Exploiting Asymmetric Systems with Flexible System Software

Kallia Chronaki

Department of Computer Architecture

Universitat Politècnica de Catalunya

This dissertation is submitted for the degree of

Doctor of Philosophy

April 2018

BLANK

ABSTRACT

In 2013, U.S. data centres accounted for 2.2% of the country's total electricity consumption, a figure that is projected to increase rapidly over the next decade. A significant proportion of power consumed within a data centre is attributed to the servers, and a large percentage of that is wasted as workloads compete for shared resources. Many data centres host interactive workloads (e.g., web search or e-commerce), for which it is critical to meet user expectations and user experience, called Quality of Service (QoS). There is also a wish to run both interactive and batch workloads on the same infrastructure to increase cluster utilisation and reduce operational costs and total energy consumption. Although much work has focused on the impacts of shared resource contention, it still remains a major problem to maintain QoS for both interactive and batch workloads. The goal of this thesis is twofold. First, to investigate how, and to what extent, resource contention has an effect on throughput and power of batch workloads via modelling. Second, we introduce a scheduling approach to determine on-the-fly the best configuration to satisfy the QoS for latency-critical jobs on any architecture.

BLANK

Exploiting Asymmetric Systems with Flexible System Software

by
Kallia Chronaki

A Dissertation
Presented to the Department of Computer Architecture
at
Universitat Politècnica de Catalunya
in Candidacy for the Degree of
Doctor of Philosophy.

Thesis Advisors:

Rosa M. Badia, Prof.
Universitat Politècnica de Catalunya, Spain

Marc Casas, Prof.
Universitat Politècnica de Catalunya, Spain

Barcelona, July 2018

BLANK

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Energy efficiency has become the main challenge for future High Performance Computer (HPC) designs motivating prolific research to face the *Power Wall* [**?**]. The use of asymmetric multi-core architectures is an approach towards increasing energy efficiency [**? ? ? ?**] in the HPC domain. These architectures feature different types of processing cores that target different performance and power optimization points.

Asymmetric multi-cores have been successfully deployed in the mobile domain, where simple in-order cores (*little*) have been combined with aggressive out-of-order cores (*big*) to build these systems. In a first generation of asymmetric multi-cores, the system could switch from low power to high responding operation modes, activating or de-activating the cluster of big or little cores accordingly [**?**]. In a second generation of asymmetric multi-core processors, all the cores can run simultaneously to further improve the peak performance of these systems [**?**]. Many researchers are pushing towards building future parallel systems with asymmetric multicores [**? ? ? ? ?**] and even mobile chips [**?**]. However it is unclear if current parallel scientific applications will benefit from these platforms. Load balancing and scheduling are two of the main challenges in utilizing such heterogeneous platforms as the programmer has to consider the system's asymmetry from the very beginning to obtain an efficient parallelization.

To tackle these issues the programmer has to be aware of the platform's asymmetry and modify the code of the application accordingly so that it achieves load balance. In this scenario, the portability of applications is very limited and the existing scientific applications that are developed to target homogeneous multi-cores face significant performance degradation.

The first step of this thesis will be to characterize the maturity of the asymmetric plat-

forms when running scientific applications and how different scheduling options affect performance and energy efficiency. We evaluate for the first time the suitability of currently available mobile asymmetric multi-core platforms for general purpose computing. We compare three scheduling approaches each of them taking place on a different level of the software stack.

First, is the *application-based scheduling*, where the application is threaded statically by the programmer. We demonstrate that out-of-the box parallel applications are not portable to asymmetric multi-cores. The asymmetry of the system can lead to load imbalance and the user has to define generic and highly sophisticated load balancing mechanisms within the application that need to serve symmetric or asymmetric systems. Additionally, our characterization evaluates the impact of an *OS scheduler* (GTS) [**?**] that is aware of the platform's characteristics and dynamically migrates the threads to the appropriate type of core to increase performance. Finally, we suggest the use of a *task-based* programming model [**? ? ? ? ?**]; task-based programming models with dynamic scheduling allow the specification of inter-task dependencies that enable automatic scheduling and synchronization by the runtime system. We compare these options on the PARSECSs benchmark suite [**?**] as it offers a set of highly parallel scientific applications. This study will give us a complete picture of the state of the art high-level scheduling approaches in terms of performance and energy efficiency on our ARM big.LITTLE [**?**] 8-core asymmetric system. More details about this part of work can be found in Section **??**.

Having the proof that the *task-based* scheduling approach is the most sophisticated one, we plan to further improve it by making it aware of the system's asymmetry. This task, is carried out with the introduction of three asymmetry-aware scheduling policies within a task-based programming model. These scheduling policies have the knowledge of the types of cores of the system and, according to the application characteristics, they schedule tasks either on the fast or the slow cores of the system. More specifically, our task-based scheduling policies separate the tasks into groups of critical and non-critical tasks. Then the fast cores are responsible for the execution of the critical tasks while the slow cores execute the non-critical tasks. The main difference between these scheduling policies, apart from the implementation, is the critical task consideration and it is described in more detail in Section 3.2.

We further plan to modify the runtime system of the task-based programming model so that it takes advantage of the asymmetry of the platform. Specifically, we plan to explore the use of assistant cores for the runtime activity. Since in some cases the runtime introduces overheads, it is useful to devote one core for this activity (symbiotic core). For this, we plan to first evaluate all the available possibilities of the static assignment of the runtime thread to

one core of the system. This way we will be aware of the performance and energy outcome of executing the runtime activity on a big or on a little core. We then plan to implement mechanisms for dynamic runtime thread migration according to the current execution circumstances. This would involve the application's characteristics, such as number of tasks or whether the application is dependency intensive (has many dependencies between tasks).

Finally, we plan to incorporate our scheduling and thread migration techniques into a novel asymmetry-aware task-based runtime system. This runtime system will provide high performance and energy efficiency to the future asymmetric multi-cores and will combine the mechanisms of the use of symbiotic cores together with the scheduling policies to offer high adaptivity and portability to the current parallel applications. The runtime will use one assistant runtime thread that will migrate to either a fast or a slow core of the system according to the current circumstances and, at the same time, will perform scheduling according to the task criticality. We expect that this will significantly boost the portability, performance and energy efficiency of parallel scientific applications on the new asymmetric multi-core systems. We describe the above steps in more detail in Chapter **??**.

CHAPTER 2

BACKGROUND

2.1 THE ARM BIG.LITTLE ARCHITECTURE

The ARM big.LITTLE [**?** **?**] is a state-of-the-art AMC architecture that has been successfully deployed in the mobile market. The observation that mobile devices typically combine phases with low and high computational demands motivated this original design. ARM big.LITTLE combines simple in-order cores with aggressive out-of-order cores in the same System-on-Chip (SoC) to provide high performance and low power. *Big* and *little* cores support the same architecture so they can run the same binaries and therefore easily combined within the same system. Current cores implementing the ARMv7-A and ARMv8-A ISA support big.LITTLE configurations.

The little cores in a big.LITTLE system are designed targeting energy efficiency. Current implementations have relatively short pipelines with up to dual-issue in-order execution. L1 caches are split for instructions and data and can be dimensioned according to the target domain from 8 to 64 KB in size [**?**]. The big cores are designed for high performance. Current designs have deeper pipelines with up to seven-issue out-of-order execution, increased number of functional units and improved floating-point capabilities. L1 data cache is up to 64 KB and L1 instruction cache is up to 64 KB[**?** **?** **?**]. Little and big cores are typically integrated in a hierarchical manner. A set of cores form a cluster that may include a cache that is shared among cores in the cluster [**?**]. Then, multiple clusters can be interconnected through an on-chip network and share a last-level cache and connection to main memory and peripherals.

In this work, we use of one of the commercially available development boards featuring a big.LITTLE architecture: the Hardkernel Odroid-XU3 development board. As shown in

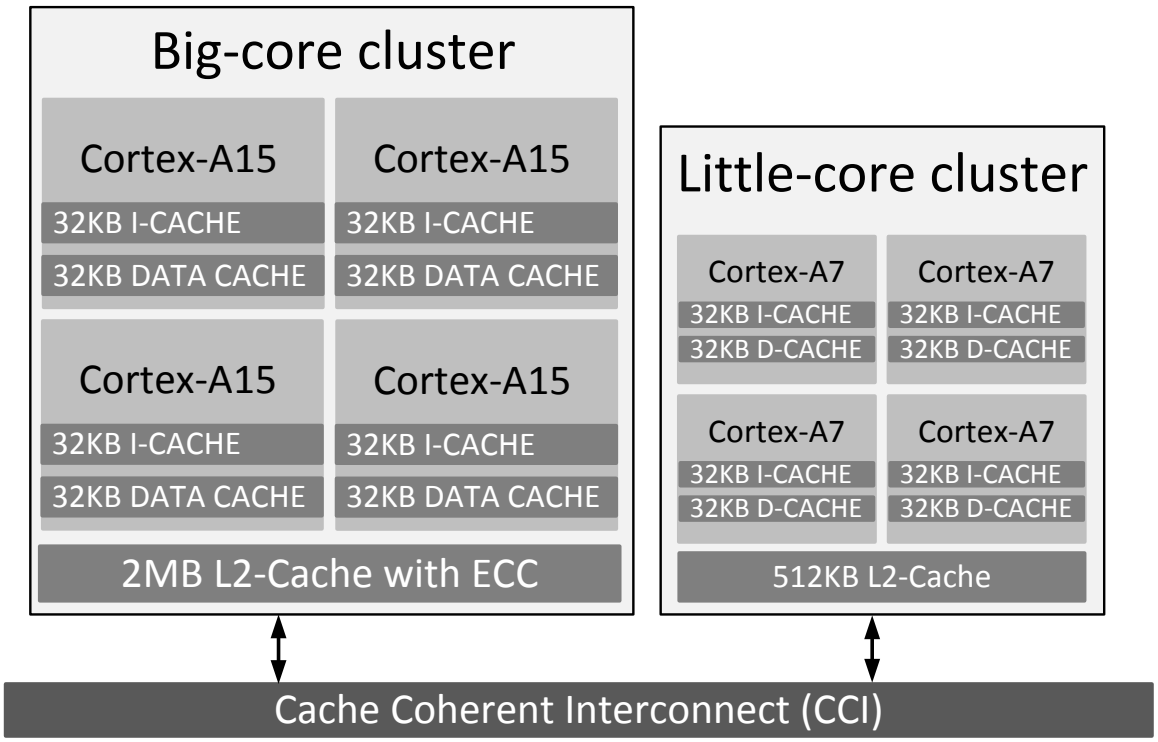


Fig. 2.1 Samsung Exynos 5422 processor with ARM big.LITTLE architecture.

Figure 2.1, the Odroid-XU3 includes an 8-core Samsung Exynos 5422 chip with four ARM Cortex-A15 cores and four Cortex-A7 cores. The four Cortex-A15 share a 2 MB 16-way 64-byte-cache-line L2 cache, while the Cortex-A7 cores share a 512 KB L2 cache. A single memory controller provides access to 2 GB of LPDDR3 RAM with dual 32-bit channels at 1866 MT/s. The reason we use this platform instead of the more up-to-date Juno platform [**?**] is that even if the latter features the more advanced Cortex A53 and Cortex A57 cores, it is limited to six cores instead of the 8 cores in Odroid-XU3.

The Cortex-A7 cores in this SoC support dual-issue of instructions and their pipeline length is between 8 and 10 stages. The L1 instruction cache is 32KB two-way set associative, with virtually indexed and physically tagged cache-lines that can hold up to 8 instructions. The core supports instruction prefetch by predicting the outcome of branches; the prefetch unit can fetch up to a maximum of four instructions per cycle. The L1 data cache is four-way set associative with physically-indexed and physically-tagged cache lines and uses a pseudo-random replacement policy [**?**]. Dynamic Voltage and Frequency Scaling (DVFS) techniques adjust the frequency of the little cores from 200MHz up to 1.4GHz.

The Cortex-A15 cores in this SoC support triple-issue of instructions and their pipeline length is between 15 and 24 stages [**?**]. The L1 instruction and data caches of the Cortex-A15 are both 32 KB and 2-way set-associative with 64 byte cache lines. The processor

supports speculative instruction execution by maintaining a 2-level global history-based dynamic predictor with a branch target buffer [**?**]. The instruction decode unit performs register renaming to remove the Write-After-Write and the Write-After-Read hazards, and promote instruction reordering [**?**]. The instruction dispatch unit analyzes instruction dependences before issuing them for execution. The integer execute unit includes 2 Arithmetic Logical Units with support for operand forwarding. DVFS techniques vary the frequency of the big cores from 200 MHz up to 2 GHz. For the rest of the paper, we refer to Cortex-A15 cores as *big* and to Cortex-A7 cores as *little*.

CHAPTER 3

STUDY OF ASYMMETRIC SYSTEMS

3.1 INTRODUCTION

The future of parallel computing is highly restricted by energy efficiency [**?**]. Energy efficiency has become the main challenge for future processor designs, motivating prolific research to face the *power wall*. Using heterogeneous processing elements is one of the approaches to increase energy efficiency [**? ?**]. Asymmetric multi-core (AMC) systems is an interesting case of heterogeneous systems to utilize for energy efficiency. These systems maintain different types of cores that support the same instruction-set architecture. The different core types are designed to target different (performance or power) optimization points [**? ? ?**].

AMCs have been mainly deployed for the mobile market. Mobile processors are also utilized in HPC platforms aiming to energy savings [**?**]. Asymmetric mobile SoCs combine low-power simple cores (*little*) with fast out-of-order cores (*big*) to achieve high performance while keeping power dissipation low. Another area where AMCs have been successful is the supercomputing market. The Sunway TaihuLight supercomputer topped the Top500 list in 2016 using AMCs. In this setup, big cores, that offer support for speculation to exploit Instruction-Level Parallelism (ILP), run the master tasks such as the OS and runtime system. Little cores are equipped with wide Single Instruction Multiple Data (SIMD) units and lean pipeline structures for energy efficient execution of compute-intensive code.

Like in other heterogeneous systems, load balancing and scheduling are fundamental challenges that must be addressed to effectively exploit all the resources in AMC platforms [**? ? ? ? ? ?**]. Mobile applications rely on multi-programmed workloads to balance the load in the system, while supercomputer applications rely on hand-tuned code

to extract maximum performance. However, these approaches are not always suitable for general-purpose parallel applications.

