

UC Irvine

UC Irvine Previously Published Works

Title

Intelligent Management of Mobile Systems through Computational Self-Awareness

Permalink

<https://escholarship.org/uc/item/8m93m9wj>

Authors

Donyanavard, Bryan
Rahmani, Amir M
Jantsch, Axel
[et al.](#)

Publication Date

2020-07-31

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Intelligent Management of Mobile Systems through Computational Self-Awareness

Bryan Donyanavard, Amir M. Rahmani, Axel Jantsch, Onur Mutlu, and Nikil Dutt

Abstract Runtime resource management for many-core systems is increasingly complex. The complexity can be due to diverse workload characteristics with conflicting demands, or limited shared resources such as memory bandwidth and power. Resource management strategies for many-core systems must distribute shared resource(s) appropriately across workloads, while coordinating the high-level system goals at runtime in a scalable and robust manner.

To address the complexity of dynamic resource management in many-core systems, state-of-the-art techniques that use heuristics have been proposed. These methods lack the formalism in providing robustness against unexpected runtime behavior. One of the common solutions for this problem is to deploy classical control approaches with bounds and formal guarantees. Traditional control theoretic methods lack the ability to adapt to (1) changing goals at runtime (i.e., *self-adaptivity*), and (2) changing dynamics of the modeled system (i.e., *self-optimization*).

In this chapter, we explore adaptive resource management techniques that provide self-optimization and self-adaptivity by employing principles of computational self-awareness, specifically *reflection*. By supporting these self-awareness properties, the system can reason about the actions it takes by considering the significance of competing objectives, user requirements, and operating conditions while executing unpredictable workloads.

Bryan Donyanavard
UC Irvine, e-mail: bdonyana@uci.edu

Amir M. Rahmani
UC Irvine e-mail: amirr1@uci.edu

Axel Jantsch
TU Vienna e-mail: axel.jantsch@tuwien.ac.at

Onur Mutlu
ETH Zurich e-mail: onur.mutlu@inf.ethz.ch

Nikil Dutt
UC Irvine e-mail: dutt@uci.edu

1 Introduction

Battery powered-devices are the most ubiquitous computers in the world. Users expect the devices to support high performance applications running on same device, sometimes at the same time. The devices support a wide range of applications, from interactive maps and navigation, to web browsers and email clients. In order to meet the performance demands of the complex workloads, increasingly powerful hardware platforms are being deployed in battery-powered devices. These platforms include a number of configurable knobs that allow for a tradeoff between power and performance, e.g., dynamic voltage and frequency scaling (DVFS), core gating, idle cycle injection, etc. These knobs can be set and modified at runtime based on workload demands and system constraints. Heterogeneous manycore processors (HMPs) have extended this principle of dynamic power-performance tradeoffs by incorporating single-ISA, architecturally differentiated cores on a single processor, with each of the cores containing a number of independent tradeoff knobs. All of these configurable knobs allow for a large range of potential tradeoffs. However, with such a large number of possible configurations, HMPs require intelligent runtime management in order to achieve application goals for complex workloads while considering system constraints. Additionally, the knobs may be interdependent, so the decisions must be coordinated. In this chapter, we explore the use of computational self-awareness to address challenges of adaptive resource management in mobile multiprocessors.

1.1 Computational Self-awareness

Self-aware computing is a new paradigm that does not strictly introduce new research concepts, but unifies overlapping research efforts in disparate disciplines [50]. The concept of self-awareness from psychology has inspired research in autonomous systems and neuroscience, and existing research in fields such as adaptive control theory support properties of self-awareness. This chapter addresses key challenges for achieving computational self-awareness that can make the design, maintenance and operation of complex, heterogeneous systems adaptive, autonomous, and highly efficient. Computational self-awareness is the ability of a computing system to recognize its own state, possible actions and the result of these actions on itself, its operational goals, and its environment, thereby empowering the system to become autonomous [42]. An infrastructure for system introspection and reflective behavior forms the foundation of self-aware systems.

1.1.1 Reflection

Reflection can be defined as *the capability of a system to reason about itself and act upon this information* [83]. A reflective system can achieve this by maintaining

a representation of itself (i.e., a self-model) within the underlying system, which is used for reasoning. Reflection is a key property of self-awareness. Reflection enables decisions to be made based on both *past* observations, as well as predictions made from past observations. Reflection and prediction involve two types of models: (1) a self-model of the subsystem(s) under control, and (2) models of other policies that may impact the decision-making process. Predictions consider *future* actions, or events that may occur before the next decision, enabling "what-if" exploration of alternatives. Such actions may be triggered by other resource managers running with a shorter period than the decision loop. The top half of Figure 1 (in blue) shows prediction enabled through reflection that can be utilized in the decision making process of a feedback loop. The main goal of the prediction model is to estimate system behavior based on potential actuation decisions. This type of prediction is most often performed using linear regression-based models [64, 71, 1, 81] due to their simplicity, while others employ a binning-based approach in which metrics sensed at runtime are used to classify workloads into categories [51, 18].

