tesla C1060 and Tesla C2050. Of the two GPUs, Tesla C2050 provides higher performance.

Banerjee et al. [2012] present a work-division technique for on-demand pseudorandom number generation. Their random number generator uses parallel random walks on an expander graph. In this problem, few random bits are required to select a neighbor on the graph. In their technique, these bits are generated on a CPU and are then transferred to a GPU in an asynchronous manner and then the GPU generates the random number. Since each walk can be performed independent of each other, massive parallelism can be obtained and using asynchronous transfer, latency of communication can be overlapped with that of computation. Their technique performs better than both CPU-only and GPU-only implementations. Their evaluation platform has a Core i7 980 CPU and a Tesla C1060 GPU.

Li et al. [2012] propose the use of heterogeneous computing to accelerate Double-Precision General Matrix Multiplication (DGEMM) algorithm taking into account the case where the size of matrices is too large to fit in GPU memory. Their technique divides the matrices into submatrices and multiplies them block by block. Also it separates the GPU computation from write-back of results. Thus, both data transfer to a GPU and from a GPU are effectively overlapped with computation on a GPU, so that the CPU, GPU, and PCIe can work in a pipelined manner, which hides the latency of computation. Their experimental system has a Xeon X5650 CPU and a Radeon HD5970 GPU and achieves up to 844 GFLOPS performance.

3.4. Use of MapReduce Framework

Shirahata et al. [2010] propose a map task scheduling technique for HCSs. If a MapReduce job, whose tasks can be executed on both a CPU and a GPU, is submitted, a task scheduler uses profiles collected from dynamic monitoring of map task's behavior to assign the map task to a suitable PU with a view to minimize the overall MapReduce task execution time. They implement their technique using Hadoop system. The experimental system uses dual-core Opteron 880 CPUs and Tesla S1070 GPUs.

Tsoi and Luk [2010] describe a heterogeneous compute cluster named Axel. Axel uses Nonuniform Node, Uniform System (NNUS) architecture, where heterogeneous PUs (viz., CPU, GPU, and Field-Progammable Gate Array (FPGA)) are hosted in a single node, and all nodes are connected through a system bus. Axel uses MapReduce framework where a GPU and FPGA work on the Map part in parallel and a CPU works on the Reduce part. Further, it accounts for processing capability, local memory and communication capability for processing units to assign works to them and thus, leverage the parallelism of the GPU, the specialization of FPGA, and the scalability of CPU clusters. The experimental system uses quad-core Phenom X4 9650 CPUs and Tesla C1060 GPUs.

Chen et al. [2012] present two approaches for accelerating MapReduce applications on an HCS. The first approach, called *map dividing*, dynamically schedules workload to each scheduling unit on both a CPU and a GPU, and each PU conducts map and reduce simultaneously. The second approach, called *pipelining*, runs the map and reduces stages on different PUs, i.e., each PU only executes one stage of MapReduce. They use one core of the CPU as the scheduler; and the remaining CPU cores and each streaming multiprocessor of GPU as one scheduling unit. To achieve load balance while keeping scheduling costs low, they also propose a runtime tuning method to adjust task block sizes for each scheduling unit. The experiments are done using AMD Fusion APU (A8-3850), which integrates a quad-core CPU and an HD6550D GPU.

4. PROGRAMMING LANGUAGES, FRAMEWORKS, AND RELATED DEVELOPMENT TOOLS

4.1. A Classification of Research Works Based on Programming Languages Used

Table III classifies the works discussed in this article based on the programming language used. Different languages offer different trade-offs between ease of programming, ability to write optimized code, ability to target multiple PUs and products from different vendors, etc. For example, OpenCL can be used for both CPUs and GPUs, and hence, applications written in OpenCL can be easily scheduled on any PU. By comparison, CUDA works only on NVIDIA GPUs and AMD Core Math Library (ACML) and Brook+ can be used to program AMD GPUs only. Similarly, OpenMP, Pthread, Thread Building Block (TBB), and Math Kernel Library (MKL) are used to write parallel programs on CPUs. From Table III, it is clear that for GPU programming, CUDA is most widely used, which is due to its similarity with C/C++ and the ability to write optimized code. For CPU programming, OpenMP is most widely used due to its portability and ease of programming by virtue of use of compiler directives. Note that MKL is internally parallelized by OpenMP threading; for the sake of clarity, we mention it separately in Table III. In Section 4.4, we briefly discuss OpenCL due to its capability to provide heterogeneous computing and omit the discussion of other languages/libraries mentioned in Table III for the sake of brevity.

Table IV summarizes the works on programming frameworks, compiler techniques, and runtime systems that facilitate heterogeneous computing. We now discuss a few of them.

Table III. Programming Languages Used for GPU and/or CPU in Different Research Works

Classification	References
OpenCL	Albayrak et al. [2012], Bernabé et al. [2013], Binotto et al. [2011], Boyer et al. [2013], Conti et al. [2012], Daga et al. [2011], Danalis et al. [2010], Delorme [2013], Gregg et al. [2010], Gregg et al. [2011], Grewe and O'Boyle [2011], Hetherington et al. [2012], Kofler et al. [2013], Lee et al. [2013a], Liu et al. [2012], Nakasato et al. [2012], Phothilimthana et al. [2013], Shen et al. [2013], Spafford et al. [2012], Ukidave et al. [2013], and Veldema et al. [2011]
CUDA (and its libraries) on GPU	Acosta et al. [2010], Agulleiro et al. [2012], Agullo et al. [2011], Álvarez-Melcón et al. [2013], Anzt et al. [2011], Augonnet et al. [2011], Balevic and Kienhuis [2011], Banerjee and Kothapalli [2011], Banerjee et al. [2012], Becchi et al. [2010], Belviranli et al. [2013], Benner et al. [2010], Benner et al. [2011], Bernabé et al. [2013], Bhaskaran-Nair et al. [2013], Binotto et al. [2010], Boratto et al. [2012], Breß et al. [2013], Chai et al. [2013], Chen et al. [2010], Choudhary et al. [2012], Clarke et al. [2012], da S Junior et al. [2010], Deshpande et al. [2011], Diamos and Yalamanchili [2008], Dziekonski et al. [2011], Endo et al. [2010], Gao et al. [2012], Gharaibeh et al. [2012], Hampton et al. [2010], Hardy et al. [2009], Hartley et al. [2008, 2010], Hawick and Playne [2013], He and Hong [2010], Hermann et al. [2010], Hong et al. [2011], Hu et al. [2011], Humphrey et al. [2012], Huo et al. [2011], Jetley et al. [2010], Jiang and Agrawal [2012], Jiménez et al. [2009], Korwar et al. [2013], Kothapalli et al. [2013], Lang and Rünger [2013], Lecron et al. [2011], Lee et al. [2012], Li et al. [2011], Liu and Luk [2012], Liu et al. [2009], Liu et al. [2011], Ltaief et al. [2011], Lu et al. [2012a], Lu et al. [2012b], Luo et al. [2011], Ma et al. [2012], Ma et al. [2013], Mariano et al. [2012], Matam et al. [2012], Meredith et al. [2011], Munguia et al. [2012], Muramatsu et al. [2011], Murarașu et al. [2012], Nakasato et al. [2012], Nigam and Narayanam [2012], Ogata et al. [2008], Ohshima et al. [2007], Padoin et al. [2012], Pajot et al. [2011], Panetta et al. [2009], Papadrakakis et al. [2011], Park et al. [2011], Phothilimthana et al. [2013], Pienaar et al. [2011], Pienaar et al. [2012], Rahimian et al. [2010], Ravi et al. [2012], Rofouei et al. [2008], Shen et al. [2010], Shimokawabe et al. [2011], Shirahata et al. [2010], Shukla and Bhuyan [2013], Siegel et al. [2010], Silberstein and Maruyama [2011], Singh and Aruni [2011], So et al. [2011], Sottile et al. [2013], Stefanski [2013], Stpiczynski [2011], Stpiczynski and Potiopa [2010], Su et al. [2013], Sun et al. [2012], Takizawa et al. [2008], Tan et al. [2012], Teodoro et al. [2009], Teodoro et al. [2012], Teodoro et al. [2013], Toharia et al. [2012], Tomov et al. [2010], Tsoi and Luk [2010], Tsuda and Nakamura [2011], Udupa et al. [2009], Venkatasubramanian and Vuduc [2009], Verner et al. [2011], Wang et al. [2013a], Wang et al. [2013b], Wang et al. [2014], Wen et al. [2012], Wu et al. [2012], Wu et al. [2013], Xiao et al. [2011], Xu et al. [2012], Yang et al. [2013], Yao et al. [2010], and Zhong et al. [2012]
Brook+ on GPU	Park et al. [2011] and Wang and Song [2011]
ACML on GPU	Li et al. [2012] and Yang et al. [2010]
OpenMP on CPU	Álvarez-Melcón et al. [2013], Ayguade et al. [2009], Banerjee and Kothapalli [2011], Boratto et al. [2012], Chai et al. [2013], Clarke et al. [2012], Conti et al. [2012], Daga et al. [2011], Deshpande et al. [2011], Hong et al. [2010], Korwar et al. [2013], Kothapalli et al. [2013], Lang and Rünger [2013], Li et al. [2011], Li et al. [2013], Liu and Luk [2012], Lu et al. [2012a], Lu et al. [2012b], Mariano et al. [2012], Nigam and Narayanam [2012], Park et al. [2011], Rahimian et al. [2010], Shen et al. [2010], Sottile et al. [2013], Teodoro et al. [2013], Veldema et al. [2011], Wen et al. [2012], Wu et al. [2012], and Yang et al. [2013]
Pthread on CPU	Balevic and Kienhuis [2011], Becchi et al. [2010], Bernabé et al. [2013], Humphrey et al. [2012], Singh and Aruni [2011], Su et al. [2013], Wang et al. [2013b], Wang et al. [2013b], and Xiao et al. [2011]
Intel MKL on CPU	Agullo et al. [2011], Benner et al. [2010, 2011], Boratto et al. [2012], Clarke et al. [2012], Dziekonski et al. [2011], Horton et al. [2011], Ltaief et al. [2011], Matam et al. [2012], Meredith et al. [2011], Spafford et al. [2012], Tomov et al. [2010], and Yang et al. [2010]
Intel TBB on CPU	Becchi et al. [2010], Breß et al. [2013], Luk et al. [2009], and Pienaar et al. [2011]

Table IV. A Classification of Development Tools for HCSs

Classification	References
Programming frameworks, languages, or libraries	Augonnet et al. [2011], Ayguade et al. [2009], Diamos and Yalamanchili [2008], Gelado et al. [2010], Hong et al. [2010], Humphrey et al. [2012], Insieme Compiler [2014], Jablin et al. [2012], Kim et al. [2012], Lee et al. [2013], Li et al. [2013], Mistry et al. [2013a], Odajima et al. [2012], OpenACC Standard [2014], OpenMP 4.0 [2014], Pai et al. [2010], Pandit and Govindarajan [2014], Pienaar et al. [2011], Saha et al. [2009], Scogland et al. [2012], Scogland et al. [2014], Spafford et al. [2010], Stone et al. [2010], Udupa et al. [2009], and Veldema et al. [2011]
Compiler-level techniques for mapping to CPU/GPU	Kofler et al. [2013], Luk et al. [2009], Phothilimthana et al. [2013], Pienaar et al. [2012], Prasad et al. [2011], Ravi et al. [2010], and Takizawa et al. [2008]
Design of OpenCL and related languages, frameworks, or extensions	Gummaraju et al. [2010], Kim et al. [2012], Mistry et al. [2013a], Phothilimthana et al. [2013], Shen et al. [2013], Spafford et al. [2010], Stone et al. [2010], and Sun et al. [2012]

4.2. Programming Frameworks

Diamos and Yalamanchili [2008] propose Harmony, which utilizes performance estimates to schedule applications in an HCS. They propose online monitoring of kernels and describe a dependence-driven scheduling that analyzes how applications share data and decides on PU selection based on which applications can run without blocking. Their evaluations are performed on a system with an Athlon64 CPU and a GeForce 8800 GT GPU.