In this paper, we evaluate several execution models on an AMC using the PARSEC benchmark suite [**?**]. This suite includes parallel applications from multiple domains such as finance, computer vision, physics, image processing and video encoding. We quantify the performance loss of executing the applications *as-is* on all cores in the system. These applications were developed on homogeneous platforms and are bound to suffer from load imbalance on parallel regions that statically distribute the work evenly across cores without considering their performance disparity.

To overcome this matter, we consider two possible solutions at the OS and runtime levels to exploit AMCs effectively. The first solution delegates scheduling to the OS. We evaluate the built-in heterogeneity-aware OS scheduler currently used in existing mobile platforms that automatically assigns threads to different core types based on CPU utilization.

The second solution is to transfer the responsibility to the runtime system so it dynamically schedules work to different core types based on work progress and core availability. We evaluate the impact of using an inherently load-balanced execution model such that of task-based programming models. Recent examples [**? ? ? ? ? ? ? ? ?**] include clauses to specify inter-task dependencies and remove most barriers which are the major source of load imbalance on AMCs. Another approach of scheduling in the runtime system is to change the existing statically-scheduled work-sharing constructs for the applications implemented in OpenMP to use dynamic scheduling.

This paper provides the first to our knowledge comprehensive evaluation of representative parallel applications on a real AMC platform: the Odroid-XU3 development board with ARM big.LITTLE architecture. We analyze the effectiveness of the aforementioned scheduling solutions in terms of performance, power and energy. We show why parallel applications are not ready to run on AMCs and how OS and runtime schedulers can overcome these issues depending on the application characteristics. Further we point out in which aspects the built-in OS scheduler falls short to effectively utilize the AMC. Finally, we show how the runtime system approach overcomes these issues, and improves the OS and static threading approaches by 13% and 23% respectively.

The rest of this document is organized as follows: Section 5.2 describes the evaluated AMC processor, while Section 3.2 provides information on scheduling at the OS and runtime system levels. Section 5.4 describes the experimental framework. Section 5.5 shows the performance and energy results and associated insights.Finally, Section 5.6 discusses related work and Section 5.7 concludes this work.

3.2 SCHEDULING IN ASYMMETRIC MULTI-CORES

Scheduling a set of processes on an AMC system is more challenging than the traditional process scheduling on SMCs. An efficient OS scheduler has to take into account the different characteristics of the cores and act accordingly [**?**]. There have been three mainstream OS schedulers for ARM big.LITTLE systems: *cluster switching*, *in-kernel switch* and *global task scheduling*, described in the next sections. In the case of parallel applications, *dynamic scheduling at the runtime system level* can be exploited to balance the workload among the different cores and is described in section 3.2.3.

3.2 CLUSTER SWITCHING AND IN-KERNEL SWITCH

In the Cluster Switching (CS) approach [**?**], only one of the clusters is active at any given time: either the cluster with little cores or the cluster with big cores executes. Thus, the OS scheduler operates on a *de-facto* symmetric multi-core with only four cores, namely the cores of the current active cluster. The policy to change the operating cluster is based on CPU utilization. When idle, background processes are executed on the little cores. When CPU utilization surpasses a threshold, all processes (foreground and background) are migrated to the big cluster. When running on the big cluster, if CPU utilization decreases below a given lower threshold, the entire workload is moved to the little cluster.

In the In-Kernel Switch (IKS) approach [**?**], each little core is paired with a big core and it is seen as a single core. On idle, background processes are run on little cores. When the CPU utilization on a given little core surpasses a threshold, the execution on that core is migrated to the big core. When the CPU utilization decreases on that big core below a given threshold, the execution migrates to the associated little core. Thus, at the same time, little and big cores may co-execute, but only one of each pair is active at a given point in time, effectively exploiting just half of the cores concurrently. For both CS and IKS, an enhanced `cpufreq` driver manages the switching within each core pair.

3.2 GLOBAL TASK SCHEDULING

The Global Task Scheduling (GTS) [**?**] allows running applications on all cores in the asymmetric multi-core. In GTS, all cores are available and visible to the OS scheduler, and this scheduler is aware of the characteristics of the core types. Each process is assigned to a core type depending on its CPU utilization: high CPU utilization processes are scheduled to big cores and low CPU utilization processes to little cores. GTS also migrates processes between big and little cores when their CPU utilization changes. As a result, cores are active

depending on the characteristics of the workload.

The key benefit of GTS is that it can use all the cores simultaneously, providing higher peak performance and more flexibility to manage the workload. In GTS tasks are directly migrated to cores without needing the intervention of the `cpufreq` daemon, reducing response time and minimizing the overhead of context switches. As a consequence, Samsung reported 20% improvement in performance over CS for mobile benchmarks [**?**]. Also, GTS supports clusters with different number of cores (e.g. with 2 big cores and 4 little cores), while IKS requires to have the same number of cores per cluster.

3.2 Dynamic Scheduling in the Runtime

Current programming models for shared memory systems such as OpenMP rely on a runtime system to manage the execution of the parallel application. In this work, we make use of two types of programming models: loop- and task-based. Loop-based scheduling distributes the iterations of a loop among the threads available in the system, following a traditional *fork-join* model. OpenMP supports loop-based scheduling through its *parallel for* directives. This clause implies a barrier synchronization at the end of the loop[1], and supports either static or dynamic loop scheduling.

With static loop scheduling, the iterations of a loop are divided to as many chunks as the number of cores. Then, every core executes the assigned chunk, leading to a low-overhead static scheduling. In addition, OpenMP supports dynamic loop scheduling. It generates more chunks than cores, and assigns them to the available cores at runtime. This is more suitable to asymmetric multi-core systems where the cores are not similar and a static iteration assignment would cause load imbalance.

Recent advances in programming models recover the use of task-based programming models to simplify parallel programming of multi-cores [**? ? ? ? ?**]. In these models the programmer splits the code in sequential pieces of work (tasks) and specifies the data dependencies among them. With this information the runtime system schedules tasks and manages synchronization. These models ease programmability [**? ? ? ? ? ?**], and also increase performance by avoiding global synchronization points.

To evaluate this approach we make use of OpenMP tasking support [**?**]. OpenMP allows expressing tasks and data dependences between them using equivalent code annotations. It conceives the parallel execution as a *task dependence graph* (TDG), where nodes are sequential pieces of code (tasks) and the edges are control or data dependences between them. The runtime system builds this TDG at execution time and dynamically schedules tasks to the available cores. Tasks become ready as soon as their input dependencies are satisfied.

[1] unless specified otherwise with the `nowait` clause

Table 3.1 Benchmarks used from the PARSEC benchmark suite and their measured performance ratio between big and little cores

Benchmark	Description	Input	Parallelization	Perf ratio
blackscholes	Calculates the prices of a portfolio analytically with the Black-Scholes partial differential equation.	10,000,000 options	data-parallel	2.18
bodytrack	Computer vision application which tracks a 3D pose of a marker-less human body with multiple cameras through an image sequence.	4 cameras, 261 frames, 4,000 particles, 5 annealing layers	pipeline	4.16
canneal	Simulated cache-aware annealing to optimize routing cost of a chip design.	2.5 million elements, 6,000 steps	unstructured	1.73
dedup	Compresses a data stream with a combination of global compression and local compression in order to achieve high compression ratios.	351 MB data	pipeline	2.67
facesim	Takes a model of a human face and a time sequence of muscle activation and computes a visually realistic animation of the modeled face.	100 frames, 372,126 tetrahedra	data-parallel	3.40
ferret	Content-based similarity search of feature-rich data (audio, images, video, etc.)	3,500 queries, 59,695 images database, find top 50 images	pipeline	3.59
fluidanimate	Extended Smoothed Particle Hydrodynamics method to simulate an incompressible fluid for interactive animations.	500 frames, 500,000 particles	data-parallel	3.32
streamcluster	Solves the online clustering problem.	200K points per block, 5 block	data-parallel	3.48
swaptions	Intel RMS workload; uses the Heath-Jarrow-Morton framework to price a portfolio of swaptions.	128 swaptions, 1 million simulations	data-parallel	2.78

The scheduling of the ready tasks is done in a first-come-first-served manner, using a FIFO scheduler. Even though this scheduler is not aware of the task computational requirements or the core type and its performance and power characteristics, it can balance the load as long as there are ready tasks available thanks to the lack of global synchronization.

3.3 Experimental Methodology

3.3 Metrics

All the experiments in this paper are performed on the Hardkernel Odroid XU3 described in Section 5.2. To avoid machine overheating, we make use of the `cpufreq` driver to set big cores at 1.6GHz and little cores at 800MHz.

We evaluate seven configurations with different numbers of *little* (L) and *big* (B) cores, denoted `L+B`. For each configuration and benchmark, we report the average performance of five executions in the application parallel region. Then, we report the application speedup over its execution time on one little core. Equation 3.1 shows the formula to compute this

speedup.

$$\text{Speedup(L, B, method)} = \frac{\text{Exec. time(1, 0, method)}}{\text{Exec. time(L, B, method)}} \tag{3.1}$$

In this platform, there are four separated current sensors to measure, in real time, the power consumption of the A15 cluster, the A7 cluster, the GPU and DRAM. To gather power and energy measurements, a background daemon reads the machine power sensors periodically during the application execution with negligible overhead. Sensors are read at their refresh rate, every 270ms, and the values of A7 and A15 clusters' sensors are collected. With the help of timestamps, we correlate the power measurements with the application parallel region in a *post-mortem* process[2]. The reported power consumption is the average power tracked during five executions of each configuration, considering the application parallel region only. We then report average power in Watts along the execution.

Finally, in terms of energy and Energy Delay Product (EDP), we report the total energy and EDP of the benchmarks region of interest normalized to the run on four little cores with static threading. Equations 3.2 and 3.3 show the formulas for these calculations.

$$\text{Normalized Energy(L, B, method)} = \frac{\text{Energy(L, B, method)}}{\text{Energy(4, 0, static-threading)}} \tag{3.2}$$

$$\text{Normalized EDP(L, B, method)} = \frac{\text{EDP(L, B, method)}}{\text{EDP(4, 0, static-threading)}} \tag{3.3}$$

3.3 APPLICATIONS

With the prevalence of many-core processors and the increasing relevance of application domains that do not belong to the traditional HPC field, comes the need for programs representative of current and future parallel workloads. The PARSEC benchmark suite [**?** **?**] features state-of-the-art, computationally intensive algorithms and very diverse workloads from different areas of computing. In our experiments, we make use of the original PARSEC codes together with a task-based implementation of nine benchmarks of the suite [**?**].

Table 3.1 describes the benchmarks included in the study along with their respective inputs, parallelization strategy and performance ratio between big and little cores per application. We are using native inputs, which are real input sets for native execution, except for `dedup`, as the entire input file of 672 MB and the intermediate data structures do not fit in the memory system of our platform. Instead, we reduce the size of the input file to 351 MB.

[2]The parallel region duration is several orders of magnitude longer than the reading frequency of power sensors

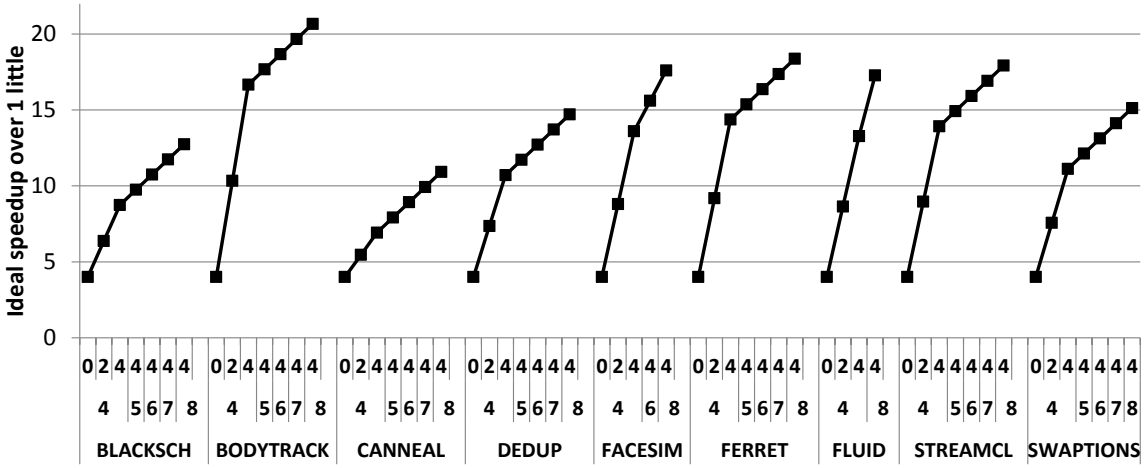


Fig. 3.1 Ideal speedup over 1 little core according to Equation 3.4. Numbers at the bottom of x axis show the total number of cores, numbers above it show the number of big cores

The original codes make use of the `pthreads` parallelization model for all the selected benchmarks. The taskified applications follow the same parallelization strategy implemented with OpenMP 4.0 task annotations. The task-based implementation is done following two basic ideas: i) remove barriers where possible, by adding explicit data-dependencies; and ii) remove application-specific load balancing mechanisms, such as application-specific pools of threads implemented in `pthreads` and delegate this responsibility to the runtime.

When running on the big.LITTLE processor, each benchmark exhibits different performance ratios between big and little cores. These ratios tell us how many times faster a big core is compared to a little core. We measure the performance ratio of each application by executing it first on one big core and then on one little core, which corresponds to Speedup(0, 1, task-based) in Equation 3.1. Table 3.1 also includes the observed performance ratio for each application. Bodytrack is the application that benefits the most from running on the big core with a performance ratio of 4.16×. The out-of-order execution of the big core together with an increased number of in-flight instructions significantly improves the performance of this application. In contrast, canneal is the benchmark with the lowest performance ratio, 1.73×, as this is a memory-intensive benchmark that does not benefit as much from the extra computation power of the big core. In general, performance ratios are above 2.5× for seven out of nine benchmarks, reaching 3.03× on average.

Taking into account these performance ratios, we can estimate the ideal speedup of the platform for each workload assuming a perfect parallelization strategy. Equation 3.4 shows the equation for the ideal speedup over 1 little core computation according to the number of

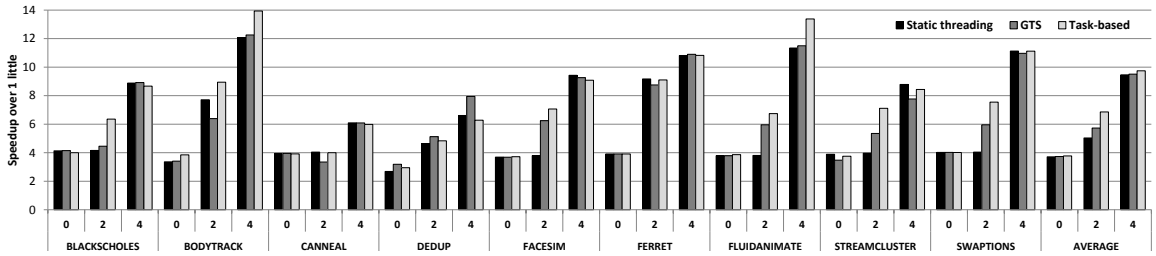


Fig. 3.2 Execution time speedup over 1 little core for systems that consist of 4 cores in total with 0, 2 and 4 big cores. Different schedulers at the application (*static threading*), OS (*GTS*) and runtime (*task-based*) levels are considered.

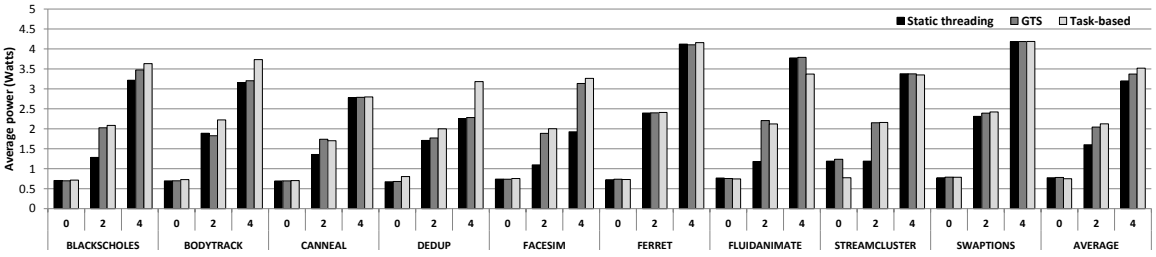


Fig. 3.3 Average power measurements on a 4-core system with 0, 2, and 4 big cores.

big (B) and little (L) cores.

$$\text{Ideal speedup(workload, B, L)} = \text{B} \times \text{Perf\_ratio(workload)} + \text{L} \tag{3.4}$$

Figure 3.1 shows the ideal speedup of the system for each application for the varying numbers of cores. This speedup assumes that the applications are fully parallel with no barriers or other synchronization points. Thus, these speedups are an upper bound of the achievable application performance.

3.4 EVALUATION

We measure execution time, power, energy and EDP of nine applications from the PARSEC benchmark suite [**?**]. We compare these metrics for three different scheduling approaches:

- *Static threading*: scheduling decisions are made at the application level. The OS is not allowed to migrate threads between the clusters of big and little cores.
- *GTS*[3]: dynamic coarse-grained OS scheduling using the GTS scheduler integrated in the Linux kernel [**? ?**] using the default PARSEC benchmarks.

[3]We choose to evaluate GTS instead of CS and IKS because it is the most advanced scheduling approach supported in the Linux kernel.

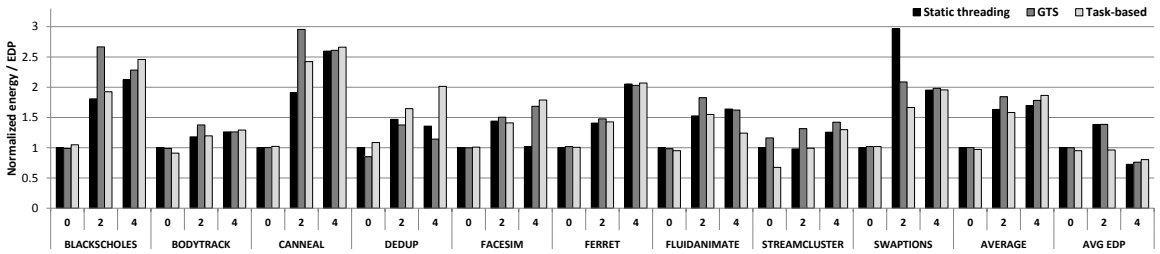


Fig. 3.4 Normalized energy consumption and average EDP on a 4-core system with 0, 2, and 4 big cores. Static threading on 4 little cores is the baseline in both cases.

- *Task-based*: dynamic fine-grained scheduling at the runtime level with the task-based implementations of the benchmarks provided in PARSECSs [**?**].

3.4 Exploiting Parallelism in AMCs

This section examines the opportunities and challenges that current AMCs offer to emerging parallel applications. With this objective, we first evaluate a system with a constant number of four cores, changing the level of asymmetry to evaluate the characteristics of each configuration. In these experiments, all applications run with the original parallelization strategy that relies on the user to balance the application (*Static threading*). We also evaluate the OS-based dynamic scheduling (*GTS*) and the task-based runtime dynamic scheduling (*Task-based*) for the same applications. The system configurations evaluated in this section are: i) Four little cores (`0+4`); ii) Two big and two little cores (`2+2`); and iii) Four big cores (`4+0`)

For these configurations, Figure 3.2 shows the speedup of the PARSEC benchmarks with respect to running on a single little core. Figure 3.3 reports the average power dissipated on the evaluated platform. Finally, Figure 3.4 shows the total energy consumed per application for the same configurations. Energy results are normalized to the energy measured with four little cores (higher values imply higher energy consumptions). Average EDP results are also included in this figure.

Focusing on the average performance results, we notice that all approaches perform similarly for the homogeneous configurations. Specifically, applications obtain the best performance on the configuration *4+0*, with an average speedup of 9.5× over one little core. When using four little cores, an average speedup of 3.8× is reached for all approaches. This shows that all the approaches are effective for this core count. In the configuration *2+2*, *Static threading* slightly improves performance (5.0× speedup), while *GTS* and *Task-based* reach significantly higher speedups: 5.9× and 6.8×, respectively.

Contrarily, in terms of power and energy, the most efficient configuration is running

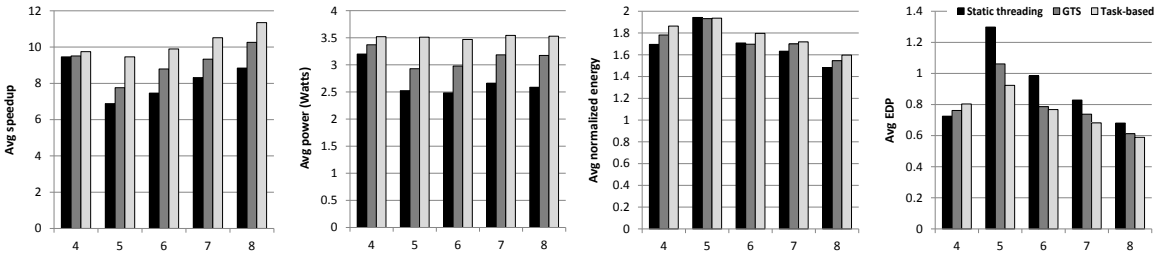


Fig. 3.5 Average results when running on 4 to 8 cores with 4 of them big. Speedup is over 1 little core. Static threading on 4 little cores is the baseline of energy consumption and EDP

with four little cores, as the performance ratio between the different cores is inversely proportional to the power ratio [**?**]. On average, all the approaches reach a power dissipation of 0.75W for the `0+4` configuration, while *Task-based* reaches 3.5W for the `4+0` configuration which is the one with the highest average power dissipation. In configuration `2+2`, average energy values for *Static threading* and *Task-based* are nearly the same, as the increase in power from 1.6W to 2.1W is compensated by a significant improvement in performance of 30%.

Finally, in terms of EDP using the four big cores is the optimal, as the performance improvements compensate the increase in total energy. In configuration `2+2`, *Task-based* achieves the same EDP results as in `0+4`, but with 81% better performance. For the asymmetric configuration, *Task-based* achieves the best performance-energy combination since its dynamic scheduling is effectively utilizing the little cores.

Next, we focus on the obtained results per benchmark. For applications with an extensive use of barriers (blackscholes, facesim, fluidanimate, streamcluster and swaptions) or with a memory intensive pattern (canneal), the extra computational power offered by the big cores in configuration `2+2` is not exploited. As a result with *Static threading* performance is only slightly improved by 1% on average when moving from `0+4` to the `2+2` configuration. This slight improvement comes at the cost of much more power and energy consumption (79% and 77% respectively). These results are explained three-fold: i) load is distributed homogeneously among threads in some applications; ii) extensive usage of barriers force big cores to wait until little cores reach the barrier; and iii) high miss rates in the last-level cache cause frequent pipeline stalls and prevent to fully exploit the computational power of big cores. To alleviate these problems, the programmer should develop more advanced parallelization strategies that could benefit from AMCs, as performed in the remaining applications, or rely on dynamic scheduling at OS or runtime levels.

The three remaining applications are parallelized using a pipeline model (bodytrack, dedup, and ferret) with queues for the data-exchange between pipeline stages and application-specific load balancing mechanisms designed by the programmer. As a result, *Static schedul-*

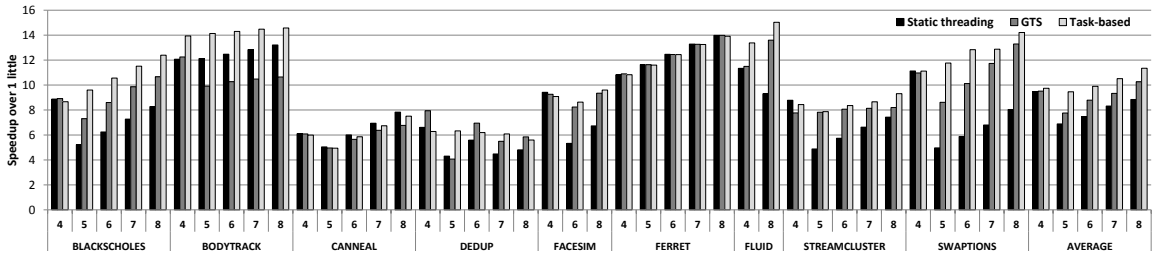


Fig. 3.6 Speedup over 1 little core when running on 4 to 8 cores and 4 of them are big

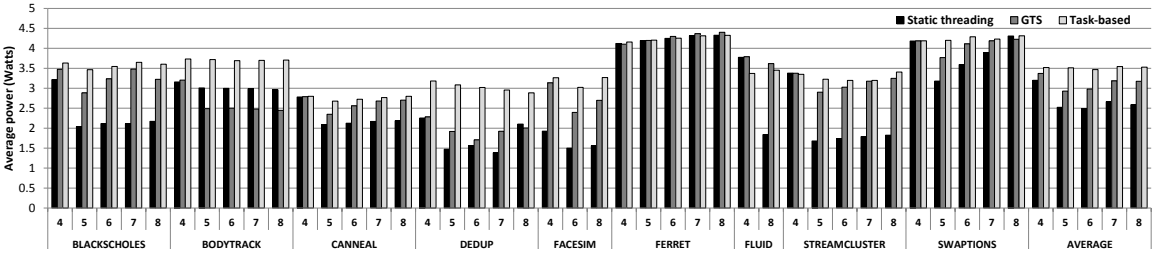


Fig. 3.7 Average power when running on 4 to 8 cores and 4 of them are big

ing with these applications benefits from the extra computational power of the big cores in the configuration 2+2. These mechanisms are not needed in the *Task-based* code; in this approach the code of the application is simplified and the runtime automatically allows the overlapping of the different pipeline stages. Thus, on the asymmetric configuration, *Task-based* further improves the obtained performance, reaching a 13% average improvement over *GTS*. Clearly, these applications benefit in performance by the increased number of big cores, while power and energy are increasing since the big cores are effectively utilized.

Generally, relying on the programmer to statically schedule asymmetric configurations does not report good results, as it is very hard to predict the system's behaviour at application-level. Only applications that implement advanced features with user-level schedulers and load balancing techniques, can benefit from asymmetry, at the cost of programmability effort. Relying on the OS scheduler is a suitable alternative without code modifications, but relying on the runtime to dynamically schedule tasks on the asymmetric processor achieves much better performance, power and energy results.

3.4 Adding Little Cores to an SMC

In the following experiments, we explore if an application running on a symmetric multi-core (SMC) with big cores can benefit from adding small cores that help in its execution. Having more computational resources increases the ideal speedup a parallel application can reach, but it also introduces challenges at application, runtime and OS level. Thus, we examine how many small cores have to be added to the system to compensate the cons of

having to deal with AMCs.

To evaluate this scenario, we explore configurations `4+0`, `4+1`, `4+2`, `4+3` and `4+4`. In these experiments, the number of big cores remains constant (four), while the number of little cores increases from 0 to 4. First we focus on the average results of speedup, power, energy and EDP, shown in Figure 3.5.

The speedup chart of Figure 3.5 shows that *Static threading* does not benefit from adding little cores to the system. In fact, this approach brings an average 6% slowdown when adding four little cores for execution (`4+4`). This is a result of the static thread scheduling; because the same amount of work is assigned to each core, when the big cores finish the execution of their part, they become idle and under-utilized. GTS achieves a limited speedup of 8% with the addition of four little cores to the `4+0` configuration. The addition of a single little core brings a 22% slowdown (from `4+0` to `4+1`) and requires three additional little cores to reach the performance of the symmetric configuration (`4+3`). Finally, the *Task-based* approach always benefits from the extra computational power as the runtime automatically deals with load imbalance. Performance improvements keep growing with the additional little cores, reaching an average improvement of 15% over the symmetric configuration when 4 extra cores are added.

The power chart of Figure 3.5 shows oppositional benefits among the three approaches. We can see that *Static threading* and *GTS* benefit from asymmetry, effectively reducing average power consumption. *Static threading* reduces power consumption when moving from the `4+0` to the `4+4` system by 23% while *GTS* does so by 6.2%. On the other hand, the *task-based* approach keeps the big cores busy for most of the time so it maintains the average power nearly constant.