Hong et al. [2010] propose a framework called MapCG that facilitates portability between a CPU and a GPU at the level of source code. Without requiring modification, a program can be compiled and executed on either a CPU or GPU using a MapReduce programming model. Use of OpenCL enables using a single kernel code version for both PUs in place of providing separate kernel versions for them. Their experimental platform uses a six-core Opteron CPU and a GeForce GTX 280 GPU.

Pai et al. [2010] present a programming framework, named PLASMA, that enables writing of portable Single-Instruction, Multiple-Data (SIMD) programs. PLASMA uses an Intermediate Representation (IR), which provides succinct and clean abstractions (i.e., free from details of any particular SIMD architecture) to enable programs to be compiled on different PUs. Then, using a runtime, these programs can be automatically multithreaded and executed on different PUs, such as a GPU, a CPU, and a Cell BE; for example, a `for` loop in a kernel can be split across a CPU and a GPU. The runtime takes care of load balancing and distributed memory and also ensures that before any computation, data are moved from the CPU to the GPU or vice versa. Their experiments show that the performance of portable PLASMA code compares well to the platform-specific codes. They perform experiments using a system with a quad-core Xeon E5440 CPU and a GeForce 8800 GTS GPU.

Pienaar et al. [2011] present a runtime framework for the execution of workloads represented as parallel-operator directed acyclic graphs (PO-DAGs) on HCSs. They identify four criteria, viz., suitability, locality, availability, and criticality, which are important to consider while performing workload division for achieving good performance. Suitability shows which PU is most suited, that is, provides the best performance for a task. Locality shows whether the data required for a task is present in the memory of a PU and if so, by scheduling the task on that PU, data transfer cost can be avoided. Availability shows when a PU will become available for scheduling, thus it might be advantageous to wait for a suitable PU to become free rather than to schedule in a greedy manner. Criticality shows how the execution of a task affects the overall

execution time, thus by scheduling the kernels that are on critical path on the most suitable PU, the makespan can be reduced. Their method uses analytical models for estimating communication time and online history information for estimating kernel execution time; and based on these schedules the tasks on different PUs to achieve the best possible performance. The experiments have been conducted for three classes of processors: netbook (Intel Atom 330 + NVIDIA ION), laptop (Core 2 Duo + GeForce 320M), and server (Xeon 5500 + Tesla C1060).

Veldema et al. [2011] propose a framework for HC in the context of OpenMP adapted to Java, called ClusterJaMP. They propose an array package that provides replicated and partitioned arrays, using which a parallel-for loop can be distributed over PUs. At the beginning of a program, they execute a microbenchmark and a few bandwidth tests, and use this information to estimate relative performance of PUs. Using this, their technique dynamically changes the number and type (GPU or CPU) of PUs to achieve an optimal communication/computation ratio for achieving the best possible performance. They perform experiments using an HCS with a Xeon 5550 CPU and a Tesla M1060 GPU.

Scogland et al. [2012] propose a runtime system for automatically dividing an Accelerated OpenMP region across different PUs. The runtime also handles ensuring data transfer to PUs for input and output. Accelerated OpenMP uses `hetero` clause as a compiler directive, using which a programmer can control whether heterogeneous computing is used; and if so, how many loop iterations are assigned to a CPU. They propose enhancing this directive by also providing information on the type of scheduler (static or dynamic), ratio of performance of PUs, and a factor termed as *div* (explained shortly). The static scheduler uses a predetermined work division based on a specified ratio, while the dynamic scheduler updates this ratio after first execution of parallel region and uses this to make better decisions in future executions. To get an even better estimate of the value of the ratio, the *div* parameter is used to control the number of iterations to be used for getting the estimate of the ratio. Evaluations are performed using a 12-core Opteron 6174 CPU and a Tesla C2050 GPU.

Jiang and Agrawal [2012] present an approach called MATE-CG for accelerating MapReduce applications on parallel heterogeneous environments. Apart from allowing CPU-only and GPU-only execution, MATE-CG runtime also supports dividing the work between a CPU and a GPU. The amount of data to be processed by the CPU and the GPU is decided by a partitioning parameter, which can be decided at runtime using an autotuning approach based on the iterative characteristic of data-intensive applications. By collecting profiling data over the first few phases, the value of this parameter can be found with small overhead. The experimental system uses quad-core Xeon E5520 CPUs and Tesla C2050 GPUs.

Humphrey et al. [2012] present a Uintah runtime system, which provides a scheduler for heterogeneous computing. Uintah works on an abstract task-graph representation of parallel computation and communication to take into account data dependencies between tasks. Each task has a C++ method associated with it, which performs the actual computation. A GPU task is represented by an additional C++ method, which is used for setting up a GPU kernel and invocation. Uintah's heterogeneous task scheduler accounts for the dependencies of tasks. It also determines the order of execution and ensures correct interprocess communication. Further, it ensures that any variable is not simultaneously used by multiple running tasks. Further, by using CUDA asynchronous API, Uintah scheduler overlaps data transfer with computation on PUs, and also ensures that before a GPU task is run, the required data are present in the GPU memory. The experiments are performed on two systems, one of which has six-core Xeon X5660 CPUs with Tesla M2090 GPUs and another has 16-core Opteron 6200 CPUs with Tesla 20-series GPUs.

Odajima et al. [2012] propose a framework for heterogeneous computing using StarPU with XcalableMP-dev parallel programming language. XcalableMP is a directive-based language extension that supports parallel computing over different nodes in a distributed memory system. XcalableMP-dev extends XcalableMP for clusters equipped with heterogeneous nodes. StarPU is used as the execution engine of XcalableMP-dev. They demonstrate the use of their framework for the N-body problem, where the loops are distributed between PUs in proportion to their performance. Their experimental system uses Opteron 6134 CPUs and Tesla C2050 GPUs.

4.3. Compiler-Level Techniques

Luk et al. [2009] propose the Qilin framework for mapping computations on an HCS. It provides an API for writing parallelizable programs. Qilin uses a training phase, in which a performance model is created for each task on each PU. Each kernel version is executed with different inputs and a linear model is fitted to the observed runtimes. Using this information, optimal workload division is computed and dynamic compilation is done to instantiate the chosen distribution. Similar to OpenMP, Qilin is built on top of C/C++; however, unlike OpenMP, which can exploit parallelism only on CPUs, Qilin can exploit parallelism on both CPUs and GPUs. Their experimental machine uses a Core 2 Quad CPU along with a GeForce 8800 GTX GPU.

Ravi et al. [2010] propose a compiler and runtime framework for mapping a MapReduce class of applications to a system composed of multicore CPUs and GPUs. Their framework starts with simple C functions that has annotations added to it. Using this, their framework automatically generates the middleware API code for the CPU and GPU simultaneously. Afterward, the runtime system uses a work-sharing based approach to dynamically partition the work between CPU cores and the GPU. In work sharing, the scheduler adds the work in a global work list. An idle processor can obtain works from this list. They also observe that since GPUs incur large overhead of kernel invocation and data transfer, the chunk size of work allocated to them should be large. In contrast, CPU cores are relatively slower but also have smaller latency, and hence, to achieve better load balancing, the chunk size allocated to them should be small. Their evaluation system uses an Opteron 8350 CPU with a GeForce 9800 GTX GPU.

Prasad et al. [2011] propose a compiler for compiling MATLAB programs to achieve execution on heterogeneous processors. After identifying the data parallel regions of the program, the compiler composes them into kernels. Further, the identified kernels are mapped to a suitable PU so that the execution of kernels happens synergistically and the data transfer overhead is minimized. Their compiler does not require the users to perform manual annotation to specify the arrays whose operations are mapped to the GPU. Use of their compiler provides a performance advantage over native execution of MATLAB. Their experimental system uses a quad-core Xeon processor and one of the following GPUs: GeForce 8800 GTS or Tesla S1070. Their results show that, in general, Tesla S1070 provides higher speedup than GeForce 8800 GTS. They also note that a few large-sized problems could not be run on GeForce 8800 GTS due to its limited memory capacity.

Pienaar et al. [2012] propose an Automatic Heterogeneous Pipelining (AHP) framework for HCSs. AHP takes a C++ program that has been annotated with compiler directives by the programmer regarding targets of pipelining. Using program annotations, AHP extracts a task graph for each program region. It also profiles the tasks on PUs of HCS to estimate task execution time and intertask communication time. Then AHP works to iteratively identify pipeline stages and transforms the task graph to reflect the pipeline structure. Afterward, it maps the pipelined task graph to the HCS and generates code for the optimized heterogeneous pipeline that uses frameworks such as

Intel TBB, Cilk, OpenMP, and CUDA. Their evaluation platform uses a quad-core Xeon E5520 CPU and a NVIDIA Tesla C1060 GPU.

Kofler et al. [2013] propose a workload-division technique that uses an Insieme source-to-source compiler [Insieme Compiler 2014] to translate an OpenCL program for a single device into an OpenCL program for multiple devices. Afterward, workload partitioning is performed using on an offline-generated model, which is based on both static program features (e.g., number of branches, floating point operations, etc.) and dynamic program features (e.g., data transfer size, overhead, etc.), capabilities of the PU, and the input data. To achieve the best possible workload division, they use a Principle Component Analysis (PCA) method. The two HCSs used in their experiments have (1) an Opteron 6168 CPU with a Radeon HD5870 GPU and (2) a Xeon X5650 CPU with a GeForce GTX480 GPU.

4.4. Design of OpenCL and Related Frameworks

OpenCL [Stone et al. 2010] is an open standard and has been adopted by several vendors. It facilitates both task parallelism and data parallelism. For each PU, a command queue can be created to which tasks can be submitted. Use of parallel programming can be achieved on each PU individually and on all PUs collectively. It provides a higher abstraction programming framework and hence, it enables the programmer to write programs that are portable across a wide variety of processing units, although it may not be able to achieve the highest possible performance on a PU.

Spafford et al. [2010] present a library called Maestro for data orchestration on multiple OpenCL-enabled PUs. The OpenCL task queue model requires the programmer to manage separate task queues for each PU. This, in turn, requires the programmer to have detailed knowledge of each PU. Maestro uses a single high-level task queue and based on the runtime information, it can automatically transfer data and divide work among PUs. Maestro uses an autotuning approach to optimize OpenCL code on each PU and offline profiling to measure relative performance of each PU to decide optimal work division. Maestro also facilitates overlapping data transfer with computation to achieve pipelining. Their experimental HCS uses an Opteron 246 CPU with a Radeon HD 5870 GPU.

Kim et al. [2012] present SnuCL that enables OpenCL applications to run in a distributed manner on a cluster. SnuCL makes all the OpenCL compatible PUs (e.g., a CPU or a GPU) on a cluster logically appear on a single local node. SnuCL internally uses MPI but does not require the use of MPI in the application source. Using SnuCL, OpenCL applications written for a single node can run on the entire cluster without any modification. Thus, the application can utilize the PUs of a compute node similar to the PUs in the host node. This helps the programmer in utilizing all the PUs in a cluster for heterogeneous computing. Evaluations are performed on a CPU/GPU heterogeneous cluster whose compute node has a Xeon X5660 CPU with a GeForce GTX 480 GPU.

5. TECHNIQUES FOR IMPROVING ENERGY EFFICIENCY

Table V summarizes the techniques for saving energy in HCSs (DVFS = dynamic voltage/frequency scaling). Note that while many other techniques that aim to improve performance may also save energy, in this table we only show those techniques that have been evaluated on the basis of energy efficiency. We now discuss a few of these techniques.

5.1. Workload Partitioning Based Techniques

Liu and Luk [2012] propose an approach for utilizing CPUs, GPUs, and FPGAs for improving energy efficiency of scientific computing and demonstrate its use for

Table V. A Classification of Approaches Used for Energy Saving in HCSs

Classification	References
By intelligent workload partitioning and performance improvement	Choi et al. [2013], Hamano et al. [2009], Liu and Luk [2012], Liu et al. [2012], Luk et al. [2009], Ma et al. [2012], Rofouei et al. [2008], Silberstein and Maruyama [2011], and Takizawa et al. [2008]
By DVFS on both CPUs and GPUs	Liu et al. [2012], Ma et al. [2012], and Wang and Song [2011]
By DVFS on CPUs only	Anzt et al. [2011] and Liu et al. [2011]
By resource scaling or low-power modes	Liu et al. [2011] and Silberstein and Maruyama [2011]

High-Performance Linpack (HPL) benchmark. Their technique uses profiling to record the runtime power consumption and execution time (which includes data transfer time and computation time) of all PUs, along with the power consumption of the communication channel. Using these parameters, the problem of finding the right workload division among PUs while optimizing energy efficiency is modeled as a linear programming problem. Their experimental HCS has a Xeon W3505 CPU and a Tesla C2070 GPU.