The reduction in power, results to reduced average energy in the case of *Static threading* in configuration `4+4`, as shown on the energy chart of Figure 3.5. As discussed in Section 3.4.1, little cores are more energy efficient than big cores, at the cost of reduced performance. In all the approaches, at least two extra little cores are needed to reduce energy. In configuration `4+4`, energy is reduced by 14% for *Static threading*, 15% for *GTS*, and 16% for *Task-based*. Consequently, we can state that asymmetry reduces overall energy consumption.

To see the impact on both performance and energy efficiency we plot the average EDP on the rightmost chart of Figure 3.5. In this chart the lower values are the better. The *task-based* approach is the one that has the best performance-energy combination for the asymmetric configurations since it maintains the lowest EDP for all cases. *Static threading* manages to reduce the average EDP by 6% while *GTS* and *task based* approaches do so by 24% and 36% respectively.

Figure 3.6 shows a more detailed exploration of the performance results. As Table 3.1 shows, the applications with barrier synchronization are blackscholes, facesim, fluidanimate, streamcluster and swaptions. For these applications the most efficient system configuration with the *Static threading* approach is the `4+0`. Little cores increase execution time due to load imbalance effects. Since the big cores reach barriers earlier, power is reduced for these applications, as shown in Figure 3.7. Energy reduction is less significant with a few extra little cores as the performance degradation is higher, but as the number of little cores increases, energy is reduced.

Applications with more advanced load balancing techniques like pipelined parallelism (bodytrack, dedup and ferret), benefit of the asymmetric hardware and balance the load among all the cores. As a result, performance improves as we increase the number of little cores. In the case of bodytrack, *GTS* reduces performance by 15% when adding four little cores. We attribute this to the cost of the thread migration from one core to the other in contrast to the *Static threading* approach that does not add such overheads. In the case of dedup, results show more variability. This benchmark is very I/O intensive and, depending on the type of core that executes these I/O operations, performance drastically changes. In order to deal with this problem, a smarter dynamic scheduling mechanism would be required. Finally, canneal does not scale according to its ideal speedup reported on Figure 3.1 as it has a memory intensive pattern that limits performance.

Figure 3.7 shows the average power. The barrier-synchronized applications (blackscholes, facesim, fluidanimate, streamcluster and swaptions) reduce power because of their imbalance; since big cores have long idle times with the *Static threading* approach, they do not spend the same power as *GTS* and *Task-based*. For pipeline-parallel applications, both bodytrack and ferret maintain nearly the same power levels among the configurations for each scheduling approach. Dedup is an exception, as the results highly depend on the core that executes the aforementioned I/O operations. Yet, the effect of the lower power for *Static threading* is observed in all the benchmarks and is because the big cores are under-utilized.

This section proves that adding little cores to an SMC with big cores presents significant challenges for the application, OS and runtime developers. Little cores increase load imbalance and can degrade performance as a result. Relying on the programmer to deal with this asymmetry is complex, but a dynamic OS scheduler such as *GTS* helps in mitigating these problems, providing an average performance increase of 10%. However, the optimal performance results are obtained with the *Task-based* approach, as they improve static threading by 23% on average. In terms of power and energy, the AMC provides significant benefits, although the SMC with little cores remains the most energy-efficient configuration. The answer to the question of which system configuration provides the best power-performance

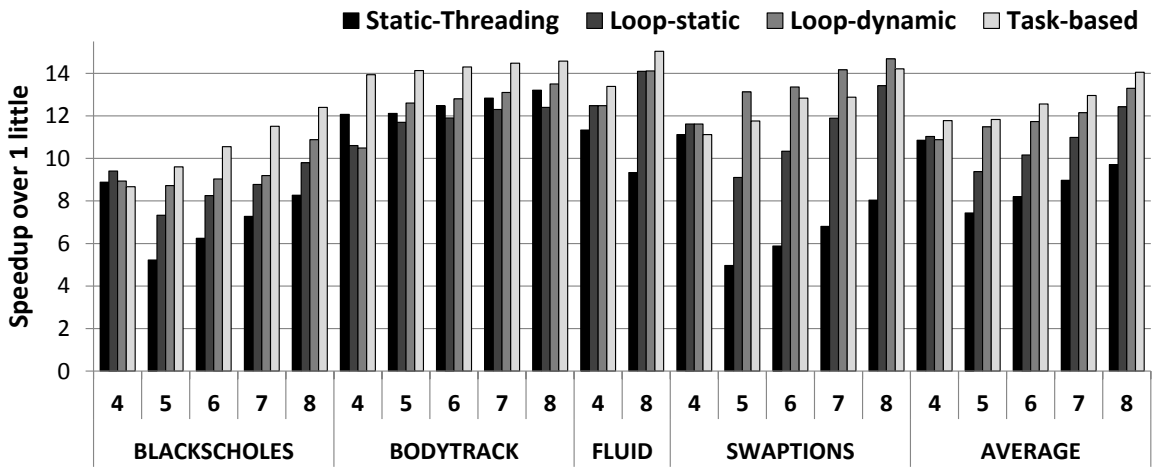


Fig. 3.8 Speedup over 1 little core when running on 4 to 8 cores and 4 of them are big. Four different programming models are considered: Static threading using `pthreads`, parallel loops with static scheduling (loop static), parallel loops with dynamic scheduling (loop dynamic), and a task-based solution with dynamic scheduling (task-based).

balance, can be found on the average EDP chart of Figures 3.4 and 3.5, and is the use of the entire 8-core system with the *Task based* approach.

3.4 Programming Models for AMCs

As we saw in the previous section, current implementations of parallel applications are not ready to fully take advantage of an AMC system. Applications that are statically threaded using the low-level `pthreads` library usually suffer from load imbalance since their implementations assume that the work has to be equally distributed among the available cores. Implementing advanced load balancing schemes, such as work pools, in `pthreads` requires a significant development effort.

As an alternative, many parallel applications are implemented using loop-based scheduling with the OpenMP *parallel for* directives. In this case, the runtime library is in charge of scheduling work to the available threads in the system, either statically or dynamically, as described in Section 3.2.3.

We compare these solutions to the task-based approach evaluated in the previous sections. Figure 3.8 shows the results obtained from running blackscholes, bodytrack, fluidanimate and swaptions on all the scheduling models: static threading, static loop scheduling, dynamic loop scheduling and task-based scheduling. We chose these applications as they are the only ones implemented using the OpenMP loop directives.

Looking at the average results in Figure 3.8, we can observe that the task-based solution achieves the best results when the system is asymmetric. Task-based improves the static

threading by up to 59% on 5 cores, while dynamic loop scheduling improves by up to 54%. The OpenMP version with static scheduling reaches an average 26% improvement over the static-threading approach with pthreads.

Taking a closer look to the results we observe that for bodytrack, an application with sophisticated parallelization techniques, static-threading achieves better results than loop-static. This is because the static-threading implementation contains specific parallelization techniques that cannot be completely expressed using the loop-static method. The loop-dynamic method improves performance for bodytrack by up to 4% due to the runtime decisions of the iteration execution, but the optimal solution is offered by the task-based approach that achieves up to 16% improvement over static-threading, due to the flexibility in expressing irregular parallelization strategies.

Blackscholes, fluidanimate and swaptions, consist of independent tasks and are a good fit for loop parallelism. The first observation is that both applications benefit from the loop-static approach on an SMC with 4 big cores. Moreover, the task-based approach is still the optimal for blackscholes and fluidanimate, reaching up to 83% improvement over static threading for 5 cores, while for swaptions loop-dynamic is the most appropriate, improving the baseline by up to 2.6×. The difference in the benefits of these applications relies on the task granularity; blackscholes consists of 6400 tasks that are about a hundred times smaller than each one of the 128 tasks of swaptions. This shows that loop-dynamic is more efficient on coarse-grained applications. Finally, fluidanimate, that is also a fine-grained application that consists of 128 500 tasks, also benefits from the task-based approach. For this benchmark, static and dynamic loop scheduling achieve similar performance; this is due to the limited parallelism per parallel region, as the loop-based implementation consists of multiple barriers between small parallel regions, fact that diminishes the effect of dynamic vs static scheduling.

3.5 Related Work

There has been a lot of studies on AMC systems. Some works focus on the system design, while other works explore the challenges that appear in efficiently utilizing such a heterogeneous system. Kumar et al [**?**] present the idea of an AMC system and proposed a feedback-based way to dynamically migrate processes among the different cores. To determine the core that most effectively executed a workload, Kumar et al [**?**] proposed the use of sampling. This method minimizes the execution time of each single thread and increases performance. Other studies focused on the pipeline design of such AMCs and the area that should be devoted to each component in the system [**? ?**]. Other works on AMCs focus on

hardware support for critical section detection [**?**] or bottleneck detection [**?** **?**]. These approaches are orthogonal to the ones evaluated in this paper and could benefit from them to further improve the final performance of the system.

Process scheduling on AMCs is one of the most challenging topics in this area of study. Bias scheduling [**?**] is an OS scheduler that characterizes the running threads according to their memory or execution intensity. It then schedules the computation intensive threads to the big cores of the system while the memory intensive threads to the little cores of the system. The experimental evaluation is done on Intel Xeon processors and the heterogeneous system is emulated by changing the configuration of three out of the four cores of the processor. Cong et al propose the Energy-Efficient [**?**] OS scheduler based on energy estimation. The evaluation is performed on the Intel QuickIA [**?**] platform that integrates an Intel Xeon with an Atom processor. Van Craeynest et al. [**?**] propose the fairness-aware OS scheduler that focuses on AMC architectures. The performance impact estimation (PIE) scheduler [**?**] is based on the impact of MLP and ILP on the overall CPI and focuses on improving performance. The scheduler predicts the impact of each different core-type of the system on the MLP, ILP and it assumes hardware support for CPI. Rodrigues et al [**?**] propose a thread scheduling technique that estimates power and performance when deciding to assign a thread to a specific core of the heterogeneous system. Finally, Energy-Aware Scheduling (EAS) is an on-going effort in the Linux community to introduce the energy factor in the OS scheduler [**?** **?**]. It is based on performance and power profiling to set performance and power capacities and let the Linux completely fair scheduler assign slots to processes considering the different core capacities. EAS is not yet part of the Linux kernel and, therefore, GTS is the most sophisticated state of the art scheduling method in production on current big.LITTLE processors.

Similar to OS scheduling approaches there have been many task scheduling approaches that are directed for utilizing AMCs. The Levelized Min Time [**?**] heuristic first clusters the tasks that can execute in parallel (*levels*) and then it assigns priorities to them, according to their execution time. The Dynamic Level Scheduling algorithm [**?**] assigns the tasks to the processors according to their *dynamic level* (DL). Heterogeneous Economical Duplication (HED) [**?**] duplicates the tasks in order to be executed on more than one cores but it then removes the redundant duplicates if they do not affect the makespan. CATS scheduler [**?**] is designed for AMCs like big.LITTLE and dynamically schedules the *critical* tasks to the big cores of the system to increase performance. Topcuoglu et al proposed the Heterogeneous Earliest Finish Time (HEFT) scheduler that statically assigns each task to the processor that will finish it at the earliest possible time. To do so, it keeps records with the task costs for each processor type. They also proposed the Critical Path on a Processor (CPOP) algorithm

[**?**] that maintains a list of tasks and statically identifies and schedules the tasks belonging to the critical path to the processor that minimizes the sum of their execution times. The Longest Dynamic Critical Path (LDCP) algorithm [**?**] identifies the tasks that belong to the critical path and schedules them with higher priority.

All these works reflect the remarkable research that is taking place on AMCs. However we consider that their experimental evaluation is limited for three main reasons: i) The evaluation is done through a simulator or emulation of an AMC [**? ? ? ? ? ? ? ? ? ? ? ? ?**]; ii) The evaluated applications are either random task dependency graph generators or scientific kernels and micro-benchmarks [**? ? ?**]. iii) Their evaluation does not focus on power and energy consumption [**? ? ? ? ?**].

This paper includes a unique evaluation of performance, power and energy on a real AMC of real parallel applications. This paper also reflects the impact of using different big and little core counts which is not present in previous works [**?**].

3.6 CONCLUSIONS

In this extensive evaluation of highly parallel applications on an ARM big.LITTLE AMC system we showed that current implementations of parallel applications using pthreads are not ready to fully utilize an AMC. Implementing highly sophisticated parallelization strategies such as parallel pipelines (ferret) to exploit AMCs at the application level requires a significant programming effort and is not applicable to all workloads. The built-in GTS heterogeneity-aware OS scheduler only partially mitigates the slowdown of static threading when using both big and little cores. Both dynamically-scheduled loop- and task-based versions achieve higher performance with increased utilization which results in increased power. This leads to similar energy consumption as static threading and GTS, which ends up with better results in EDP.

Overall, GTS and static threading are not suitable solutions to run intensive multi-threaded applications on AMCs. Dynamic scheduling is essential to distribute the load across different core types. A loop-based implementation with dynamic scheduling is appropriate when the parallel work granularity is large and the potential imbalance at the tail of the loop is insignificant compared to the overall parallel region duration. A task-based implementation with inter-task dependencies allows removing barriers, which is the preferred solution, especially when the granularity of parallel regions is small.

5

REFERENCES

CHAPTER 4

TASK-BASED SCHEDULING SOLUTIONS

CHAPTER 5

RUNTIME OVERHEADS MIGRATION

5.1 INTRODUCTION

Since the end of Dennard scaling [**?**] and the subsequent stagnation of CPU clock frequencies, computer architects and programmers rely on multi-core designs to achieve the desired performance levels. While multi-core architectures constitute a solution to the CPU clock stagnation problem, they bring important challenges both from the hardware and software perspectives. On the hardware side, multi-core architectures require sophisticated mechanisms in terms of coherence protocols, consistency models or deep memory hierarchies. Such requirements complicate the hardware design process. On the software side, multi-core designs significantly complicate the programming burden compared to their single-core predecessors. The different CPUs are exposed to the programmer, who has to make sure to use all of them efficiently, as well as using the memory hierarchy properly by exploiting both temporal and spatial locality. This increasing programming complexity, also known as the Programmability Wall [**?**], has motivated the advent of sophisticated programming paradigms and runtime system software to support them.

Task-based parallelism [**? ? ? ?**] has been proposed as a solution to the Programmability Wall and, indeed, the most relevant shared memory programming standards, like OpenMP, support tasking constructs [**?**]. The task based model requires the programmer to split the code into several sequential pieces, called tasks, as well as explicitly specifying their input and output dependencies. The task-based execution model (or runtime system) consists of a master thread and several worker threads. The master thread goes over the code of the application and creates tasks once it encounters source code annotations identifying them. The runtime system manages the pool of all created tasks and schedules them across the threads once their input dependencies are satisfied. To carry out the task management process, the parallel runtime system creates and maintains a Task Dependency Graph (TDG). In this graph nodes represent tasks and edges are dependencies between them. Once a new task is created, a new node is added to the TDG. The connectivity of this new node is defined by the data dependencies of the task it represents, which are explicitly specified in the application's source code. When the execution of a task finalizes, its corresponding node is removed from the TDG, as well as its data dependencies.

This task-based runtime system constitutes of a software layer that enables parallel programmers to decouple the parallel code from the underlying parallel architecture where it is supposed to run on. As long as the application can be decomposed into tasks, the task-based execution model is able to properly manage it across homogeneous many-core architectures or heterogeneous designs with different core types. A common practice in the high performance domain is to map a single thread per core, which enables the tasks running on that

thread to fully use the core capacity. Finally, another important asset of task-based parallelism is the possibility of automatically managing executions on accelerators with different address spaces. Since the input and output dependencies of tasks are specified, the runtime system can automatically offload a task and its dependencies to an accelerator device (e.g., GPU) without the need for specific programmer intervention [**?**]. Additional optimizations in terms of software pre-fetching [**?**] or more efficient coherence protocols [**?**] can also be enabled by the task-based paradigm.

Despite their advantages, task-based programming models also induce computational costs. For example, the process of task creation requires the traversal of several indexed tables to update the status of the parallel run by adding the new dependencies the recently created tasks bring, which produces a certain overhead. Such overhead constitutes a significant burden, especially on architectures with several 10's or 100's of cores where tasks need to be created at a very fast rate to feed all of them. This paper proposes the Task Generation Express (TaskGenX) approach. Our proposal suggests that the software and hardware are designed to eliminate the most important bottlenecks of task-based parallelism without hurting their multiple advantages. This paper focuses on the software part of this proposal and draws the requirements of the hardware design to achieve significant results. In particular, this paper makes the following contributions beyond the state-of-the-art:

- A new parallel task-based runtime system that decouples the most costly routines from the other runtime activities and thus enables them to be off-loaded to specific-purpose helper cores.

- A detailed study of the requirements of a specific-purpose helper core able to accelerate the most time consuming runtime system activities.

- A complete evaluation via trace-driven simulation considering 11 parallel OpenMP codes and 25 different system configurations, including homogeneous and heterogeneous systems. Our evaluation demonstrates how TaskGenX achieves average speedups of 3.1× when compared against currently use state-of-the-art approaches.

The rest of this document is organized as follows: Section 5.2 describes the task-based execution model and its main bottlenecks. Section 5.3 describes the new task-based runtime system this paper proposes as well as the specialized hardware that accelerates the most time-consuming runtime routines. Section 5.4 contains the experimental set-up of this paper. Section 5.5 describes the evaluation of TaskGenX via trace-driven simulation. Finally, Section 5.6 discusses related work and Section 5.7 concludes this work.

```
1       ...
2    //task_clause
3    memalloc(&task, args, size);
4    createTask(deps, task, parent, taskData);
5       ...
```

Listing 5.1 Compiler generated pseudo-code equivalence for task annotation.

```
1  void createTask(DepList dList, Task t,
2                  Task parent, Data args) {
3    initAndSetupTask(task1, parent, args);
4    insertToTDG(dList, task1);
5  }
```

Listing 5.2 Pseudo-code for task creation.

5.2 Background and Motivation

5.2 Task-based Programming Models

Task-based parallel programming models [**?** **?** **?** **?**], are widely used to facilitate the programming of parallel codes for multi-core systems. These programming models offer annotations that the programmer can add to the application's sequential code. One type of these annotations is the task annotations with dependency tracking which OpenMP [**?**] supports since its 4.0 release [**?**]. By adding these annotations, the programmer decomposes the application into *tasks* and specifies the input and output data dependencies between them. A compiler is responsible to translate the annotations into code by adding calls to the programming model's runtime system. The runtime system consists of software threads and is responsible for the efficient execution of the tasks with respect to the data dependencies as well as the availability of resources.

When the compiler encounters a task annotation in the code, it transforms it to the pseudo-code shown in Listing 5.1. `Memalloc` is performing the memory allocation for the task and its arguments. Next is a runtime call, which is the `createTask`, responsible for the linking of the task with the runtime system. At this point a task is considered *created* and below are the three possible states of a task inside the runtime system:

- *Created:* A task is initialized with the appropriate data and function pointers and it is inserted in the Task Dependency Graph (TDG). The insertion of a task in the TDG implies that the data dependencies of the tasks have been identified and the appropriate data structures have been created and initialized.

- *Ready:* When all the data dependencies of a created task have been satisfied, the task

is ready and it is inserted in the *ready queue* where it waits for execution.

- *Finished:* When a task has finished execution and has not been deleted yet.

The runtime system creates and manages the software threads for the execution of the tasks. Typically one software thread is being bound to each core. One of the threads is the *master thread*, and the rest are the *worker threads*. The master thread starts executing the code of Listing 5.1 sequentially. The allocation of the task takes place first. What follows is the task creation, that includes the analysis of the dependencies of the created task and the connection to the rest of the existing dependencies. Then, if there are no task dependencies, which means that the task is *ready*, the task is also inserted in the ready queue and waits for execution.

Listing 5.2 shows the pseudo-code for the task creation step within the runtime. The `createTask` function is first initializing the task by copying the corresponding data to the allocated memory as well as connecting the task to its parent task (`initAndSetupTask`). After this step, the task is ready to be inserted in the TDG. The TDG is a distributed and dynamic graph structure that the runtime uses to keep the information about the current tasks of the application. The insertion of a task in the TDG is done by the `insertToTDG` function. This function takes as arguments a list with all the memory addresses that are to be written or read by the task (`dList`), and the task itself. Listing 5.3 shows the pseudo-code for the TDG insertion. If for a task the `dList` is empty (line 2), this means that there are no memory addresses that need to be tracked during the execution; thus, the task is marked as *ready* by pushing it to the *ready queue* (line 3). Each entry of `dList` contains the actual memory address as well as the access type (read, write or read-write). The runtime keeps a distributed unified dependency tracking structure, the `depMap` where it stores all the tracked memory addresses together with their writer and reader tasks. For each item in the `dList` the runtime checks if there is an existing representation inside the `depMap` (line 8). If the memory address of an entry of the `dList` is not represented in the `depMap`, it is being added as shown in line 9. If the address of a `dList` item exists in the `depMap`, this means that a prior task has already referred to this memory location, exhibiting a data dependency. According to the access type of `d`, the readers and the writers of the specific address are updated in the `depMap` (lines 10-15).

To reduce the lookup into the `depMap` calls, every time the contents of a memory address are modified, the tasks keep track of their *successors* as well as the number of *predecessors*. The *successors* of a task are all the tasks with inputs depending on the output of the current task. The *predecessors* of a task are the tasks whose output is used as input for the current task. When a `read` access is identified, the task that is being created is added to the list of successors of the last writer task, as shown on line 20 of Listing 5.2.

```
1  void insertToTDG(DepList dList, Task t) {
2   if( dList is empty ) {
3     readyQ->push(t);
4     return;
5   }
6   Dependency entry;
7   for( d in dList ) {
8     entry = depMap[d.address()];
9     if(entry==NULL) depMap.add(entry, t);
10    if(d.accessType() == "write")
11      entry.addLastWriter(t);
12    if(d.accessType() == "read") {
13      entry.addReader(t);
14      entry.lastWriter()->addSuccessor(t);
15    }
16  }
17 }
```

Listing 5.3 Pseudo-code for TDG insertion

As tasks are executed, the dependencies between them and their successors are satisfied. So the successor tasks that are waiting for input, eventually become *ready* and are inserted to the ready queue. When a task goes to the *finished* state, the runtime has to perform some actions in order to prepare the successor tasks for execution. These actions are described in Listing 5.4. The runtime first updates the `depMap` to remove the possible references of the task as reader or writer (line 2). Then, if the task does not have any successors, it can safely be deleted (line 3). If the task has successors, the runtime traverses the successor list and for each successor task it decreases its predecessor counter (lines 5-6). If for a successor task its predecessor counter reaches zero, then this task becomes *ready* and it is inserted in the *ready queue* (lines 7-8). The runtime activity takes place at the task state changes. One state change corresponds to the task creation, so a task from being just allocated it becomes *created*. At this point the runtime prepares all the appropriate task and dependency tracking data structures as well as inserts the task into the TDG. The second change occurs when a task from being *created* it becomes *ready*; this implies that the input dependencies of this task are satisfied so the runtime schedules and inserts the task into the ready queue. The third change occurs when a running task finishes execution. In this case, following our task states, the task from being *ready* it becomes *finished*; this is followed by the runtime updating the dependency tracking data structures and scheduling possible successor tasks that become ready. For the rest of the paper we will refer to the first state change runtime activity as the task creation overheads (*Create*). For the runtime activity that takes place for the following two state changes (and includes scheduling and dependence analysis) we will

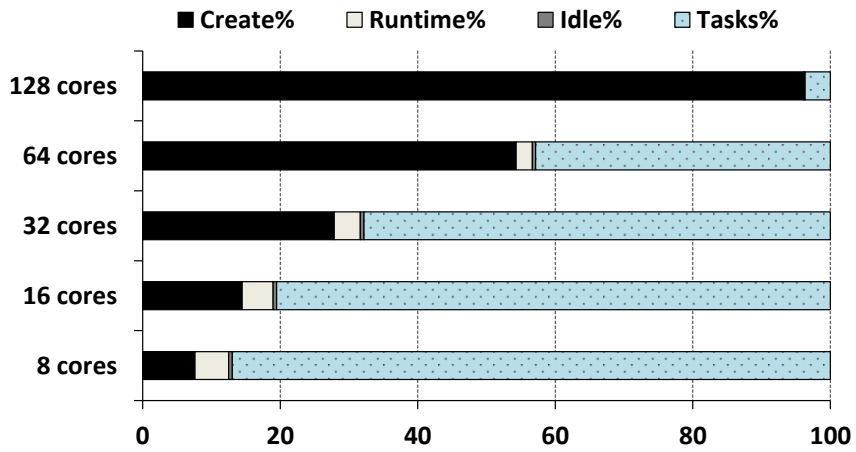


Fig. 5.1 Master thread activity for Cholesky as we increase the number of cores.

use the term runtime overheads (*Runtime*).

5.2 MOTIVATION

Figure 5.1 shows the runtime activity of the master thread during the execution of the Cholesky[1] benchmark on 8, 16, 32, 64 and 128 cores[2]. The execution time represented here is the wall clock time during the parallel region of the benchmark. Each one of the series represents a different runtime overhead from the ones described above. The percentage of time spent on task creation is increasing as we increase the number of cores. This is because the creation overhead is invariant of core count: the more we reduce the application's execution time by adding resources the more important this step becomes in terms of execution time. In contrast, the task execution time percentage is decreased as we increase the number of cores because the computational activity is being shared among more resources. One way to reduce the task creation overhead is by introducing nested parallelism. In this programming technique, every worker thread is able to generate tasks thus the task creation is spread among cores and its overhead is reduced. However, not all applications can be implemented with this parallelization technique and there are very few applications using this scheme. *Runtime* decreases as we increase the number of cores because this activity is also shared among the resources. This is because this part of the runtime takes place once the tasks finish execution and new tasks are being scheduled. So the more the resources, the less the runtime activity per thread, therefore less activity for the master thread.

Our motivation for this work is the bottleneck introduced by task creation as shown in

[1]Details about the benchmarks used are in Section 5.4
[2]The experimental set-up is explained in Section 5.4

```
1  void task_finish(Task *t) {
2    depMap.removeReaderWriter(t);
3    if(t->successors.empty()) delete t;
4    else {
5      for( succ in t->successors ) {
6        succ.decreasePredecessors();
7        if(succ.numPredecessors == 0)
8          readyQ->push(succ);
9      }
10   }
11 }
```

Listing 5.4 Pseudo-code for task_finish runtime activity.

Figure 5.1. Our runtime proposal decouples this piece of the runtime and accelerates it on a specialized hardware resulting in higher performance.

5.3 TASK GENERATION EXPRESS

In this paper we propose a semi-centralized runtime system that dynamically separates the most computationally intensive parts of the runtime system and accelerates them on specialized hardware. To develop the TaskGenX we use the OpenMP programming model [**?**], [**?**]. The base of our implementation is the Nanos++ runtime system responsible for the parallel execution and it is used in this paper as a replacement of the entire OpenMP's default runtime.

Nanos++ [**?**] is a distributed runtime system that uses dynamic scheduling. As most task-based programming models, Nanos++ consists of the master and the worker threads. The master thread is launching the parallel region and creates the tasks that have been defined by the programmer[3]. The scheduler of Nanos++ consists of a *ready queue* (*TaskQ*) that is shared for reading and writing among threads and is used to keep the tasks that are ready for execution. All threads have access to the *TaskQ* and once they become available they try to pop a task from the *TaskQ*. When a thread finishes a task, it performs all the essential steps described in Section 5.2.1 to keep the data dependency structures consistent. Moreover, it pushes the tasks that become ready to the *TaskQ*.