Takizawa et al. [2008] propose Stream Programming with Runtime Autotuning (SPRAT), a runtime environment for improving energy efficiency by workload division. SPRAT uses a performance model that accounts for both the computation time of PUs and the communication time; and dynamically selects a PU for executing a kernel such that the system energy is minimized. Their experimental system uses a Core 2 Quad CPU and one of the three GPUs, which, in decreasing order of performance, are GeForce 8800 GTX, GeForce 8800 GT, and GeForce 8600 GTS.

Hamano et al. [2009] propose an HCT that schedules tasks to PUs with the aim of optimizing energy efficiency. Their technique takes into account the performance of tasks on different PUs and the energy consumption of PUs when they are busy and idle. Using this, their technique estimates the energy efficiency for different mapping of tasks to different PUs and chooses the mapping that leads to minimum energy consumption. Their evaluation platform has a Phenom 9850 quad-core CPU along with a GeForce 8800 GTS GPU.

5.2. DVFS Based Techniques

Wang and Song [2011] present a technique for saving energy in HCSs. Their technique models the problem of workload division and voltage scaling as an integer linear programming problem, with the objective of minimizing energy consumption (of both PUs and system buses) for a given performance constraint. Voltage scaling and workload division affect each other and also have an effect on the performance and energy consumption of PUs. Their model also accounts for the communication cost. With the optimal workload division, minimum execution time is obtained with highest voltage levels. Compared to this time, for different percentages of increase in execution time (i.e., performance loss), the energy consumption can be calculated, using which the trade-off between performance and energy consumption can be explored. They perform experiments on a machine equipped with a Core I7-920 quad-core CPU and a Radeon HD 4870 4870 GPU.

Anzt et al. [2011] propose a technique for saving energy in HCSs while executing iterative linear solvers. After starting a GPU kernel, the CPU stays in a busy-wait loop and performs no useful work. During this time, CPU energy can be saved by using Dynamic Voltage/Frequency Scaling (DVFS) with little performance loss. Further, when the execution time of the kernel is large enough, by noting that time once, the CPU can be transitioned to a sleep state for that duration in future iterations, that

is, future calls to the kernel since the solver takes nearly the same time in different iterations. Their experimental HCS uses an eight-core Opteron 6128 CPU and a Tesla C1060 GPU.

Ma et al. [2012] propose a framework for power management of HCSs. Their technique divides the workload between a CPU and a GPU based on workload characteristics to achieve load balancing and reduce idling. Afterward, the frequency and voltage of the CPU and the frequency of the GPU are adjusted to achieve energy savings with minimal performance degradation. Their experimental system uses a dual-core AMD Phenom II X2 CPU along with a GeForce 8800 GTX GPU.

Liu et al. [2012] discuss a workload-division technique for saving energy in an HCS while also meeting task deadlines. Their technique first maps tasks on a CPU or a GPU, such that their deadlines are met and load balancing can be achieved, and then applies DVFS to both PUs to save energy. When the average-case execution times are shorter than worst-case execution times, extra slack is generated that can be further exploited using DVFS to save extra energy. Their experiments are performed using a Xeon 5160 CPU and a Radeon HD 5770 GPU.

5.3. Resource Scaling Based Techniques

Liu et al. [2011] propose a technique based on a waterfall model for improving the energy efficiency of large-scale heterogeneous clusters, in which each node may have several CPU-GPU pairs. Their technique transitions the node into one of the three possible power states, namely, busy (all CPUs and GPUs of a node are working), spare (at least a single CPU-GPU pair is free), and sleep (all CPU-GPU pairs are free). At the time of reduced workload, a node in the sleep state is powered off and at the time of additional workload, a node is woken up. Further, their technique schedules the tasks on an available CPU-GPU pair with the view to minimize their execution time, and scales the voltage of the CPU to save energy while meeting task deadlines. Finally, to achieve load balancing, their technique migrates a small fraction of the GPU's share of task to the CPU in the same CPU-GPU pair when required. Their performance models are based on a Tianhe-1A supercomputer, which uses a six-core Xeon X5670 CPU and a Tesla M2050 GPU.

Silberstein and Maruyama [2011] present an algorithm for optimizing the energy efficiency of an HCS while running applications comprised of multiple interdependent tasks. Their algorithm takes into account the energy consumption of each task on a CPU and a GPU, along with the data transfer cost and constructs a schedule with provably minimal total consumed energy. They assume that both a CPU and a GPU can be powered off when idle and incur no overhead when they are powered on. Their experimental platform uses a Core 2 Quad CPU and a GeForce GTX 285 GPU.

6. FUSED CPU-GPU CHIPS AND COMPARISON WITH DISCRETE SYSTEMS

Discrete GPUs have separate memory spaces from the CPU and hence, data transfer happens over a connection bus (e.g., PCIe bus), which incurs large overhead. Also, these systems require large programming effort, since the programmer has to manage the data that is manipulated by both the CPU and the GPU. To address these limitations, researchers have designed fused HCSs that feature shared memory space between the CPU and the GPU. This reduces their data transfer time, which especially benefits the applications that require large communication between the CPU and the GPU. Table VI summarizes the works that compare fused CPU-GPU chips with discrete CPU-GPU systems. We now discuss a few of these works.

Delorme [2013] present two strategies for implementing radix sort on the fused HCS. The coarse-grained implementation divides the data to be sorted between a CPU and a GPU at the beginning of algorithm. Afterward, they individually sort the

Table VI. Classification of Works Related to Fused HCSs

Classification	References
Comparison of fused HCSs with discrete HCSs	Boyer et al. [2013], Conti et al. [2012], Daga et al. [2011], Delorme [2013], Hetherington et al. [2012], Lee et al. [2013a], Spafford et al. [2012], Sun et al. [2012], and Ukidave et al. [2013]
Data transfer overhead evaluation	Boyer et al. [2013], Conti et al. [2012], Daga et al. [2011], Delorme [2013], Hetherington et al. [2012], Lee et al. [2013a], and Spafford et al. [2012]
Energy efficiency evaluation	Spafford et al. [2012] and Ukidave et al. [2013]

data and finally, their outputs are merged to produce single sorted output. The fine-grained implementation with dynamic partitioning shares data after each pass of the sorting algorithm. In this implementation, algorithm repartitions data after each step and thus, a piece of data does not belong to a specific PU, rather the entire data must be visible to both PUs during kernel execution. This implementation requires more communication and synchronization than the coarse-grained implementation; however, it also provides better performance due to better load balancing. He also shows that both these implementations outperform the state-of-the-art GPU radix sort implementation, which shows the merit of utilizing a CPU for performing computations in an HCS. They perform experiments using two APUs, viz., an AMD A6-3650 APU (with a quad-core AMD CPU and a Radeon HD 6530D GPU) and an AMD A8-3850 APU (with a quad-core AMD CPU and a Radeon HD 6550D GPU), of which the A8-3850 APU shows higher performance.

Hetherington et al. [2012] compare the performance of Memcached (a key-value store application) on discrete HCS with that on fused HCS. They observe that on discrete systems, the time of data transfer between the CPU and the GPU becomes a significant fraction of the total execution time, which negatively affects the performance of the GPU. On excluding the data transfer overhead, however, the discrete GPU provides a large performance gain. Their experiments use two fused HCSs, viz., a Llano A8-3850 APU (with a Radeon HD 6550D GPU) and a Zacate E-350 (with a Radeon HD 6310 GPU) and one Radeon HD 5870 discrete GPU. They note that the fused chips such as Llano and Zacate have lower computation power compared to the discrete GPU; however, their ability to avoid the transfer of data enables them to provide higher performance than the discrete GPU chip.

Similarly, Spafford et al. [2012] show that low CPU-GPU data transfer overhead is a key advantage of fused HCS, which makes its energy efficiency better than that of a discrete GPU. At the same time, the penalty of contention between a CPU and a GPU is higher for a fused HCS (due to fused memory hierarchy) than for a discrete GPU. Further, compared to a discrete CPU with a nearly similar power budget, the fused HCS uses a simpler CPU microarchitecture design and hence, it provides lower performance on computationally intensive tasks than the discrete CPU. Their experiments use a A8-3850 Llano APU as the fused HCS and a Radeon HD 5670 and a FirePro v8800 as two discrete GPUs.

Fused HCSs (or APUs) replace the PCIe data transfer cost by fast memory block transfers between the CPU and GPU memory partitions. Daga et al. [2011] show that due to this, for a kernel benchmark, an AMD E-series Zacate APU with only 80 GPU cores provides more than 3× improvement over a discrete AMD Radeon HD 5870 GPU that has a total of 1600 more powerful GPU cores. They also show that compared to a discrete GPU, an APU reduces the parallelization overhead.

Lee et al. [2013a] compare the performance of memory accesses on a fused HCS and a discrete GPU for different memory access patterns. They note that on the fused HCS, a CPU memory access from the GPU is nearly as fast as a GPU memory access. This provides the opportunity for the GPU to directly access data on the CPU without

copying, and it especially benefits the applications with little data reuse. For their experiments, they use an A8-3850 APU as the fused HCS and a Radeon HD5870 as the discrete GPU.

Ukidave et al. [2013] study the performance and energy consumption of discrete and fused HCSs for different FFT implementations with different input sizes. They show that with growing input data size, the energy efficiency of these HCSs increases due to better utilization of the data-parallel resources on the GPUs. They also show that the power consumption increases with the number of OpenCL kernel calls and increased use of the GPU fetch unit. They perform experiments using two discrete GPUs, viz., a GeForce GTX 480 and a Radeon HD 7770, and two fused systems, viz., an AMD A8-3850 APU and a Core i7-3770 Ivy Bridge processor (which uses Intel Series-4000 embedded GPU).

7. APPLICATION AREAS OF AND BENCHMARKS FOR HETEROGENEOUS COMPUTING

7.1. Application Domains of Heterogeneous Computing

Table VII classifies the works discussed in this article based on their application domains. These domains are by no means exhaustive, as heterogeneous computing can be applied in several other domains that are also computationally intensive and/or provide the opportunity for parallelization.

7.2. Benchmarks for Heterogeneous Computing

Due to their unique characteristics, HCSs require special benchmarks to gain insights into their functioning. To fulfill this need, several benchmark suites have been proposed that help in meaningfully comparing their architectures and programming environments against similar systems. Table VIII summarizes these benchmarks. Notice that some of these benchmarks provide implementations in different languages which allow comparison using those programming models. We now briefly discuss these benchmark suites.

SHOC (Scalable HeterOgeneous Computing) benchmark [Danalis et al. 2010] provides both low-level microbenchmarks (to evaluate architectural features of the system) and application kernels (to evaluate the features of the system such as intranode and internode communication between PUs). In addition to the serial version, SHOC provides an embarrassingly parallel version (which executes on different PUs or nodes of a cluster, but have no communication between PUs or nodes), and a true parallel version (which measures multiple nodes, with single or multiple PUs per node, and also involves communication). Thus, SHOC benchmarks scale effectively across a single PU to a large cluster. For the same fundamental algorithm, Parboil [Stratton et al. 2012] provides versions of varying levels of optimizations. This feature enables the compiler writers to evaluate source and compiler optimizations on different architectures.

The choice of programs in Rodinia [Che et al. 2009] is based on Berkeley's dwarf taxonomy and is aimed at covering different types of parallel communication patterns and synchronization techniques. Valar [Mistry et al. 2013b] benchmark suite provides OpenCL applications for studying interaction of processing units in HCSs. Programs in this suite are categorized based on whether the majority of algorithm execution is done on a single PU, or on multiple PUs with or without substantial communication. Also, the programs are characterized by whether their behavior is latency sensitive, streaming type, or has a Quality-of-Service (QoS) requirement.