5.3 IMPLEMENTATION

TaskGenX relieves the master and worker threads from the intensive work of task creation

[3]Nanos++ also supports nested parallelism so any of the worker threads can potentially create tasks. However the majority of the existing parallel applications are not implemented using nested parallelism.

```
1 void SRTloop() {
2   while( true ) {
3     while(RRQ is not empty)
4       executeRequest( RRQ.pop() );
5     if( runtime.SRTstop() ) break;
6   }
7   return;
8 }
```

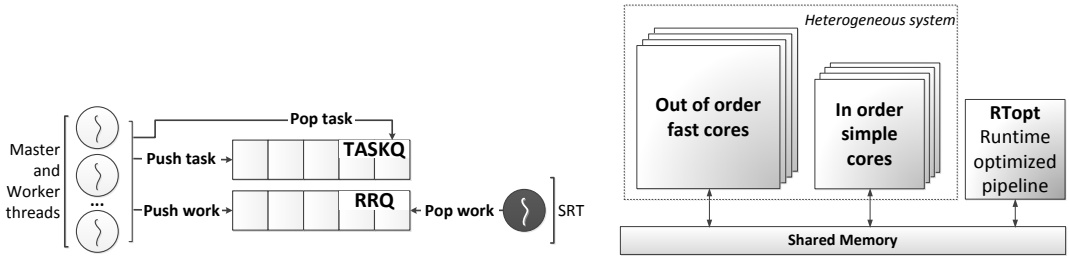
Listing 5.5 Pseudo-code for the SRT loop.

by offloading it on the specialized hardware. Our runtime, apart from the master and the worker threads, introduces the Special Runtime Thread (SRT). When the runtime system starts, it creates the SRT and binds it to the task creation accelerator, keeping its thread identifier in order to manage the usage of it. During runtime, the master and worker threads look for ready tasks in the task ready queue and execute them along with the runtime. Instead of querying the ready queue for tasks, the SRT looks for runtime activity requests in the Runtime Requests Queue (*RRQ*) and if there are requests, it executes them.

Figure 5.2a shows the communication infrastructure between threads within TaskGenX. Our system maintains two queues; the Ready Task Queue (*TaskQ*) and the Runtime Requests Queue (*RRQ*). The *TaskQ* is used to keep the tasks that are ready for execution. The *RRQ* is used to keep the pending runtime activity requests. The master and the worker threads can push and pop tasks to and from the *TaskQ* and they can also add runtime activity to the *RRQ*. The special runtime thread (SRT) pops runtime requests from the *RRQ* and executes them on the accelerator.

When the master thread encounters a task clause in the application's code, after allocating the memory needed, it calls the `createTask` as shown in Listing 5.2 and described in Section 5.2.1. TaskGenX decouples the execution of `createTask` from the master thread. To do so, TaskGenX implements a wrapper function that is invoked instead of `createTask`. In this function, the runtime system checks if the SRT is enabled; if not then the default behaviour takes place, that is, to perform the creation of the task. If the SRT is enabled, a *Create* request is generated and inserted in the *RRQ*. The *Create* runtime request includes the appropriate info to execute the code described in Listing 5.2. That is, the dependence analysis data, the address of the allocated task, its parent and its arguments.

While the master and worker threads are executing tasks, the SRT is looking for *Create* requests in the *RRQ* to execute. Listing 5.5 shows the code that the SRT is executing until the end of the parallel execution. The special runtime thread continuously checks whether there are requests in the *RRQ* (line 3). If there is a pending creation request, the SRT calls the `executeRequest` (line 4), which extracts the appropriate task creation data from the



(a) Communication mechanism between master/workers and SRT threads.

(b) SoC architecture including three types of cores: out of order, in-order and RTopt.

Fig. 5.2

creation request and performs the task creation by calling the `createTask` described in Listing 5.2. When the parallel region is over, the runtime system informs the SRT to stop execution. This is when the SRT exits and the execution finishes (line 5).

5.3 Hardware Requirements

As described in the previous section, TaskGenX assumes the existence of specialized hardware that accelerates the task creation step. The goal of this paper is not to propose a detailed micro-architecture of the specialized hardware; instead we sketch the high-level hardware requirements for the TaskGenX set-up, in the hope to be an insightful and useful influence for hardware designers. The SRT is bound to the task creation accelerator and executes the requests in the RRQ. Previous studies have proposed custom accelerators for the runtime activity [**? ? ? ? ? ?**]. These proposals significantly accelerate (up to three orders of magnitude) different bottlenecks of the runtime system[4]. These special purpose designs can only execute runtime system activity.

As an alternative, in our envisioned architecture we propose to have a general purpose core that has been optimized to run the runtime system activity more efficiently. The runtime optimized (RTopt) core can be combined with both homogeneous or heterogeneous systems and accelerate the runtime activity. Figure 5.2b shows the envisioned architecture when RTopt is combined with an asymmetric heterogeneous system. This architecture has three core types that consist of simple in-order cores, fast out-of-order cores and an RTopt core for the SRT. RTopt can optimize its architecture, having a different cache hierarchy, pipeline configuration and specialized hardware structures to hold and process the SRT. As a result, the RTopt executes the SRT much faster than the other cores. The RTopt can also execute tasks, but will achieve limited performance compared to the other cores as its hardware

[4]Section 5.6 further describes these proposals.

structures have been optimized for a specific software (the SRT).

To evaluate our approach we study the requirements of the RTopt in order to provide enough performance for TaskGenX. Based on the analysis by Etsion et al. [**?**], there is a certain *task decode rate* that leads to optimal utilization of the multi-core system. This rule can be applied in the case of TaskGenX for the *task creation rate*, i.e., the frequency of task generation of the runtime system. If the *task creation rate* is higher than the *task execution rate*, then for a highly parallel application the resources will always have tasks to execute and they will not remain idle. To achieve a high *task creation rate*, we can accelerate the task creation cost. Equation 5.1 shows the maximum optimal task creation cost, $C_{opt}(x)$ in order to keep x cores busy, without starving due to task creation.

$$C_{opt}(x) = avg.\ task\ duration/x \tag{5.1}$$

If C_{gp} is the cost of task creation when it is performed on a general purpose core, then the RTopt has to achieve a speedup of $r = C_{gp}/C_{opt}(x)$ to achieve full utilization of the system. Section 5.4.2 performs an analysis based on these requirements for the evaluated applications. As we will see in Section 5.4.2, a modest and implementable value of $r = 16\times$ is enough to significantly accelerate execution on a 512-core system.

Finally, if TaskGenX executes on a regular processor without the RTopt core, the SRT is bound to a regular core without any further modification. In this scenario, applications will not significantly benefit from having a separate SRT.

5.4 Experimental Methodology

5.4 Applications

Table 5.1 shows the evaluated applications, the input sizes used, and their characteristics. All applications are implemented using the OpenMP programming model [**?**]. We obtain Cholesky and QR from the BAR repository [**?**] and we use the implementations of the rest of the benchmarks from the PARSECSs suite [**?**]. More information about these applications can be found in [**?**] and [**?**]. As the number of cores in SoCs is increasing, so does the need of available task parallelism [**?**]. We choose the input sizes of the applications so that they create enough fine-grained tasks to feed up to 512 cores. The number of tasks per application and input as well as the average per-task CPU cycles can be found on Table 5.1.

Table 5.1 Evaluated benchmarks and relevant characteristics

Application	Problem size	#Tasks	Avg task CPU cycles (thousands)	Per task overheads (CPU cycles)			Measured perf. ratio	r	Parallel model
				Create	All	Deps + Sched			
Cholesky factorization	32K 256 32K 128	357 762 2829058	753 110	15221 17992	73286 58820	58065 40828	3.5	10.34 83.74	dependencies
QR factorization	16K 512 16K 128	11 442 707 265	518 570 3 558	17595 21642	63008 60777	45413 39135	6.8	0.01 3.11	dependencies
Blackscholes	native	488 202	348	29141	85438	56297	2.3	42.87	data-parallel
Bodytrack	native	329 123	383	9 505	18979	9474	4.2	12.70	pipeline
Canneal	native	3 072 002	67	25781	50094	24313	2.0	197.01	unstructured
Dedup	native	20 248	1 532	1294	9647	8353	2.7	0.43	pipeline
Ferret	native×2	84 002	29 088	38913	98457	59544	3.6	0.68	pipeline
Fluidanimate	native	128 502	16 734	30210	94079	64079	3.3	0.91	data-parallel
Streamcluster	native	3 184 654	161	6892	13693	6801	3.5	21.91	data-parallel

5.4 SIMULATION

To evaluate TaskGenX we make use of the TaskSim simulator [**?** **?**]. TaskSim is a trace driven simulator, that supports the specification of homogeneous or heterogeneous systems with many cores. The tracing overhead of the simulator is less than 10% and the simulation is accurate as long as there is no contention in the shared memory resources on a real system [**?**]. By default, TaskSim allows the specification of the amount of cores and supports up to two core types in the case of heterogeneous asymmetric systems. This is done by specifying the number of cores of each type and their difference in performance between the different types (performance ratio) in the TaskSim configuration file.

Our evaluation consists of experiments on both symmetric and asymmetric platforms with the number of cores varying from 8 to 512. In the case of asymmetric systems, we simulate the behaviour of an ARM big.LITTLE architecture [**?**]. To set the correct performance ratio between big and little cores, we measure the sequential execution time of each application on a real ARM big.LITTLE platform when running on a little and on a big core. We use the Hardkernel Odroid XU3 board that includes a Samsung Exynos 5422 chip with ARM big.LITTLE. The big cores run at 1.6GHz and the little cores at 800MHz. Table 5.1 shows the measured performance ratio for each case. The average performance ratio among our 11 workloads is 3.8. Thus in the specification of the asymmetric systems we use as performance ratio the value 4.

To simulate our approaches using TaskSim we first run each application/input in the TaskSim trace generation mode. This mode enables the online tracking of task duration and synchronization overheads and stores them in a trace file. To perform the simulation,

TaskSim uses the information stored in the trace file and executes the application by providing this information to the runtime system. For our experiments we generate three trace files for each application/input combination on a Genuine Intel 16-core machine running at 2.60GHz.

We modify TaskSim so that it features one extra hardware accelerator (per multi-core) responsible for the fast task creation (the RTopt). Apart from the task duration time, our modified simulator tracks the duration of the runtime overheads. These overheads include: (a) task creation, (b) dependencies resolution, and (c) scheduling. The RTopt core is optimized to execute task creation faster than the general purpose cores; to determine how much faster a task creation job is executed we use the analysis performed in Section 5.3.2.

Using Equation 5.1, we compute the $C_{opt}(x)$ for each application according to their average task CPU cycles from Table 5.1 for $x = 512$ cores. C_{gp} is the cost of task creation when it is performed on a general purpose core, namely the *Create* column shown on Table 5.1. To have optimal results for each application on systems up to 512 cores, C_{gp} needs to be reduced to $C_{opt}(512)$. Thus the specialized hardware accelerator needs to perform task creation with a ratio $r = C_{gp}/C_{opt}(512)\times$ faster than a general purpose core.

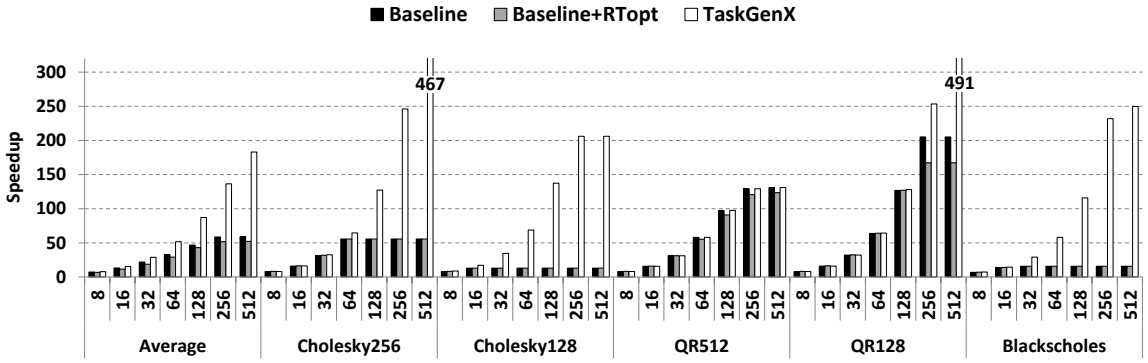
We compute r for each application shown on Table 5.1. We observe that for the applications with a large number of per-task CPU cycles and relatively small *Create* cycles (QR512, Dedup, Ferret, Fluidanimate), r is very close to zero, meaning that the task creation cost (C_{gp}) is already small enough for optimal task creation without the need of a faster hardware accelerator. For the rest of the applications, more powerful hardware is needed. For these applications r ranges from $3\times$ to $197\times$. Comparing r to the measured performance ratio of each application we can see that in most cases accelerating the task creation on a big core would not be sufficient for achieving higher task creation rate. In our experimental evaluation we accelerate task creation in the RTopt and we use the ratio of $16\times$ which is a relatively small value within this range that we consider realistic to implement in hardware. The results obtained show the average results among three different traces for each application-input.

5.5 EVALUATION

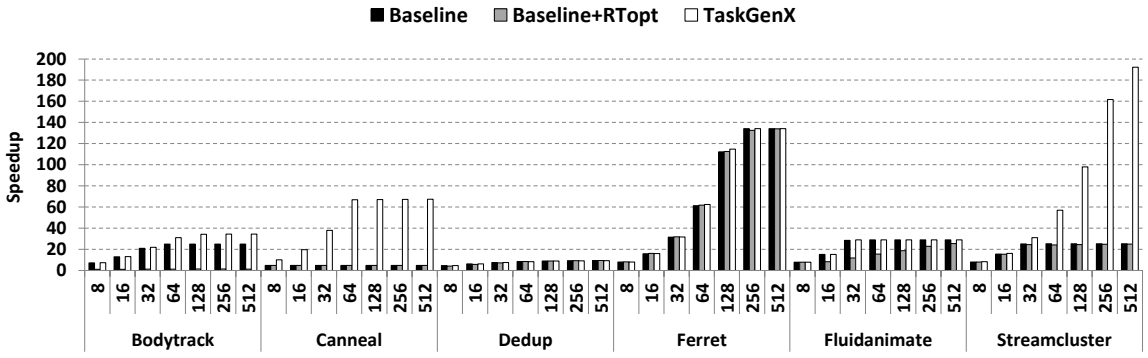
5.5 HOMOGENEOUS MULTICORE SYSTEMS

Figures 5.3a and 5.3b show the speedup over one core of three different scenarios:

- *Baseline*: the Nanos++ runtime system, which is the default runtime without using any external hardware support



(a)



(b)

Fig. 5.3 Speedup of TaskGenX compared to the speedup of Baseline and Baseline+RTopt for each application for systems with 8 up to 512 cores. The average results of (a) show the average among all workloads shown on (a) and (b)

- *Baseline+RTopt*: the Nanos++ runtime system that uses the external hardware as if it is a general purpose core
- *TaskGenX*: our proposed runtime system that takes advantage of the optimized hardware

We evaluate these approaches with the TaskSim simulator for systems of 8 up to 512 cores. In the case of Baseline+RTopt the specialized hardware acts as a slow general purpose core that is additional to the number of cores shown on the x axis. If this core executes a task creation job, it executes it 16× faster, but as it is specialized for this, we assume that when a task is executed on this core it is executed 4× slower than in a general purpose core. The runtime system in this case does not include our modifications that automatically decouple the task creation step for each task. The comparison against the Baseline+RTopt is used

only to show that the baseline runtime is not capable of effectively utilizing the accelerator. In most of the cases having this additional hardware without the appropriate runtime support results in slowdown as the tasks are being executed slower on the special hardware.