8. CONCLUSION AND FUTURE DIRECTIONS

In recent years, CPUs and GPUs are increasingly being seen as indispensable coprocessors, instead of substitutes for each other. As a result, heterogeneous computing has

Table VII. Classification of Research Works Based on their Application (or Workload) Domains

Classification	References
Maths, numerical methods, and/or algebraic routines	Agullo et al. [2011], Álvarez-Melcón et al. [2013], Anzt et al. [2011], Becchi et al. [2010], Benner et al. [2010], Benner et al. [2011], Bernabé et al. [2013], Binotto et al. [2010], Boyer et al. [2013], Clarke et al. [2012], Conti et al. [2012], Daga et al. [2011], Danalis et al. [2010], Diamos and Yalamanchili [2008], Dziekonski et al. [2011], Endo et al. [2010], Gregg et al. [2010], Gregg et al. [2011], Gummaraju et al. [2010], Hong et al. [2010], Horton et al. [2011], Jiménez et al. [2009], Kim et al. [2012], Kofler et al. [2013], Kothapalli et al. [2013], Lang and Rünger [2013], Lee et al. [2012], Li et al. [2012], Liu and Luk [2012], Ltaief et al. [2011], Luo et al. [2011], Ma et al. [2013], Mariano et al. [2012], Matam et al. [2012], Meredith et al. [2011], Muramatsu et al. [2011], Ogata et al. [2008], Ohshima et al. [2007], Pai et al. [2010], Pandit and Govindarajan [2014], Papadrakakis et al. [2011], Pienaar et al. [2011], Prasad et al. [2011], Scogland et al. [2012], Siegel et al. [2010], Spafford et al. [2012], Stefanski [2013], Stpiczynski [2011], Stpiczynski and Potiopa [2010], Su et al. [2013], Takizawa et al. [2008], Tomov et al. [2010], Ukidave et al. [2013], Veldema et al. [2011], Venkatasubramanian and Vuduc [2009], Vömel et al. [2012], Wang et al. [2013b], Wen et al. [2012], Yang et al. [2010], and Zhong et al. [2012]
Video processing, imaging, and/or computer vision	Agulleiro et al. [2012], Choudhary et al. [2012], Deshpande et al. [2011], Hartley et al. [2008, 2010], Lecron et al. [2011], Li et al. [2011], Liu et al. [2009], Mistry et al. [2013a], Nigam and Narayanam [2012], Pajot et al. [2011], Park et al. [2011], Pienaar et al. [2012], So et al. [2011], Teodoro et al. [2009], Teodoro et al. [2012], Teodoro et al. [2013], Toharia et al. [2012], Tsuda and Nakamura [2011], and Wang et al. [2013a]
Data mining, processing, and/or database systems	Banerjee and Kothapalli [2011], Banerjee et al. [2012], Becchi et al. [2010], Breß et al. [2013], Chen et al. [2012], Delorme [2013], Gelado et al. [2010], Gharaibeh et al. [2012], He and Hong [2010], Hetherington et al. [2012], Hong et al. [2011], Jablin et al. [2012], Kothapalli et al. [2013], Lee et al. [2013a], Liu et al. [2011], Munguia et al. [2012], Pandit and Govindarajan [2014], Pienaar et al. [2012], Pirk et al. [2012], Ravi et al. [2010], and Shirahata et al. [2010]
Physics	Hardy et al. [2009], Hawick and Playne [2013], Hermann et al. [2010], Humphrey et al. [2012], Jetley et al. [2010], Joselli et al. [2008], Korwar et al. [2013], Lu et al. [2012a], Nakasato et al. [2012], Panetta et al. [2009], Rahimian et al. [2010], Teodoro et al. [2013], Wu et al. [2012], and Yang et al. [2013]
Chemistry	Bhaskaran-Nair et al. [2013], Hampton et al. [2010], and Ma et al. [2013]
Bioinformatics and/or medical science	Agulleiro et al. [2012], Chai et al. [2013], Chen et al. [2010], Hartley et al. [2008, 2010], Lecron et al. [2011], Li et al. [2011], Liu et al. [2009], Rahimian et al. [2010], Shen et al. [2010], Singh and Aruni [2011], So et al. [2011], Wang et al. [2013a], Xiao et al. [2011], and Yao et al. [2010]
Smith-Waterman algorithm	Chen et al. [2010], Luk et al. [2009], and Singh and Aruni [2011]
N-body simulations	Hu et al. [2011], Jetley et al. [2010], Joselli et al. [2008], Nakasato et al. [2012], Rahimian et al. [2010], and Tsoi and Luk [2010]
Electromagnetics	Álvarez-Melcón et al. [2013], Dziekonski et al. [2011], Gao et al. [2012], and Stefanski [2013]

been actively researched and utilized in a variety of applications ranging from natural sciences to engineering, etc. Several challenges still remain, and we believe that research in the near future will provide solutions to them. We now briefly mention some directions for future research.

Many algorithm-level techniques require the programmer to manually partition the workload between PUs, identify the subtasks suitable for each PU, and/or profile the performance of each PU. This is a tedious process that does not scale well beyond small

Table VIII. A Classification of Benchmark Suites for Heterogeneous Computing

Name	Languages/Versions	Scaled Using MPI?	Example Application Fields
SHOC	Serial, OpenCL, and CUDA	Yes	Scientific computing, linear algebra, molecular dynamics, etc.
Parboil	Serial, OpenMP, OpenCL, and CUDA	No	Image processing, biomolecular simulation, astronomy, and fluid dynamics
Rodinia	OpenCL, OpenMP, and CUDA	No	Medical imaging, bioinformatics, data mining, and molecular dynamics
Valar	OpenCL	No	Computer vision, digital signal processing, data mining, and physics

problems. Fully automatic toolchains are required to extend the benefits of heterogeneous computing to a wide range of application domains.

Several existing large-scale codes are written in either CPU or GPU languages. Porting them to HCSs will incur large overhead and is also error-prone. A significant amount of work needs to be done before recently introduced HCS programming frameworks can reach the level of maturity, robustness, and popularity enjoyed by the conventional CPU programming languages.

Mobile computing systems, which now number nearly equal to the population of the Earth [International Telecommunication Union 2012], generally provide multimedia services under real-time performance constraints and hence, they are attractive targets for acceleration through the use of heterogeneous computing. However, mobile systems also impose very stringent cost and power budgets [Mittal 2014c]. While the large cost and power consumption of HCSs may be acceptable in high-end computing systems, using them in mobile systems calls for an order-of-magnitude improvement in their energy efficiency.

While fused HCSs address several limitations of discrete HCSs, they also introduce new trade-offs and require the design of suitable data-access strategies to take full advantage of their fast interconnection. Moreover, existing fused HCSs are not strictly superior to discrete GPUs on all parameters and hence, despite their potential, extensive evaluation and development on them has been lacking. Novel solutions are required at device, architecture, programming, and system levels before the fused HCSs can become commonplace in mainstream computing systems.

In this article, we have synthesized the research work on heterogeneous computing by showing the broad spectrum of heterogeneous computing techniques and their key research ideas and applications. We surveyed research works at different levels of abstraction, viz., algorithm, runtime, compiler, programming model, etc. We discussed both both fused and discrete CPU-GPU systems along with the benchmark suits proposed to evaluate them. It is hoped that this article will be highly beneficial to computer architects, researchers, and application developers and will inspire novel ideas and open promising research avenues.

REFERENCES

Alejandro Acosta, Robert Corujo, Vicente Blanco, and Francisco Almeida. 2010. Dynamic load balancing on heterogeneous multicore/multi-GPU systems. In *International Conference on High Performance Computing and Simulation (HPCS)*. 467–476.

Jose Ignacio Agulleiro, Francisco Vazquez, Ester M. Garzon, and Jose J. Fernandez. 2012. Hybrid computing: CPU+ GPU co-processing and its application to tomographic reconstruction. *Ultramicroscopy* 115 (2012), 109–114.

Emmanuel Agullo, Cédric Augonnet, Jack Dongarra, Mathieu Faverge, Hatem Ltaief, Samuel Thibault, and Stanimire Tomov. 2011. QR factorization on a multicore node enhanced with multiple GPU accelerators. *IEEE International Parallel & Distributed Processing Symposium*, 932–943.

Omer Erdil Albayrak, Ismail Akturk, and Ozcan Ozturk. 2012. Effective kernel mapping for OpenCL applications in heterogeneous platforms. In *41st International Conference on Parallel Processing Workshops (ICPPW)*. IEEE, 81–88.

Alejandro Álvarez-Melcón, Domingo Giménez, Fernando D. Quesada, and Tomás Ramírez. 2013. Hybrid-parallel algorithms for 2D Green's functions. *Procedia Computer Science* 18 (2013), 541–550.

Hartwig Anzt, Vincent Heuveline, José I. Aliaga, Maribel Castillo, Juan C. Fernandez, Rafael Mayo, and Enrique S. Quintana-Orti. 2011. Analysis and optimization of power consumption in the iterative solution of sparse linear systems on multi-core and many-core platforms. In *International Green Computing Conference and Workshops (IGCC)*. IEEE, 1–6.

Cédric Augonnet, Samuel Thibault, Raymond Namyst, and Pierre-André Wacrenier. 2011. StarPU: A unified platform for task scheduling on heterogeneous multicore architectures. *Concurrency and Computation: Practice and Experience* 23, 2 (2011), 187–198.

Eduard Ayguade, Rosa M. Badia, Daniel Cabrera, Alejandro Duran, Marc Gonzalez, Francisco Igual, Daniel Jimenez, Jesus Labarta, Xavier Martorell, Rafael Mayo, Josep M. Perez, and Enrique S. Quintana-Ortí. 2009. A proposal to extend the OpenMP tasking model for heterogeneous architectures. In *Evolving OpenMP in an Age of Extreme Parallelism*. 154–167.

Ana Balevic and Bart Kienhuis. 2011. An efficient stream buffer mechanism for dataflow execution on heterogeneous platforms with GPUs. In *First Workshop on Data-Flow Execution Models for Extreme Scale Computing (DFM)*. IEEE, 53–57.

Dip Sankar Banerjee, Aman Kumar Bahl, and Kishore Kothapalli. 2012. An on-demand fast parallel pseudo random number generator with applications. In *InternationalParallel & Distributed Processing Symposium Workshops & PhD Forum (IPDPSW)*. 1703–1711.

Dip Sankar Banerjee and Kishore Kothapalli. 2011. Hybrid algorithms for list ranking and graph connected components. In *International Conference on High Performance Computing*. 1–10.

Michela Becchi, Surendra Byna, Srihari Cadambi, and Srimat Chakradhar. 2010. Data-aware scheduling of legacy kernels on heterogeneous platforms with distributed memory. In *22nd ACM Symposium on Parallelism in Algorithms and Architectures*. 82–91.

Mehmet E. Belviranli, Laxmi N. Bhuyan, and Rajiv Gupta. 2013. A dynamic self-scheduling scheme for heterogeneous multiprocessor architectures. *ACM Transactions on Architecture and Code Optimization (TACO)* 9, 4 (2013), 57.

Peter Benner, Pablo Ezzatti, Daniel Kressner, Enrique S. Quintana-Ortí, and Alfredo Remón. 2011. A mixed-precision algorithm for the solution of Lyapunov equations on hybrid CPU–GPU platforms. *Parallel Computing* 37, 8 (2011), 439–450.

Peter Benner, Pablo Ezzatti, Enrique S. Quintana-Ortí, and Alfredo Remón. 2010. Using hybrid CPU-GPU platforms to accelerate the computation of the matrix sign function. In *Euro-Par 2009—Parallel Processing Workshops*. 132–139.

Gregorio Bernabé, Javier Cuenca, and Domingo Giménez. 2013. Optimization techniques for 3D-FWT on systems with manycore GPUs and multicore CPUs. *Procedia Computer Science* 18, 319–328.

Kiran Bhaskaran-Nair, Wenjing Ma, Sriram Krishnamoorthy, Oreste Villa, Hubertus J. J. van Dam, Edoardo Aprà, and Karol Kowalski. 2013. Non-iterative multireference coupled cluster methods on heterogeneous CPU-GPU systems. *Journal of Chemical Theory and Computation* 9, 4 (2013), 1949–1957.