Focusing on the average results first, we can observe that TaskGenX constantly improves the baseline and the improvement is increasing as the number of cores is increased, reaching up to 3.1× improved performance on 512 cores. This is because as we increase the number of cores, the task creation overhead becomes more critical part of the execution time and affects performance even more. So, this becomes the main bottleneck due to which the performance of many applications saturates. TaskGenX overcomes it by automatically detecting and moving task creation on the specialized hardware.

Looking in more detail, we can see that for all applications the baseline has a saturation point in speedup. For example Cholesky256 saturates on 64 cores, while QR512 on 256 cores. In most cases this saturation in performance comes due to the sequential task creation that is taking place for an important percentage of the execution time (as shown in Figure 5.1). TaskGenX solves this as it efficiently decouples the task creation code and accelerates it leading to higher speedups.

TaskGenX is effective as it either improves performance or it performs as fast as the baseline (there are no slowdowns). The applications that do not benefit (QR512, Ferret, Fluidanimate) are the ones with the highest average per task CPU cycles as shown on Table 5.1. Dedup also does not benefit as the per task creation cycles are very low compared to its average task size. Even if these applications consist of many tasks, the task creation overhead is considered negligible compared to the task cost, so accelerating it does not help much.

This can be verified by the results shown for QR128 workload. In this case, we use the same input size as QR512 (which is 16K) but we modify the block size, which results in more and smaller tasks. This not only increases the speedup of the baseline, but also shows even higher speedup when running with TaskGenX reaching very close to the ideal speedup and improving the baseline by 2.3×. Modifying the block size for Cholesky, shows the same effect in terms of TaskGenX over baseline improvement. However, for this application, using the bigger block size of 256 is more efficient as a whole. Nevertheless, TaskGenX improves the cases that performance saturates and reaches up to 8.5× improvement for the 256 block-size, and up to 16× for the 128 block-size.

Blackscholes and Canneal, are applications with very high task creation overheads compared to the task size as shown on Table 5.1. This makes them very sensitive to performance degradation due to task creation. As a result their performance saturates even with limited core counts of 8 or 16 cores. These are the ideal cases for using TaskGenX as such bot-

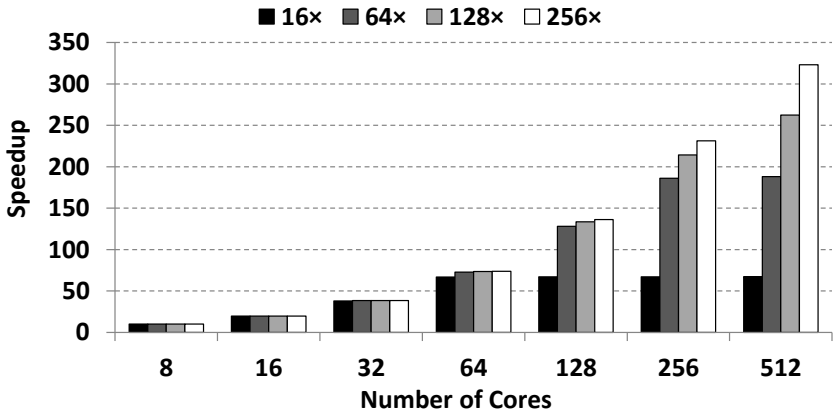


Fig. 5.4 Canneal performance as we modify r; x-axis shows the number of cores.

tlenecks are eliminated and performance is improved by 15.9× and 13.9× respectively. However, for Canneal for which the task creation lasts a bit less than half of the task execution time, accelerating it by 16 times is not enough and soon performance saturates at 64 cores. In this case, a more powerful hardware would improve things even more. Figure 5.4 shows how the performance of Canneal is affected when modifying the task creation performance ratio, r between the specialized hardware and general purpose. Using hardware that performs task creation close to 256× faster than the general purpose core leads to higher improvements.

Streamcluster has also relatively high task creation overhead compared to the average task cost so improvements are increased as the number of cores is increasing. TaskGenX reaches up to 7.6× improvement in this case.

The performance of Bodytrack saturates on 64 cores for the baseline. However, it does not approach the ideal speedup as its pipelined parallelization technique introduces significant task dependencies that limit parallelism. TaskGenX still improves the baseline by up to 37%. This improvement is low compared to other benchmarks, firstly because of the nature of the application and secondly because Bodytrack introduces nested parallelism. With nested parallelism task creation is being spread among cores so it is not becoming a sequential overhead as happens in most of the cases. Thus, in this case task creation is not as critical to achieve better results.

5.5 Heterogeneous Multicore Systems

At this stage of the evaluation our system supports two types of general purpose processors, simulating an asymmetric multi-core processor. The asymmetric system is influenced by the ARM big.LITTLE architecture [**?**] that consists of big and little cores. In our simulations,

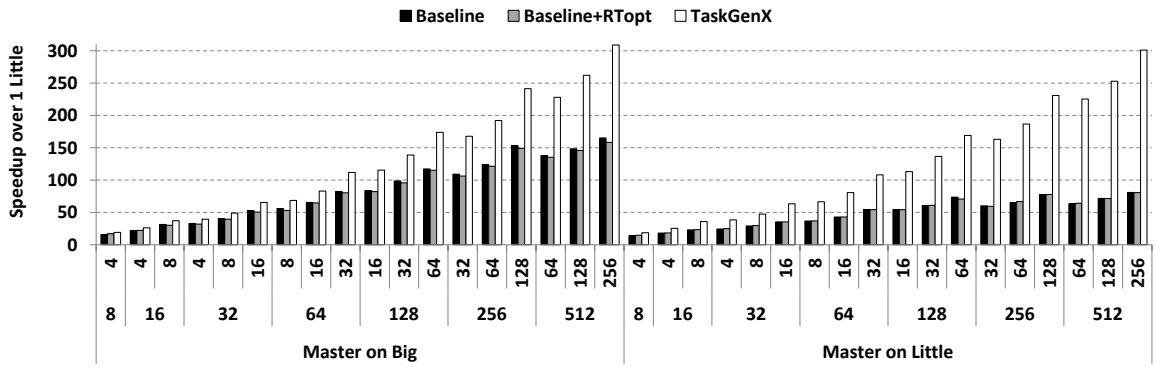


Fig. 5.5 Average speedup among all 11 workloads on heterogeneous simulated systems. The numbers at the bottom of x axis show the total number of cores and the numbers above them show the number of big cores. Results are separated depending on the type of core that executes the master thread: a big or little core.

we consider that the big cores are four times faster than the little cores of the system. This is based on the average measured performance ratio, shown on Table 5.1, among the 11 workloads used in this evaluation.

In this set-up there are two different ways of executing a task-based application. The first way is to start the application's execution on a big core of the system and the second way is to start the execution on a little core of the system. If we use a big core to load the application, then this implies that the master thread of the runtime system (the thread that performs the task creation when running with the baseline) runs on a fast core, thus tasks are created faster than when using a slow core as a starting point. We evaluate both approaches and compare the results of the baseline runtime and TaskGenX.

Figure 5.5 plots the average speedup over one little core obtained among all 11 workloads for the Baseline, Baseline+RTopt and TaskGenX. The chart shows two categories of results on the x axis, separating the cases of the master thread's execution. The numbers at the bottom of x axis show the total number of cores and the numbers above show the number of big cores.

The results show that moving the master thread from a big to a little core degrades performance of the baseline. This is because the task creation becomes even slower so the rest of the cores spend more idle time waiting for the tasks to become ready. TaskGenX improves performance in both cases. Specifically when master runs on big, the average improvement of TaskGenX reaches 86%. When the master thread runs on a little core, TaskGenX improves performance by up to 3.7×. This is mainly due to the slowdown caused by the migration of master thread on a little core. Using TaskGenX on asymmetric systems

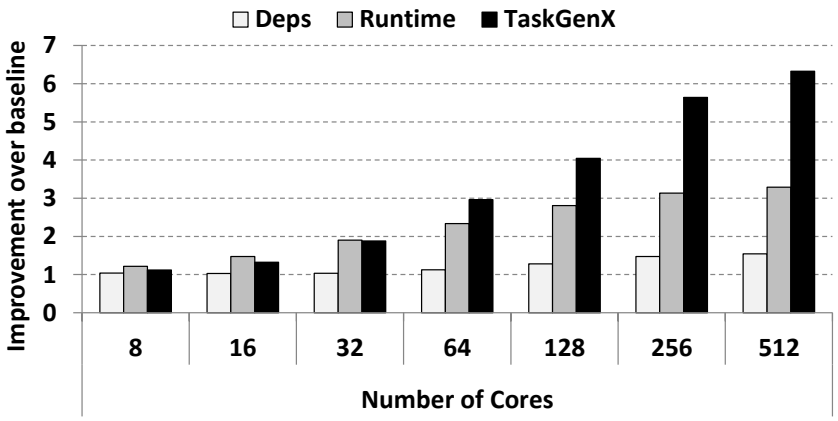


Fig. 5.6 Average improvement over baseline; *x*-axis shows the number of cores.

achieves approximately similar performance regardless of the type of core that the master thread is running. This makes our proposal more portable for asymmetric systems as the programmer does not have to be concerned about the type of core that the master thread migrates.

5.5 Comparison to Other Approaches

As we saw earlier, TaskGenX improves the baseline scheduler by up to 6.3× for 512 cores. In this section we compare TaskGenX with other approaches. To do so, we consider the proposals of Carbon [**?**], Task Superscalar [**?**], Picos++ [**?**] and Nexus# [**?**]. We group these proposals based on the part of the runtime activity they are offloading from the CPU. Carbon and Task Superscalar are runtime-driven meaning that they both accelerate all the runtime and scheduling parts. The task creation, dependence analysis as well as the scheduling, namely the ready queue manipulation, are transferred to the RTopt with these approaches. These overheads are represented on Table 5.1 under ALL. For the evaluation of these approaches one RTopt is used optimized to accelerate all the runtime activities. The second group of related designs that we compare against is the dependencies-driven, which includes approaches like Picos++ and Nexus#. These approaches aim to accelerate only the dependence analysis part of the runtime as well as the scheduling that occurs when a dependency is satisfied. The RTopt in this case is optimized to accelerate these activities. For example, when a task finishes execution, and it has produced input for another task, the dependency tracking mechanism is updating the appropriate counters of the reader task and if the task becomes ready, the task is inserted in the ready queue. The insertion into the ready queue is the scheduling that occurs with the dependence analysis. These overheads are represented on Table 5.1 under *Deps+Sched*.

Figure 5.6 shows the average improvement in performance for each core count over the performance of the baseline scheduler on the same core count. *Runtime* represents the runtime driven approaches and the *Deps* represents the dependencies driven approaches as described above. X-axis shows the number of general purpose cores; for every core count one additional RTopt core is used.

Accelerating the scheduling with *Runtime*-driven is as efficient as TaskGenX for a limited number of cores, up to 32. This is because they both accelerate task creation which is an important bottleneck. *Deps*-driven approaches on the other hand are not as efficient since in this case the task creation step takes place on the master thread.

Increasing the number of cores, we observe that the improvement of the *Runtime*-driven over the baseline is reduced and stabilized close to 3.2× while TaskGenX continues to speedup the execution. Transferring all parts of the runtime to RTopt with the *Runtime*-driven approaches, leads to the serialization of the runtime. Therefore, all scheduling operations (such as enqueue, dequeue of tasks, dependence analysis etc) that typically occur in parallel during runtime are executed sequentially on the RTopt. Even if RTopt executes these operations faster than a general purpose core, serializing them potentially creates a bottleneck as we increase the number of cores. TaskGenX does not transfer other runtime activities than the task creation, so it allows scheduling and dependence analysis operations to be performed in a distributed manner.

Deps driven approaches go through the same issue of the serialization of the dependency tracking and the scheduling that occurs at the dependence analysis stage. The reason for the limited performance of *Deps* compared to *Runtime* is that *Deps* does not accelerate any part of the task creation. Improvement over the baseline is still significant as performance with *Deps* is improved by up to 1.5×.

TaskGenX is the most efficient software-hardware co-design approach when it comes to highly parallel applications. On average, it improves the baseline by up to 3.1× for homogeneous systems and up to 3.7× for heterogeneous systems. Compared to other state of the art approaches, TaskGenX is more effective on a large number of cores showing higher performance by 54% over *Runtime* driven approaches and by 70% over *Deps* driven approaches.

5.6 Related Work

Our approach is a new task-based runtime system design that enables the acceleration of task creation to overcome important bottlenecks in performance. Task-based runtime systems have intensively been studied. State of the art task-based runtime systems include the

OpenMP [**?**], OmpSs [**?**], StarPU [**?**] and Swan [**?**]. All these models support tasks and maintain a TDG specifying the inter-task dependencies. This means that the runtime system is responsible for the task creation, the dependence analysis as well as the scheduling of the tasks. However, none of these runtime systems offers automatic offloading of task creation.

The fact that task-based programming models are so widely spread makes approaches like ours very important and also gives importance to studies that focus on adding hardware support to boost performance of task-based runtime systems. Even if their work focuses more on the hardware part of the design, their contributions are very relative to our study as we can distinguish which parts of the hardware is more beneficial to be accelerated.