Alecio P. D. Binotto, Christian Daniel, Daniel Weber, Arjan Kuijper, Andre Stork, Carlos Pereira, and Dieter Fellner. 2010. Iterative SLE solvers over a CPU-GPU platform. In *International Conference on High Performance Computing and Communications*. 305–313.

Alecio P. D. Binotto, Carlos E. Pereira, Arjan Kuijper, Andre Stork, and Dieter W. Fellner. 2011. An effective dynamic scheduling runtime and tuning system for heterogeneous multi and many-core desktop platforms. In *IEEE 13th International Conference on High Performance Computing and Communications (HPCC)*. 78–85.

Murilo Boratto, Pedro Alonso, Carla Ramiro, and Marcos Barreto. 2012. Heterogeneous computational model for landform attributes representation on multicore and multi-GPU systems. *Procedia Computer Science* 9 (2012), 47–56.

Michael Boyer, Kevin Skadron, Shuai Che, and Nuwan Jayasena. 2013. Load balancing in a changing world: Dealing with heterogeneity and performance variability. In *ACM International Conference on Computing Frontiers*.

Alexander Branover, Denis Foley, and Maurice Steinman. 2012. AMD fusion APU: Llano. *IEEE Micro,* 32, 2 (2012), 28–37.

Sebastian Breß, Felix Beier, Hannes Rauhe, Kai-Uwe Sattler, Eike Schallehn, and Gunter Saake. 2013. Efficient co-processor utilization in database query processing. *Information Systems* 38, 8 (2013), 1084–1096.

Jun Chai, Huayou Su, Mei Wen, Xing Cai, Nan Wu, and Chunyuan Zhang. 2013. Resource-efficient utilization of CPU/GPU-based heterogeneous supercomputers for Bayesian phylogenetic inference. *The Journal of Supercomputing* 66, 1 (2013), 364–380.

Shuai Che, Michael Boyer, Jiayuan Meng, David Tarjan, Jeremy W. Sheaffer, Sang-Ha Lee, and Kevin Skadron. 2009. Rodinia: A benchmark suite for heterogeneous computing. In *IEEE International Symposium on Workload Characterization*. 44–54.

Bo Chen, Yun Xu, Jiaoyun Yang, and Haitao Jiang. 2010. A new parallel method of Smith-Waterman algorithm on a heterogeneous platform. In *Algorithms and Architectures for Parallel Processing*. Springer, 79–90.

Linchuan Chen, Xin Huo, and Gagan Agrawal. 2012. Accelerating mapreduce on a coupled CPU-GPU architecture. In *SC'12 Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. 25:1–25:11.

Hong Jun Choi, Dong Oh Son, Seung Gu Kang, Jong Myon Kim, Hsien-Hsin Lee, and Cheol Hong Kim. 2013. An efficient scheduling scheme using estimated execution time for heterogeneous computing systems. *The Journal of Supercomputing* 65, 2 (2013), 886–902.

Siddharth Choudhary, Shubham Gupta, and P. J. Narayanan. 2012. Practical time bundle adjustment for 3D reconstruction on the GPU. In *Trends and Topics in Computer Vision*, Lecture Notes in Computer Science, Volume 6554. Springer, Berlin, 423–435.

David Clarke, Aleksandar Ilic, Alexey Lastovetsky, and Leonel Sousa. 2012. Hierarchical partitioning algorithm for scientific computing on highly heterogeneous CPU+ GPU clusters. In *Euro-Par Parallel Processing*, Lecture Notes in Computer Science, Volume 7484. Springer, Berlin, 489–501.

Christian Conti, Diego Rossinelli, and Petros Koumoutsakos. 2012. GPU and APU computations of finite time lyapunov exponent fields. *Jouranl of Computational Physics* 231, 5 (2012), 2229–2244.

J. R. da S. Junior, Esteban W. Clua, Anselmo Montenegro, and Paulo A. Pagliosa. 2010. Fluid simulation with two-way interaction rigid body using a heterogeneous GPU and CPU environment. In *Brazilian Symposium on Games and Digital Entertainment (SBGAMES)*. IEEE, 156–164.

Mayank Daga, Ashwin M. Aji, and Wu-chun Feng. 2011. On the efficacy of a fused CPU+ GPU processor (or APU) for parallel computing. In *Symposium on Application Accelerators in High-Performance Computing (SAAHPC)*. IEEE, 141–149.

Satish Damaraju, Varghese George, Sanjeev Jahagirdar, Tanveer Khondker, Robert Milstrey, Sanjib Sarkar, Scott Siers, Israel Stolero, and Arun Subbiah. 2012. A 22nm IA multi-CPU and GPU system-on-chip. In *IEEE International Solid-State Circuits Conference Digest of Technical Papers*. 56–57.

Anthony Danalis, Gabriel Marin, Collin McCurdy, Jeremy S. Meredith, Philip C. Roth, Kyle Spafford, Vinod Tipparaju, and Jeffrey S. Vetter. 2010. The scalable heterogeneous computing (SHOC) benchmark suite. In *Proceedings of the 3rd Workshop on General-Purpose Computation on Graphics Processing Units (GPGPU'10)*. 63–74.

Michael Christopher Delorme. 2013. *Parallel Sorting on the Heterogeneous AMD Fusion Accelerated Processing Unit*. Master of Applied Science Thesis, University of Toronto.

Aditya Deshpande, Ishan Misra, and P. J. Narayanan. 2011. Hybrid implementation of error diffusion dithering. In *Proceedings of the 2011 18th International Conference on High Performance Computing*. 1–10.

Gregory F. Diamos and Sudhakar Yalamanchili. 2008. Harmony: An execution model and runtime for heterogeneous many core systems. In *Proceedings of the 17th International Symposium on High Performance Distributed Computing (HPDC'08)*. 197–200.

Shuai Ding, Jinru He, Hao Yan, and Torsten Suel. 2009. Using graphics processors for high performance IR query processing. In *Proceedings of the 18th International Conference on World Wide Web (WWW'09)*. 421–430.

Adam Dziekonski, Adam Lamecki, and Michal Mrozowski. 2011. Tuning a hybrid GPU-CPU V-Cycle multilevel preconditioner for solving large real and complex systems of FEM equations. *IEEE Antennas and Wireless Propagation Letters* 10 (2011), 619–622.

Toshio Endo, Akira Nukada, Satoshi Matsuoka, and Naoya Maruyama. 2010. Linpack evaluation on a supercomputer with heterogeneous accelerators. In *2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS'10)*. 1–8.

Eric J. Fluhr, Joshua Friedrich, Daniel Dreps, Victor Zyuban, Gregory Still, Christopher Gonzalez, Allen Hall, David Hogenmiller, Frank Malgioglio, Ryan Nett, and others. 2014. 5.1 POWER8^{TM}: A 12-core server-class processor in 22nm SOI with 7.6 Tb/s off-chip bandwidth. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC'14)*. 96–97.

Peng Cheng Gao, Yu Bo Tao, Zhi Hui Bai, and Hai Lin. 2012. Mapping the SBR and TW-ILDCs to heterogeneous CPU-GPU architecture for fast computation of electromagnetic scattering. *Progress In Electromagnetics Research* 122 (2012), 137–154.

Michael T. Garba and Horacio González-vélez. 2012. Asymptotic peak utilisation in heterogeneous parallel CPU/GPU pipelines: A decentralised queue monitoring strategy. *Parallel Processing Letters* 22, 2 (2012).

Eric Gardner. 2014. https://software.intel.com/en-us/articles/what-disclosures-has-intel-made-about-knights-landing.

Isaac Gelado, John E. Stone, Javier Cabezas, Sanjay Patel, Nacho Navarro, and Wen-mei W. Hwu. 2010. An asymmetric distributed shared memory model for heterogeneous parallel systems. In *ACM SIGARCH Computer Architecture News*, 38 1 (March 2010), 347–358.

Abdullah Gharaibeh, Lauro Beltrão Costa, Elizeu Santos-Neto, and Matei Ripeanu. 2012. A yoke of oxen and a thousand chickens for heavy lifting graph processing. In *Proceedings of the 21st International Conference on Parallel Architectures and Compilation Techniques (PACT'12)*. ACM, New York, NY, 345–354.

Green500. 2014. Green500 Supercomputers. Retrieved from www.green500.org.

Chris Gregg, Michael Boyer, Kim Hazelwood, and Kevin Skadron. 2011. Dynamic heterogeneous scheduling decisions using historical runtime data. In *Proceedings of the 2nd Workshop on Applications for Multi- and Many-Core Processors*.

Chris Gregg, Jeff Brantley, and Kim Hazelwood. 2010. Contention-aware scheduling of parallel code for heterogeneous systems. In *Proceedings of the USENIX Workshop on Hot Topics in Parallelism (HotPar'10)*.

Chris Gregg and Kim Hazelwood. 2011. Where is the data? Why you cannot debate CPU vs. GPU performance without the answer. In *Proceedings of the 2011 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'11)*. 134–144.

Dominik Grewe and Michael F. P. O'Boyle. 2011. A static task partitioning approach for heterogeneous systems using OpenCL. In *Proceedings of the 20th International Conference on Compiler Construction: Part of the Joint European Conferences on Theory and Practice of Software*. Springer, Berlin, 286–305.

Jayanth Gummaraju, Laurent Morichetti, Michael Houston, Ben Sander, Benedict R. Gaster, and Bixia Zheng. 2010. Twin peaks: A software platform for heterogeneous computing on general-purpose and graphics processors. In *Proceedings of the 19th International Conference on Parallel Architectures and Compilation Techniques (PACT'10)*. ACM, New York, NY, 205–216.

Sumit Gupta. 2014. http://blogs.nvidia.com/blog/2014/03/25/gpu-roadmap-pascal/.

Tomoaki Hamano, Toshio Endo, and Satoshi Matsuoka. 2009. Power-aware dynamic task scheduling for heterogeneous accelerated clusters. In *Proceedings of the IEEE International Symposium on Parallel & Distributed Processing*. 1–8.

Scott S. Hampton, Sadaf R. Alam, Paul S. Crozier, and Pratul K. Agarwal. 2010. Optimal utilization of heterogeneous resources for biomolecular simulations. In *Proceedings of the 2010 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*. 1–11.

David J. Hardy, John E. Stone, and Klaus Schulten. 2009. Multilevel summation of electrostatic potentials using graphics processing units. *Parallel Comput.* 35, 3 (2009), 164–177.

Timothy D. R. Hartley, Umit Catalyurek, Antonio Ruiz, Francisco Igual, Rafael Mayo, and Manuel Ujaldon. 2008. Biomedical image analysis on a cooperative cluster of GPUs and multicores. In *Proceedings of the 22nd Annual International Conference on Supercomputing (ICS'08)*. 15–25.

Timothy D. R. Hartley, Erik Saule, and Umit V. Catalyurek. 2010. Automatic dataflow application tuning for heterogeneous systems. In *Proceedings of the 2010 International Conference on High Performance Computing (HiPC'10)*. 1–10.

Kenneth Arthur Hawick and Daniel P. Playne. 2013. Parallel algorithms for hybrid multi-core CPU-GPU implementations of component labelling in critical phase models. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'13)*. 45–51.

Zhengyu He and Bo Hong. 2010. Dynamically tuned push-relabel algorithm for the maximum flow problem on CPU-GPU-hybrid platforms. In *Proceedings of the 2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS'10)*. 1–10.

Everton Hermann, Bruno Raffin, François Faure, Thierry Gautier, and Jérémie Allard. 2010. Multi-GPU and multi-CPU parallelization for interactive physics simulations. In *Proceedings of the Euro-Par 2010-Parallel Processing*, Lecture Notes in Computer Science, Volume 6272. Springer, Berlin, 235–246.

Tayler H. Hetherington, Timothy G. Rogers, Lisa Hsu, Mike O'Connor, and Tor M. Aamodt. 2012. Characterizing and evaluating a key-value store application on heterogeneous CPU-GPU systems. In *Proceedings of the 2012 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'12)*. 88–98.

Chuntao Hong, Dehao Chen, Wenguang Chen, Weimin Zheng, and Haibo Lin. 2010. MapCG: Writing parallel program portable between CPU and GPU. In *Proceedings of the 19th International Conference on Parallel Architectures and Compilation Techniques (PACT'10)*. 217–226.

Sungpack Hong, Tayo Oguntebi, and Kunle Olukotun. 2011. Efficient parallel graph exploration on multi-core CPU and GPU. In *International Conference on Parallel Architectures and Compilation Techniques (PACT'11)*. 78–88.

Mitch Horton, Stanimire Tomov, and Jack Dongarra. 2011. A class of hybrid lapack algorithms for multicore and GPU architectures. In *Symposium on Application Accelerators in High-Performance Computing (SAAHPC'11)*. IEEE, 150–158.

Qi Hu, Nail A. Gumerov, and Ramani Duraiswami. 2011. Scalable fast multipole methods on distributed heterogeneous architectures. In *Proceedings of the 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM, New York, NY, Article 36.

Alan Humphrey, Qingyu Meng, Martin Berzins, and Todd Harman. 2012. Radiation modeling using the Uintah heterogeneous CPU/GPU runtime system. In *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the Campus and Beyond*. ACM, New York, NY, Article 4.

Xin Huo, Vignesh T. Ravi, and Gagan Agrawal. 2011. Porting irregular reductions on heterogeneous CPU-GPU configurations. In *Proceedings of the 18th International Conference on High Performance Computing*. 1–10.

Insieme Compiler. 2014. http://www.dps.uibk.ac.at/insieme/index.html.

International Telecommunication Union. 2012. Retrieved from http://www.itu.int/dms_pub/itu-d/opb/ind/D-IND-ICTOI-2012-SUM-PDF-E.pdf.

Thomas B. Jablin, James A. Jablin, Prakash Prabhu, Feng Liu, and David I. August. 2012. Dynamically managed data for CPU-GPU architectures. In *Proceedings of the 10th International Symposium on Code Generation and Optimization (CGO'12)*. 165–174.

Pritish Jetley, Lukasz Wesolowski, Filippo Gioachin, Laxmikant V. Kalé, and Thomas R. Quinn. 2010. Scaling hierarchical N-body simulations on GPU clusters. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–11.

Wei Jiang and Gagan Agrawal. 2012. Mate-CG: A map reduce-like framework for accelerating data-intensive computations on heterogeneous clusters. In *Proceedings of the IEEE 26th International Parallel & Distributed Processing Symposium (IPDPS'12)*. 644–655.

Víctor J. Jiménez, Lluís Vilanova, Isaac Gelado, Marisa Gil, Grigori Fursin, and Nacho Navarro. 2009. Predictive runtime code scheduling for heterogeneous architectures. In *High Performance Embedded Architectures and Compilers*, Lecture Notes in Computer Science, Volume 5409. Springer, Berlin, 19–33.

Mark Joselli, Marcelo Zamith, Esteban Clua, Anselmo Montenegro, Aura Conci, Regina Leal-Toledo, Luis Valente, Bruno Feijó, Marcos d'Ornellas, and Cesar Pozzer. 2008. Automatic dynamic task distribution between CPU and GPU for real-time systems. In *Proceedings of the 11th IEEE International Conference on Computational Science and Engineering*. 48–55.

Jungwon Kim, Sangmin Seo, Jun Lee, Jeongho Nah, Gangwon Jo, and Jaejin Lee. 2012. SnuCL: An OpenCL framework for heterogeneous CPU/GPU clusters. In *Proceedings of the 26th ACM International Conference on Supercomputing (ICS'12)*. ACM, New York, NY, 341–352.

Klaus Kofler, Ivan Grasso, Biagio Cosenza, and Thomas Fahringer. 2013. An automatic input-sensitive approach for heterogeneous task partitioning. In *Proceedings of the 27th ACM International Conference on Supercomputing (ICS'13)*.

Sai Kiran Korwar, Sathish Vadhiyar, and Ravi S. Nanjundiah. 2013. GPU-enabled efficient executions of radiation calculations in climate modeling. In *Proceedings of the 20th International Conference on High Performance Computing (HiPC'13)*. IEEE, 353–361.

Kishore Kothapalli, Dip Sankar Banerjee, P. J. Narayanan, Surinder Sood, Aman Kumar Bahl, Shashank Sharma, Shrenik Lad, Krishna Kumar Singh, Kiran Matam, Sivaramakrishna Bharadwaj, Rohit Nigam, Parikshit Sakurikar, Aditya Deshpande, Ishan Misra, Siddharth Choudhary, and Shubham Gupta. 2013. CPU and/or GPU: Revisiting the GPU Vs. CPU Myth. *arXiv preprint arXiv:1303.2171*.

Jens Lang and Gudula Rünger. 2013. Dynamic distribution of workload between CPU and GPU for a parallel conjugate gradient method in an adaptive FEM. *Procedia Computer Science* 18 (2013), 299–308.

Fabian Lecron, Sidi Ahmed Mahmoudi, Mohammed Benjelloun, Saïd Mahmoudi, and Pierre Manneback. 2011. Heterogeneous computing for vertebra detection and segmentation in X-ray images. *Journal of Biomedical Imaging* 2011, Article 5 (Jan. 2011).

Changmin Lee, Won W. Ro, and Jean-Luc Gaudiot. 2012. Cooperative heterogeneous computing for parallel processing on CPU/GPU hybrids. In *16th Workshop on Interaction between Compilers and Computer Architectures (INTERACT)*. IEEE, 33–40.

Janghaeng Lee, Mehrzad Samadi, Yongjun Park, and Scott Mahlke. 2013. Transparent CPU-GPU collaboration for data-parallel kernels on heterogeneous systems. In *Proceedings of the 22nd International Conference on Parallel Architectures and Compilation Techniques (PACT'13)*. 245–256.

Kenneth Lee, Heshan Lin, and Wu-chun Feng. 2013a. Performance characterization of data-intensive kernels on AMD fusion architectures. *Computer Science—Research and Development* 28, 2–3 (May 2013), 175–184.

Victor W. Lee, Changkyu Kim, Jatin Chhugani, Michael Deisher, Daehyun Kim, Anthony D. Nguyen, Nadathur Satish, Mikhail Smelyanskiy, Srinivas Chennupaty, Per Hammarlund, Ronak Singhal, and Pradeep Dubey. 2010. Debunking the 100X GPU vs. CPU myth: An evaluation of throughput computing on CPU and GPU. In *Proceedings of the 37th Annual International Symposium on Computer Architecture (ISCA'10)*. ACM, New York, NY, 451–460.

Hung-Fu Li, Tyng-Yeu Liang, and Jun-Yao Chiu. 2013. A compound OpenMP/MPI program development toolkit for hybrid CPU/GPU clusters. In *The Journal of Supercomputing* 66, 1 (2013), 381–405.

Jiajia Li, Xingjian Li, Guangming Tan, Mingyu Chen, and Ninghui Sun. 2012. An optimized large-scale hybrid DGEMM design for CPUs and ATI GPUs. In *Proceedings of the 26th ACM International Conference on Supercomputing (ICS'12)*. 377–386.

Linchuan Li, Xingjian Li, Guangming Tan, Mingyu Chen, and Peiheng Zhang. 2011. Experience of parallelizing cryo-EM 3D reconstruction on a CPU-GPU heterogeneous system. In *Proceedings of the 20th International Symposium on High Performance Distributed Computing*. ACM, New York, NY, 195–204.

Cong Liu, Jian Li, Wei Huang, Juan Rubio, Evan Speight, and Xiaozhu Lin. 2012. Power-efficient time-sensitive mapping in heterogeneous systems. In *Proceedings of the 21st International Conference on Parallel Architectures and Compilation Techniques (PACT'12)*. ACM, New York, NY, 23–32.

Ding Liu, Ruixuan Li, Xiwu Gu, Kunmei Wen, Heng He, and Guoqiang Gao. 2011. Fast snippet generation based on CPU-GPU hybrid system. In *Proceedings of the IEEE 17th International Conference on Parallel and Distributed Systems (ICPADS'11)*. 252–259.

Qiang Liu and Wayne Luk. 2012. Heterogeneous systems for energy efficient scientific computing. In *Reconfigurable Computing: Architectures, Tools and Applications*. Springer, 64–75.

Wenjie Liu, Zhihui Du, Yu Xiao, David A. Bader, and Chen Xu. 2011. A waterfall model to achieve energy efficient tasks mapping for large scale GPU clusters. In *Proceedings of the International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW'11)*. 82–92.

Yixun Liu, Andriy Fedorov, Ron Kikinis, and Nikos Chrisochoides. 2009. Real-time non-rigid registration of medical images on a cooperative parallel architecture. In *IEEE International Conference on Bioinformatics and Biomedicine*. 401–404.

Hatem Ltaief, Stanimire Tomov, Rajib Nath, Peng Du, and Jack Dongarra. 2011. A scalable high performant Cholesky factorization for multicore with GPU accelerators. In *High Performance Computing for Computational Science–VECPAR 2010*, Lecture Notes in Computer Science, Volume 6449. Springer, Berlin, 93–101.

Fengshun Lu, Junqiang Song, Xiaoqun Cao, and Xiaoqian Zhu. 2012a. CPU/GPU computing for long-wave radiation physics on large GPU clusters. *Computers & Geosciences* 41 (April 2012), 47–55.

Fengshun Lu, Junqiang Song, Fukang Yin, and Xiaoqian Zhu. 2012b. Performance evaluation of hybrid programming patterns for large CPU/GPU heterogeneous clusters. *Computer Physics Communications* 183, 6 (2012), 1172–1181.

Chi-Keung Luk, Sunpyo Hong, and Hyesoon Kim. 2009. Qilin: Exploiting parallelism on heterogeneous multiprocessors with adaptive mapping. In *Proceedings of the 42nd International Symposium on Microarchitecture (MICRO)*. ACM, New York, NY, 45–55.

Li Luo, Chao Yang, Yubo Zhao, and Xiao-Chuan Cai. 2011. A scalable hybrid algorithm based on domain decomposition and algebraic multigrid for solving partial differential equations on a cluster of CPU/GPUs. In *Proceedings of the 2nd International Workshop on GPUs and Scientific Applications*. 45–50.

Kai Ma, Xue Li, Wei Chen, Chi Zhang, and Xiaorui Wang. 2012. GreenGPU: A holistic approach to energy efficiency in GPU-CPU heterogeneous architectures. In *Proceedings of the 41st International Conference on Parallel Processing (ICPP)*. IEEE, 48–57.

Wenjing Ma, Sriram Krishnamoorthy, Oreste Villa, Karol Kowalski, and Gagan Agrawal. 2013. Optimizing tensor contraction expressions for hybrid CPU-GPU execution. *Cluster Computing* 16, 1 (March 2013), 131–155.

Artur Mariano, Ricardo Alves, Joao Barbosa, Luis Paulo Santos, and Alberto Proenca. 2012. A (ir) regularity-aware task scheduler for heterogeneous platforms. In *Proceedings of the International Conference on High Performance Computing*.

Kiran Kumar Matam, Siva Rama Krishna Bharadwaj, and Kishore Kothapalli. 2012. Sparse matrix matrix multiplication on hybrid CPU+ GPU platforms. In *Proceedings of the High Performance Computing Conference (HiPC'12)*.

Jeremy S. Meredith, Philip C. Roth, Kyle L. Spafford, and Jeffrey S. Vetter. 2011. Performance implications of nonuniform device topologies in scalable heterogeneous architectures. *IEEE Micro* 31, 5 (2011), 66–75.

Perhaad Mistry, Yash Ukidave, Dana Schaa, and David Kaeli. 2013a. A framework for profiling and performance monitoring of heterogeneous applications. *Programmability Issues for Heterogeneous Multicores (MULTIPROG-2013)*.

Perhaad Mistry, Yash Ukidave, Dana Schaa, and David Kaeli. 2013b. Valar: A benchmark suite to study the dynamic behavior of heterogeneous systems. In *Proceedings of the 6th Workshop on General Purpose Processor using Graphics Processing Units (GPGPU'13)*. ACM, New York, NY, 54–65.

Sparsh Mittal. 2012. A survey of architectural techniques for DRAM power management. *International Journal of High Performance Systems Architecture* 4, 2 (Dec. 2012), 110–119.

Sparsh Mittal. 2014a. A survey of techniques for managing and leveraging caches in GPUs. *Journal of Circuits, Systems, and Computers (JCSC)* 23, 8 (2014).