Carbon [**?**] accelerates the scheduling of tasks by implementing hardware ready queues. Carbon maintains one hardware queue per core and accelerates all possible scheduling overheads by using these queues. Nexus# [**?**] is also a distributed hardware accelerator capable of executing the *in*, *out*, *inout*, *taskwait* and *taskwait on* pragmas, namely the task dependencies. Unlike Carbon and Nexus, TaskGenX accelerates only task creation. Moreover, ADM [**?**] is another distributed approach that proposes hardware support for the inter-thread communication to avoid going through the memory hierarchy. This aims to provide a more flexible design as the scheduling policy can be freely implemented in software. These designs require the implementation of a hardware component for each core of an SoC. Our proposal assumes a centralized hardware unit that is capable of operating without the need to change the SoC.

Task Superscalar [**?**] and Picos++ [**?**] use a single hardware component to accelerate parts of the runtime system. In the case of Task superscalar, all the parts of the runtime system are transferred to the accelerator. Picos++ [**?**] is a hardware-software co-design that supports nested tasks. This design enables the acceleration of the inter-task dependencies on a special hardware. Swarm [**?**] performs speculative task execution. Instead of accelerating parts of the runtime system, Swarm uses hardware support to accelerate speculation. This is different than our design that decouples only task creation.

Our work diverges to prior studies for two main reasons:

- The implementation of prior studies requires changes in hardware of the SoC. This means that they need an expensive design where each core of the chip has an extra component. Our proposal offers a much cheaper solution by requiring only a single specialized core that, according to our experiments, can manage the task creation for 512-core SoCs.

- None of the previous studies is aiming at accelerating exclusively task creation overheads. According to our study task creation becomes the main bottleneck as we increase the number of cores and our study is the first that takes this into account.

5.7 CONCLUSIONS

This paper presented TaskGenX, the first software-hardware co-design that decouples task creation and accelerates it on a runtime optimized hardware. In contrast to previous studies, our paper makes the observation that task creation is a significant bottleneck in parallel runtimes. Based on this we implemented TaskGenX on top of the OpenMP programming model. On the hardware side, our paper sets the requirements for the RTopt in order to achieve optimal results and proposes an asymmetric architecture that combines it with general purpose cores.

Based on this analysis we evaluate the performance of 11 real workloads using our approach with TaskSim simulator. Accelerating task creation, TaskGenX achieves up to 15.8× improvement (Cholesky128) over the baseline for homogeneous systems and up to 16× (Blackscholes) on asymmetric systems when the application is launched on a little core. Using TaskGenX on asymmetric systems offers a portable solution, as the task creation is not affected by the type of core that the master thread is bound to.

We further showed that for some cases like Canneal where task creation needs to be accelerated as much as 197× in order to steadily provide enough created tasks for execution. However, even by using a realistic and implementable hardware approach that offers 16× speedup of task creation, achieves satisfactory results as it improves the baseline up to 14×.

Comparing TaskGenX against other approaches such as Carbon, Nexus, Picos++ or TaskSuperscalar that manage to transfer different parts of the runtime to the RTopt proves that TaskGenX is the most minimalistic and effective approach. Even if TaskGenX transfers the least possible runtime activity to the RTopt hardware it achieves better results. This implies that TaskGenX requires a less complicated hardware accelerator, as it is specialized for only a small part of the runtime, unlike the other approaches that need specialization for task creation, dependency tracking and scheduling.

We expect that combining TaskGenX with an asymmetry-aware task scheduler will achieve even better results, as asymmetry introduces load imbalance.

CHAPTER 6

REAL-TIME SCHEDULING

CHAPTER 7

REAL-TIME SCHEDULING

CHAPTER 8

RELATED WORK

CHAPTER 9

STATE OF THE ART AND RELATED WORK

There has been a lot of studies on asymmetric multi-core systems. Some works focus on the design of the system, while other works explore the challenges that appear in efficiently utilizing such a heterogeneous system. Kumar et al [**?**] present the idea of an asymmetric multi-core system and proposed a feedback-based way to dynamically migrate processes among the different cores. To determine the core that most effectively executed a workload, Kumar et al [**?**] proposed the use of sampling. The proposed method minimizes the execution time of each single thread and increases performance. Other studies focused on the pipeline design of such asymmetric systems and the area that should be devoted to each component in the system [**?** **?**]. Other works on asymmetric systems focus on hardware support for critical section detection [**?**] or bottleneck detection [**?** **?**]. These approaches are orthogonal to the approaches evaluated in this paper and could benefit from them to further improve the final performance of the system.

Process scheduling on asymmetric systems is one of the most challenging topics in this area of study. Bias [**?**] scheduling is an operating system scheduler that characterizes the running threads according to their memory or execution intensity. It then schedules the computation intensive threads to the big cores of the system while the memory intensive threads to the little cores of the system. The experimental evaluation is done on Intel Xeon processors and the heterogeneous system is emulated by changing the configuration of three out of the four cores of the processor. Cong et al propose the Energy-Efficient [**?**] OS scheduler based on energy estimation. The evaluation is performed on the Intel QuickIA [**?**] platform that integrates an Intel Xeon with an Atom processor. Van Craeynest et al propose the fairness-aware OS scheduler [**?**] that focuses on asymmetric multi-core architectures. The performance impact estimation (PIE) scheduler [**?**] is based on the impact of MLP

and ILP on the overall CPI and focuses on improving performance. The scheduler predicts the impact of each different core-type of the system on the MLP, ILP and it assumes hardware support for CPI. Rodrigues et al [**?**] propose a thread scheduling technique that estimates power and performance when deciding to assign a thread to a specific core of the heterogeneous system.

Similar to OS scheduling approaches there have been many task scheduling approaches that are designed for utilizing asymmetric systems. The Levelized Min Time [**?**] heuristic first clusters the tasks that can execute in parallel (*levels*) and then it assigns priorities to them, according to their execution time. The Dynamic Level Scheduling algorithm [**?**] assigns the tasks to the processors according to their *dynamic level* (DL). Heterogeneous Economical Duplication (HED) [**?**] duplicates the tasks in order to be executed on more than one core but it then removes the redundant duplicates if they do not affect the makespan. Topcuoglu et al proposed the Heterogeneous Earliest Finish Time (HEFT) scheduler that statically assigns each task to the processor that will finish it at the earliest possible time. To do so, it keeps records with the task costs for each processor type. They also proposed the Critical Path on a Processor (CPOP) algorithm [**?**] that maintains a list of tasks and statically identifies and schedules the tasks belonging to the critical path to the processor that minimizes the sum of their execution times. The Longest Dynamic Critical Path (LDCP) algorithm [**?**] identifies the tasks that belong to the critical path and schedules them with higher priority.

All these works reflect the remarkable research that is taking place on asymmetric systems. However we consider that their experimental evaluation is limited for three main reasons:

- Their experimental evaluation is done through a simulator or emulation of an asymmetric system [**? ? ? ? ? ? ? ? ? ? ? ? ? ? ? ?**].
- The evaluated applications are either random task dependency graph generators or scientific kernels and micro-benchmarks [**? ? ?**].
- Their evaluation does not focus on power and energy consumption [**? ? ? ? ?**].

The search for efficient task scheduling on multi-core systems has been intensively studied. Most scheduling heuristics target homogeneous multiprocessors, nevertheless there exists an important number of studies in heterogeneous multiprocessors. In this section we give an overview of different categories of schedulers for heterogeneous systems, we explain some details about schedulers targeting specific systems using compute accelerators and explain details of previous works on criticality-aware schedulers.

9.0 SCHEDULERS FOR HETEROGENEOUS SYSTEMS

There are previous works on schedulers for heterogeneous systems that form four different types of schedulers: listing, clustering, guided-random, and duplication-based schedulers.

Listing schedulers [**?** **?** **?** **?** **?**] have two scheduling stages. In the first stage, each task is given a priority based on the policy defined in each algorithm. In the second stage, tasks are assigned to processors depending on their priorities. Most criticality-aware schedulers fall in this category, and we discuss them in Section 9.0.3. The scheduler proposed in this paper is also a list scheduler.

Clustering schedulers [**?** **?** **?** **?**] first separate tasks into clusters, where each cluster is to be executed on the same processor. During the clustering stage, the algorithm assumes an unlimited number of available processors in the system. If the number of clusters exceeds the number of available cores, the *merging* stage joins multiple clusters so that they match the number of available processors. An example is the Levelized Min Time [**?**] clustering scheduler. This heuristic clusters tasks that can execute in parallel according to their *level* (i.e. sibling nodes in a graph have the same level), and assigns priorities to the tasks in a cluster according to its execution time, (i.e. tasks with the highest execution time have the highest priority). The task-to processor assignment is done in decreasing order of priority.

Guided-random schedulers [**?** **?** **?**] randomize their schedules by applying policies influenced by other sciences. Genetic algorithms [**?**] group tasks into generations and schedule them according to a randomized genetic technique. Chemical reaction algorithms [**?**] mimic molecular interactions to map tasks to processors. Some of these guided-random approaches are designed for heterogeneous systems [**?** **?**]. The scheduler by Page et al. [**?**] enables dynamic scheduling of multiple-sized tasks for heterogeneous systems. However, it does not support dependencies between tasks.

Duplication-based schedulers [**?** **?** **?**] aim to eliminate communication costs between processors by scheduling tasks and their successors on the same processor. If a task has many successors, it is duplicated and executed in multiple cores prior to its successors so all successor tasks get the data from their predecessors with the lowest communication cost. This scheduling potentially introduces redundant duplications of tasks which may lead to bad schedules. The Heterogeneous Economical Duplication scheduler [**?**] performs task duplication in an economical manner as it removes the redundant duplicates if they do not affect performance.

These previous works schedule tasks statically and assume the prior knowledge of the task execution times on the different processor types in the heterogeneous system.

9.0 Schedulers for Compute Accelerators

The schedulers in the previous section target the scheduling of generic TDGs on generic heterogeneous architectures. In this section we cover schedulers that target specific systems with compute accelerators. These works are more focused on the scheduling of tasks on the target platform based on the abstractions provided by the corresponding mixture of programming models for the general-purpose processors and the compute accelerators in the system.

Most heterogeneous systems with compute accelerators nowadays combine general-purpose CPUs and GPU compute accelerators. There is a set of programming models providing abstractions to ease the development of applications on these platforms. OmpSs [**?** **?**] offers this abstraction by allowing multiple implementations of a given task to be executed on different processing units [**?**]. The scheduler then assigns the execution of a task to the best resource according to its earliest finish time. Another case is StarPU [**?**], a library that offers runtime heterogeneity support and provides priority schedulers for task-to-processor allocation. AHP [**?**] is another framework that generates software pipelines for heterogeneous systems and schedules tasks to their earliest executor, based on profiling information gathered prior to runtime.

None of these works, however, take into account the criticality of tasks regarding task dependencies, but they rather focus on the earliest execution time of individual tasks on the processor types in the specific system configuration.

9.0 Criticality-Aware Schedulers

Several previous works propose scheduling heuristics that focus on the critical path in a TDG to reduce total execution time [**?** **?** **?** **?**]. To identify the tasks in the critical path, most of these works use the concept of *upward rank* and *downward rank*. The upward rank of a task is the maximum sum of computation and communication cost of the tasks in the dependency chains from that task to an exit node in the graph. The downward rank of a task is the maximum sum of computation and communication cost of the tasks in the dependency chain from an entry node in the graph up to that task. Each task has an upward rank and downward rank for each processor type in the heterogeneous system, as the computation and communication costs differ across processor types.

The Heterogeneous Earliest Finish Time (HEFT) algorithm [**?**] maintains a list of tasks sorted in decreasing order of their upward rank. At each schedule step, HEFT assigns the task with the highest upward rank to the processor that finishes the execution of the task at the earliest possible time. Another work is the Longest Dynamic Critical Path (LDCP)

algorithm [**?**]. LDCP also statically schedules first the task with the highest upward rank on every schedule step. The difference between LDCP and HEFT is that LDCP updates the computation and communication costs on multiple processors of the scheduled task by the computation and communication cost in the processor to which it was assigned.

The Critical-Path-on-a-Processor (CPOP) algorithm [**?**] also maintains a list of tasks sorted in decreasing order as in HEFT, but in this case it is ordered according to the addition of their *upward rank* and *downward rank*. The tasks with the highest *upward rank* + *downward rank* belong to the critical path. On each step, these tasks are statically assigned to the processor that minimizes the critical-path execution time.

The main weaknesses of these works are that (a) they assume prior knowledge of the computation and communication costs of each individual task on each processor type, (b) they operate statically on the whole dependency graph, so they do not apply to dynamically scheduled applications in which only a partial representation of the dependency graph is available at a given point in time, and (c) most of them use randomly-generated synthetic dependency graphs that are not necessarily representative of the dependencies in real workloads.

CHAPTER 10

CONCLUSIONS

CHAPTER 11

CONCLUSION

The goal for this PhD thesis is to incorporate techniques of task and thread scheduling in order to fully utilize asymmetric systems. The main contributions of the thesis rely on the efficient exploitation of future asymmetric multi-core systems in terms of performance and energy efficiency as well as on the development of future asymmetric systems that fit the needs of high performance computing. Existing parallel scientific applications will become portable when moving from a traditional multi-core to an asymmetric multi-core system. Our useful observations throughout this study will also contribute and give guidelines for the design of the future multi-core asymmetric systems for high performance computing.

Our current results have shown that the state-of-the-art asymmetric multi-core systems are not ready to efficiently run out-of-the-box high performance applications and that the most efficient way is by using a task-based approach. This increases the need of research in this direction through the paths of scheduling and thread migration as described in the previous Chapters. In our first attempts to follow these paths, we have seen the high potential of the criticality-aware task schedulers to speed up dependency-intensive applications and take advantage of the asymmetric compute resources.

We are optimistic that following our second research approach of runtime thread migration will also contribute positively. The greatest challenge will be to increase performance without sacrificing energy, thus the dynamic search for the appropriate assistant core for the runtime thread has to consider all these obstacles. We expect that this approach will also influence designers to consider the use of assistant cores in the future asymmetric multi-cores. Finally, in our last and most complete approach we will need to synchronize all of our tools (e.g. scheduling and thread migration) to adapt to the runtime circumstances and boost performance with decent energy consumption.

Since a part of this work is already complete, we expect that our goals will be successfully accomplished and this study will be a useful reference for the future research.

CHAPTER 12

FUTURE DIRECTIONS

REFERENCES