Sparsh Mittal. 2014b. A survey of architectural techniques for improving cache power efficiency. *Elsevier Sustainable Computing: Informatics and Systems* 4, 1 (2014), 33–43.

Sparsh Mittal. 2014c. A survey of techniques for improving energy efficiency in embedded computing systems. *International Journal of Computer Aided Engineering and Technology (IJCAET)* 46, 4, Article 47 (April 2014) .

Sparsh Mittal and Jeffrey S. Vetter. 2015. A survey of methods for analyzing and improving GPU energy efficiency. *ACM Computing Surveys* 47, 2, Article 19 (2015).

Timothy Prickett Morgan. 2014. Oracle Cranks up the Cores to 32 with Sparc M7 Chip. Retrieved from http://www.enterprisetech.com/2014/08/13/oracle-cranks-cores-32-sparc-m7-chip/.

Lluis-Miquel Munguia, David A. Bader, and Eduard Ayguade. 2012. Task-based parallel breadth-first search in heterogeneous environments. In *Proceedings of the 19th International Conference on High Performance Computing (HiPC'12)*. 1–10.

Jun-ichi Muramatsu, Takeshi Fukaya, Shao-Liang Zhang, Kinji Kimura, and Yusaku Yamamoto. 2011. Acceleration of Hessenberg reduction for nonsymmetric eigenvalue problems in a hybrid CPU-GPU computing environment. *International Journal of Networking and Computing* 1, 2 (2011).

Alin Murarașu, Josef Weidendorfer, and Arndt Bode. 2012. Workload balancing on heterogeneous systems: A case study of sparse grid interpolation. In *Euro-Par 2011: Parallel Processing Workshops*, Lecture Notes in Computer Science, Volume 7156. Springer, Berlin, 345–354.

Naohito Nakasato, Go Ogiya, Yohei Miki, Masao Mori, and Ken'ichi Nomoto. 2012. Astrophysical particle simulations on heterogeneous CPU-GPU systems. In *arXiv preprint arXiv:1206.1199*.

Andrew Nere, Sean Franey, Atif Hashmi, and Mikko Lipasti. 2012. Simulating cortical networks on heterogeneous multi-GPU systems. *J. Parallel and Distrib. Comput.* 43, 7 (July 2012), 953–971.

Rohit Nigam and P. J. Narayanan. 2012. Hybrid ray tracing and path tracing of Bezier surfaces using a mixed hierarchy. In *Proceedings of the 8th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'12)*. Article 35, 35:1–35:8.

NVIDIA. 2015. http://www.geforce.com/hardware/desktop-gpus.

Tetsuya Odajima, Taisuke Boku, Toshihiro Hanawa, Jinpil Lee, and Mitsuhisa Sato. 2012. GPU/CPU work sharing with parallel language XcalableMP-dev for parallelized accelerated computing. In *Proceedings of the 41st International Conference on Parallel Processing Workshops (ICPPW'12)*. 97–106.

Yasuhito Ogata, Toshio Endo, Naoya Maruyama, and Satoshi Matsuoka. 2008. An efficient, model-based CPU-GPU heterogeneous FFT library. In *Proceedings of the IEEE International Symposium on Parallel and Distributed Processing (IPDPS'08)*. 1–10.

Satoshi Ohshima, Kenji Kise, Takahiro Katagiri, and Toshitsugu Yuba. 2007. Parallel processing of matrix multiplication in a CPU and GPU heterogeneous environment. In *Proceedings of the 7th International Conference on High Performance Computing for Computational Science-VECPAR 2006*. Springer, 305–318.

OpenACC Standard. 2014. Homepage. Retrieved from http://www.openacc-standard.org/.

OpenMP 4.0. 2014. Homepage. Retrieved from http://openmp.org/wp/2013/07/openmp-40/.

Edson Luiz Padoin, Laércio Lima Pilla, Francieli Zanon Boito, Rodrigo Virote Kassick, Pedro Velho, and Philippe O. A. Navaux. 2013. Evaluating application performance and energy consumption on hybrid CPU+ GPU architecture. *Cluster Computing* 16, 3 (Sept. 2013), 511–525.

Sreepathi Pai, Ramaswamy Govindarajan, and Matthew Jacob Thazhuthaveetil. 2010. PLASMA: Portable programming for SIMD heterogeneous accelerators. In *Proceedings of the Workshop on Language, Compiler, and Architecture Support for GPGPU*.

Anthony Pajot, Loïc Barthe, Mathias Paulin, and Pierre Poulin. 2011. Combinatorial bidirectional path-tracing for efficient hybrid CPU/GPU rendering. In *Computer Graphics Forum* 30, 2 (April 2011), 315–324.

Prasanna Pandit and R. Govindarajan. 2014. Fluidic kernels: Cooperative execution of OpenCL programs on multiple heterogeneous devices. In *Proceedings of the International Symposium on Code Generation and Optimization (CGO)*. Article 273, 273:273–273:283.

Jairo Panetta, Thiago Teixeira, Paulo R. P. de Souza Filho, Carlos A. da Cunha Filho, David Sotelo, Fernando M. Roxo da Motta, Silvio Sinedino Pinheiro, Ivan Pedrosa Junior, Andre L. Romanelli Rosa, Luiz R. Monnerat, Leandro T. Carneiro, and Carlos H. B. de Albrecht. 2009. Accelerating Kirchhoff migration by CPU and GPU cooperation. In *Proceedings of the 21st International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD'09)*. 26–32.

Manolis Papadrakakis, George Stavroulakis, and Alexander Karatarakis. 2011. A new era in scientific computing: Domain decomposition methods in hybrid CPU–GPU architectures. *Computer Methods in Applied Mechanics and Engineering* 200, 13 (2011), 1490–1508.

Song Jun Park, James A. Ross, Dale R. Shires, David A. Richie, Brian J. Henz, and Lam H. Nguyen. 2011. Hybrid core acceleration of UWB SIRE radar signal processing. *IEEE Transactions on Parallel and Distributed Systems* 22, 1 (2011), 46–57.

Phitchaya Mangpo Phothilimthana, Jason Ansel, Jonathan Ragan-Kelley, and Saman Amarasinghe. 2013. Portable performance on heterogeneous architectures. In *Proceedings of the 18th International Conference on Architectural Support for Programming Languages and Operating Systems*. 431–444.

Jacques A. Pienaar, Srimat Chakradhar, and Anand Raghunathan. 2012. Automatic generation of software pipelines for heterogeneous parallel systems. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC'12)*. Article 24. 1–12.

Jacques A. Pienaar, Anand Raghunathan, and Srimat Chakradhar. 2011. MDR: Performance model driven runtime for heterogeneous parallel platforms. In *Proceedings of the ACM International Conference on Supercomputing*. ACM, New York, NY, 225–234.

Holger Pirk, Thibault Sellam, Stefan Manegold, and Martin Kersten. 2012. X-device query processing by bitwise distribution. In *8th International Workshop on Data Management on New Hardware*. ACM, New York, NY, 48–54.

Usman Pirzada. 2015. Nvidia Geforce GTX TITAN X Unveiled - GM200 'Big Daddy Maxwell', 12GB VRam and 8 Billion Transistors. Retrieved from wccftech.com/nvidia-gtx-titan-x-revealed-gdc-2015/.

Matthew Poremba, Sparsh Mittal, Dong Li, Jeffrey Vetter, and Yuan Xie. 2015. DESTINY: A tool for modeling emerging 3D NVM and eDRAM caches. In *DATE*. 1543–1546.

Ashwin Prasad, Jayvant Anantpur, and R. Govindarajan. 2011. Automatic compilation of MATLAB programs for synergistic execution on heterogeneous processors. In *ACM Sigplan Notices* 46, 6 (June 2011), 152–163.

Abtin Rahimian, Ilya Lashuk, Shravan Veerapaneni, Aparna Chandramowlishwaran, Dhairya Malhotra, Logan Moon, Rahul Sampath, Aashay Shringarpure, Jeffrey Vetter, Richard Vuduc, Denis Zorin, and George Biros. 2010. Petascale direct numerical simulation of blood flow on 200k cores and heterogeneous architectures. In *Proceedings of the 2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis (SC'10)*. 1–11.

Vignesh T. Ravi and Gagan Agrawal. 2011. A dynamic scheduling framework for emerging heterogeneous systems. In *Proceedings of the 18th International Conference on High Performance Computing (HiPC'11)*. 1–10.

Vignesh T. Ravi, Wenjing Ma, David Chiu, and Gagan Agrawal. 2010. Compiler and runtime support for enabling generalized reduction computations on heterogeneous parallel configurations. In *Proceedings of the 24th ACM International Conference on Supercomputing*. ACM, New York, NY, 137–146.

Vignesh T. Ravi, Wenjing Ma, David Chiu, and Gagan Agrawal. 2012. Compiler and runtime support for enabling reduction computations on heterogeneous systems. *Concurrency and Computation: Practice and Experience* 24, 5 (2012), 463–480.

Mahsan Rofouei, Thanos Stathopoulos, Sebi Ryffel, William Kaiser, and Majid Sarrafzadeh. 2008. Energy-aware high performance computing with graphic processing units. In *Proceedings of the 2008 Conference on Power Aware Computing and Systems*.

Bratin Saha, Xiaocheng Zhou, Hu Chen, Ying Gao, Shoumeng Yan, Mohan Rajagopalan, Jesse Fang, Peinan Zhang, Ronny Ronen, and Avi Mendelson. 2009. Programming model for a heterogeneous x86 platform. In *ACM Sigplan Notices* 44, 6 (2009), 431–440.

Thomas R. W. Scogland, Wu-chun Feng, Barry Rountree, and Bronis R. de Supinski. 2014. CoreTSAR: Adaptive worksharing for heterogeneous systems. In *Supercomputing*, Lecture Notes in Computer Science, Volume 8488. Springer International Publishing, 172–186.

Thomas R. W. Scogland, Barry Rountree, Wu-chun Feng, and Bronis R. de Supinski. 2012. Heterogeneous task scheduling for accelerated OpenMP. In *Proceedings of the IEEE 26th International Parallel & Distributed Processing Symposium (IPDPS'12)*. 144–155.

Jie Shen, Ana Lucia Varbanescu, Henk Sips, Michael Arntzen, and Dick G. Simons. 2013. Glinda: A framework for accelerating imbalanced applications on heterogeneous platforms. In *Proceedings of the ACM International Conference on Computing Frontiers*. ACM, New York, NY, Article 14.

Wenfeng Shen, Daming Wei, Weimin Xu, Xin Zhu, and Shizhong Yuan. 2010. Parallelized computation for computer simulation of electrocardiograms using personal computers with multi-core CPU and general-purpose GPU. *Computer Methods and Programs in Biomedicine* 100, 1 (2010), 87–96.

Takashi Shimokawabe, Takayuki Aoki, Tomohiro Takaki, Akinori Yamanaka, Akira Nukada, Toshio Endo, Naoya Maruyama, and Satoshi Matsuoka. 2011. Peta-scale phase-field simulation for dendritic solidification on the TSUBAME 2.0 supercomputer. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC'11)*. ACM, New York, NY. Article 3.

Koichi Shirahata, Hitoshi Sato, and Satoshi Matsuoka. 2010. Hybrid map task scheduling for GPU-based heterogeneous clusters. In *Proceedings of the IEEE 2nd International Conference on Cloud Computing Technology and Science (CloudCom'10)*. 733–740.

Sambit K. Shukla and Laxmi N. Bhuyan. 2013. A hybrid shared memory heterogeneous execution platform for PCIe-based GPGPUs. In *20th International Conference on High Performance Computing*. 343–352.

Jakob Siegel, Oreste Villa, Sriram Krishnamoorthy, Antonino Tumeo, and Xiaoming Li. 2010. Efficient sparse matrix-matrix multiplication on heterogeneous high performance systems. In *2010 IEEE International Conference on Cluster Computing Workshops*. 1–8.

Mark Silberstein and Naoya Maruyama. 2011. An exact algorithm for energy-efficient acceleration of task trees on CPU/GPU architectures. In *Proceedings of the 4th Annual International Conference on Systems and Storage (SYSTOR'11)*. ACM, New York, NY, Article 7.

Jaideeep Singh and Ipseeta Aruni. 2011. Accelerating smith-waterman on heterogeneous cpu-gpu systems. In *Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering (iCBBE'11)*. 1–4.

Hayden K. H. So, Junying Chen, Billy YS Yiu, and Alfred C. H. Yu. 2011. Medical ultrasound imaging: To GPU or not to GPU? *IEEE Micro* 31, 5 (2011), 54–65.

F. Sottile, C. Roedl, V. Slavnic, P. Jovanovic, D. Stankovic, P. Kestener, and Franck Houssen. 2013. GPU implementation of the DP code. *Partnership for Advanced Computing in Europe*.

Kyle Spafford, Jeremy Meredith, and Jeffrey Vetter. 2010. Maestro: Data orchestration and tuning for OpenCL devices. In *Euro-Par 2010-Parallel Processing*, Lecture Notes in Computer Science, Volume 6272. Springer, Berlin, 275–286.

Kyle L. Spafford, Jeremy S. Meredith, Seyong Lee, Dong Li, Philip C. Roth, and Jeffrey S. Vetter. 2012. The tradeoffs of fused memory hierarchies in heterogeneous computing architectures. In *Proceedings of the 9th Conference on Computing Frontiers*. 103–112.

Tomasz P. Stefanski. 2013. Implementation of FDTD-compatible Green's function on heterogeneous CPU-GPU parallel processing system. *Progress in Electromagnetics Research* 135 (2013), 297–316.

John E. Stone, David Gohara, and Guochun Shi. 2010. OpenCL: A parallel programming standard for heterogeneous computing systems. *Computing in Science & Engineering* 12, 3 (2010), 66–73.

Przemysław Stpiczynski. 2011. Solving linear recurrences on hybrid GPU accelerated manycore systems. In *Proceedings of the Federated Conference on Computer Science and Information Systems*. 465–470.

Przemysław Stpiczynski and Joanna Potiopa. 2010. Solving a kind of BVP for ODEs on heterogeneous CPU+ CUDA-enabled GPU systems. In *Proceedings of the International Multiconference on Computer Science and Information Technology (IMCSIT'10)*. IEEE, 349–353.

John A. Stratton, Christopher Rodrigues, I-Jui Sung, Nady Obeid, Li-Wen Chang, Nasser Anssari, Geng Daniel Liu, and W. M. W. Hwu. 2012. *Parboil: A Revised Benchmark Suite for Scientific and Commercial Throughput Computing*. Technical Report. Center for Reliable and High-Performance Computing.

Yu Su, Ding Ye, and Jingling Xue. 2013. Accelerating inclusion-based pointer analysis on heterogeneous CPU-GPU systems. In *Proceedings of the 20th International Conference on High Performance Computing*. 149–158.

Enqiang Sun, Dana Schaa, Richard Bagley, Norman Rubin, and David Kaeli. 2012. Enabling task-level scheduling on heterogeneous platforms. In *Proceedings of the 5th Annual Workshop on General Purpose Processing with Graphics Processing Units*. ACM, New York, NY, 84–93.

Hiroyuki Takizawa, Katsuto Sato, and Hiroaki Kobayashi. 2008. SPRAT: Runtime processor selection for energy-aware computing. In *Proceedings of the IEEE International Conference on Cluster Computing*. 386–393.

Yu Shyang Tan, Bu-Sung Lee, Bingsheng He, and Roy H. Campbell. 2012. A map-reduce based framework for heterogeneous processing element cluster environments. In *Proceedings of the 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*. 57–64.

George Teodoro, Tahsin M. Kurc, Tony Pan, Lee A. D. Cooper, Jun Kong, Patrick Widener, and Joel H. Saltz. 2012. Accelerating large scale image analyses on parallel, CPU-GPU equipped systems. In *Proceedings of the IEEE 26th International Parallel & Distributed Processing Symposium*. 1093–1104.

George Teodoro, Tony Pan, Tahsin M. Kurc, Jun Kong, Lee A. D. Cooper, and Joel H. Saltz. 2013. Efficient irregular wavefront propagation algorithms on hybrid CPU-GPU machines. *Parallel Comput.* 39, 4–5 (2013), 189–211.

George Teodoro, Rafael Sachetto, Olcay Sertel, Metin N. Gurcan, W. Meira, Umit Catalyurek, and Renato Ferreira. 2009. Coordinating the use of GPU and CPU for improving performance of compute intensive applications. In *International Conference on Cluster Computing and Workshops*. 1–10.

Pablo Toharia, Oscar D. Robles, Ricardo SuáRez, Jose Luis Bosque, and Luis Pastor. 2012. Shot boundary detection using Zernike moments in multi-GPU multi-CPU architectures. *Journal of Parallel and Distributed Computing* 72, 9 (Sept. 2012), 1127–1133.

Stanimire Tomov, Rajib Nath, and Jack Dongarra. 2010. Accelerating the reduction to upper Hessenberg, tridiagonal, and bidiagonal forms through hybrid GPU-based computing. *Parallel Computing* 36, 12 (2010), 645–654.

Top500. 2014. Top500 Supercomputers. www.top500.org.

Kuen Hung Tsoi and Wayne Luk. 2010. Axel: A heterogeneous cluster with FPGAs and GPUs. In *Proceedings of the 18th Annual ACM/SIGDA International Symposium on Field Programmable Gate Arrays*. ACM, New York, NY, 115–124.

Fernando Tsuda and Ricardo Nakamura. 2011. A technique for collision detection and 3D interaction based on parallel GPU and CPU processing. In *Proceedings of the 2011 Brazilian Symposium on Games and Digital Entertainment (SBGAMES'11)*. 36–42.

Abhishek Udupa, R. Govindarajan, and Matthew J. Thazhuthaveetil. 2009. Synergistic execution of stream programs on multicores with accelerators. *ACM Sigplan Notices* 44, 7 (2009), 99–108.

Yash Ukidave, Amir Kavyan Ziabari, Perhaad Mistry, Gunar Schirner, and David Kaeli. 2013. Quantifying the energy efficiency of FFT on heterogeneous platforms. In *Proceedings of the 2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS'13)*. 235–244.

Ronald Veldema, Thorsten Blass, and Michael Philippsen. 2011. Enabling multiple accelerator acceleration for Java/OpenMP. In *Proceedings of the 3rd USENIX Conference on Hot Topics in Parallelism (HotPar)*.

Sundaresan Venkatasubramanian and Richard W. Vuduc. 2009. Tuned and wildly asynchronous stencil kernels for hybrid CPU/GPU systems. In *Proceedings of the 23rd International Conference on Supercomputing (ICS'09)*. ACM, New York, NY, 244–255.

Uri Verner, Assaf Schuster, and Mark Silberstein. 2011. Processing data streams with hard real-time constraints on heterogeneous systems. In *Proceedings of the international conference on Supercomputing (ICS'11)*. ACM, New York, NY, 120–129.

Jeffrey S. Vetter and Sparsh Mittal. 2015. Opportunities for nonvolatile memory systems in extreme-scale high performance computing. *Computing in Science and Engineering* 17, 2 (2015), 73–82.

Christof Vömel, Stanimire Tomov, and Jack Dongarra. 2012. Divide and conquer on hybrid GPU-accelerated multicore systems. *SIAM Journal on Scientific Computing* 34, 2 (2012), C70–C82.

Richard Vuduc, Aparna Chandramowlishwaran, Jee Choi, Murat Guney, and Aashay Shringarpure. 2010. On the limits of GPU acceleration. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Parallelism*.

Guibin Wang and Wei Song. 2011. Communication-aware task partition and voltage scaling for energy minimization on heterogeneous parallel systems. In *Proceedings of the 12th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'11)*. 327–333.

Yueqing Wang, Yong Dou, Song Guo, Yuanwu Lei, and Dan Zou. 2014. CPU–GPU hybrid parallel strategy for cosmological simulations. *Concurrency and Computation: Practice and Experience* 26, 3 (March 2014), 748–765.

Yu Wang, Haixiao Du, Mingrui Xia, Ling Ren, Mo Xu, Teng Xie, Gaolang Gong, Ningyi Xu, Huazhong Yang, and Yong He. 2013a. A hybrid CPU-GPU accelerated framework for fast mapping of high-resolution human brain connectome. *PloS ONE* 8, 5 (2013), e62789.

Zhenning Wang, Long Zheng, Quan Chen, and Minyi Guo. 2013b. CAP: Co-scheduling based on asymptotic profiling in CPU+ GPU hybrid systems. In *International Workshop on Programming Models and Applications for Multicores and Manycores*. ACM, New York, NY, 107–114.

Mei Wen, Huayou Su, Wenjie Wei, Nan Wu, Xing Cai, and Chunyuan Zhang. 2012. Using 1000+ GPUs and 10000+ CPUs for sedimentary basin simulations. In *Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER)*. 27–35.

Dieter Wendel, Ronald Kalla, Joshua Friedrich, James Kahle, Jens Leenstra, Cedric Lichtenau, Balaram Sinharoy, William Starke, and Victor Zyuban. 2010. The POWER7 processor SoC. In *Proceedings of the International Conference on IC Design and Technology (ICICDT'10)*. 71–73.

Qiang Wu, Canqun Yang, Feng Wang, and Jingling Xue. 2012. A fast parallel implementation of molecular dynamics with the Morse potential on a heterogeneous petascale supercomputer. In *International Parallel and Distributed Processing Symposium Workshops & PhD Forum*. 140–149.

Tin-Yu Wu, Wei-Tsong Lee, Chien-Yu Duan, and Tain-Wen Suen. 2013. Enhancing cloud-based servers by GPU/CPU virtualization management. In *Advances in Intelligent Systems and Applications-Volume 2*. Springer, 185–194.

Shucai Xiao, Heshan Lin, and Wu-chun Feng. 2011. Accelerating protein sequence search in a heterogeneous computing system. In *Proceedings of the 2011 IEEE International Parallel & Distributed Processing Symposium*. 1212–1222.

Ming Xu, Feiguo Chen, Xinhua Liu, Wei Ge, and Jinghai Li. 2012. Discrete particle simulation of gas-solid two-phase flows with multi-scale CPU-GPU hybrid computation. *Chemical Engineering Journal* 207-208 (Oct. 2012), 746–757.

Canqun Yang, Feng Wang, Yunfei Du, Juan Chen, Jie Liu, Huizhan Yi, and Kai Lu. 2010. Adaptive optimization for petascale heterogeneous CPU/GPU computing. In *Proceedings of the IEEE International Conference on Cluster Computing (CLUSTER)*. 19–28.

Chao Yang, Wei Xue, Haohuan Fu, Lin Gan, Linfeng Li, Yangtong Xu, Yutong Lu, Jiachang Sun, Guangwen Yang, and Weimin Zheng. 2013. A peta-scalable CPU-GPU algorithm for global atmospheric simulations. In *Proceedings of the 18th ACM/SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP'13)*. ACM, New York, NY, 1–12.

Ping Yao, Hong An, Mu Xu, Gu Liu, Xiaoqiang Li, Yaobin Wang, and Wenting Han. 2010. CuHMMer: A load-balanced CPU-GPU cooperative bioinformatics application. In *Proceedings of the International Conference on High Performance Computing and Simulation (HPCS'10)*. IEEE, 24–30.

Marcelo Yuffe, Ernest Knoll, Moty Mehalel, Joseph Shor, and Tsvika Kurts. 2011. A fully integrated multi-CPU, GPU and memory controller 32nm processor. In *Proceedings of the IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC'11)*. 264–266.

Ziming Zhong, Vladimir Rychkov, and Alexey Lastovetsky. 2012. Data partitioning on heterogeneous multicore and multi-GPU systems using functional performance models of data-parallel applications. In *Proceedings of the International Conference on Cluster Computing*. 191–199.

V. Zyuban, S. A. Taylor, B. Christensen, A. R. Hall, C. J. Gonzalez, J. Friedrich, F. Clougherty, J. Tetzloff, and R. Rao. 2013. IBM POWER7+ design for higher frequency at fixed power. *IBM Journal of Research and Development* 57, 6 (2013), 1:1–1:18.

Received August 2014; revised March 2015; accepted May 2015