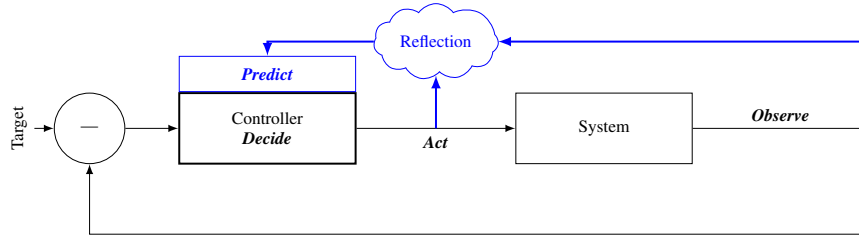


Fig. 1 Feedback loop overview. The bottom part of the figure represents a simple observe-decide-act loop. The top part (in blue) adds the reflection mechanism to this loop, enabling predictions for smart decision making.

1.2 Closed-loop Resource Management in Mobile Systems

Runtime resource management for many-core systems is increasingly challenging due to the complex interaction of: i) integrating hundreds of (heterogeneous) cores and uncore components on a single chip, ii) limited amount of system resources (e.g., power, cores, interconnects), iii) diverse workload characteristics with conflicting constraints and demands, and iv) increasing pressure on shared system resources from data-intensive workloads. As system size and capability scale, designers face a large space of configuration parameters controlled by actuation knobs, which in turn generate a very large number of cross-layer actuation combinations [103]. Making runtime decisions to configure knobs in order to achieve a simple goal (e.g., maximize performance) can be challenging. That challenge is exacerbated when considering a goal that may change throughout runtime, and consist of conflicting objectives (e.g., maximize performance while minimizing power consump-

tion). Additionally, ubiquitous mobile devices are expected to be general-purpose, supporting any combination of applications (i.e., workloads) desired by users, often without any prior knowledge of the workload.

Designers face a large space of configuration parameters that often are controlled by a limited number of actuation knobs, which in turn generate a very large number of cross-layer actuation configurations. For instance, Zhang and Hoffman [103] show that for an 8-core Intel Xeon processor, combining only a handful of actuation knobs (such as clock frequency and Hyperthreading levels) generates over 1000 different actuation configurations; they use binary search to efficiently explore the configuration space for achieving a *single* goal: cap the Thermal Design Power (TDP) while maximizing performance. Searching the configuration space is common practice in many similar single-goal, heuristic-based, runtime resource management approaches [73, 9, 90, 94, 10]. While there is a large body of literature on ad-hoc resource management approaches for processors using heuristics and thresholds [16, 44, 24, 14], rules [41, 17], solvers [68, 31], and predictive models [7, 22, 21, 19], there is a lack of formalism in providing guarantees for resource management of complex many-core systems.

Closed-loop systems have been used extensively to improve the state of a system by configuring knobs in order to achieve a goal. Closed-loop systems traditionally deploy an *Observe, Decide and Act* (ODA) feedback loop (lower half (in black) of Figure 1) to determine the system configuration. In an ODA loop, the observed behavior of the system is compared to the target behavior, and the discrepancy is fed to the controller for decision making. The controller invokes actions based on the result of the *Decide* stage.

Resource management approaches in the literature can be classified into three main classes: (1) heuristic-based-approaches [68, 31, 58, 27, 91, 89, 29, 41, 17, 26, 7, 22, 21, 16, 44, 103, 4, 96, 98, 102, 54, 86, 95, 67, 19, 24, 14, 25, 11, 13, 8, 88, 87], (2) control-theory-based approaches [57, 37, 74, 75, 60, 100, 72, 23, 55, 65, 45, 36, 84, 46, 30, 69], and (3) stochastic/machine-learning-based approaches [28, 7, 21, 15, 40]. Recent work has combined aspects of machine learning and feedback control [62]. In addition, there have been efforts to enable coordinated management in computer systems in various ways [7, 73, 9, 93, 43, 101, 90, 94, 10, 21, 25, 24, 14, 11, 85, 48, 12, 70]. These works coordinate and control multiple goals and actuators in a non-conflicting manner by adding an ad-hoc component or hierarchy to a controller.

In this chapter, we demonstrate the effectiveness of computational self-awareness in adaptive resource management for mobile processors. The self-aware resource managers discussed are implemented using classical and hierarchical control.

2 Self-optimization

Self-optimization is the ability of a system to adapt and act efficiently by itself in the face of *internal stimuli*. We consider internal stimuli as changes related to dy-

namics in the system's self-model, i.e., model inaccuracy. Internal stimuli does not necessarily include workload itself, but if the self-model is application-dependent, workload changes may be the source of internal stimuli. For example, if the system's self-model is application-dependent, and the executing application changes, a self-optimizing manager will have the ability to reason and act towards achieving the system goal(s) efficiently for the new application. However, if the system's self-model is rigid and the system dynamics used to reason and act are oversimplified, model inaccuracies may lead to undesirable or inefficient decisions when the application changes.

2.1 Background and Motivation

Dynamic voltage/frequency scaling (DVFS) has been established as an effective technique to improve the power-efficiency of chip-multiprocessors (CMPs) [34]. In this context, numerous closed-loop control-theoretic solutions for chip power management [74, 38, 61, 66, 56, 97] have been proposed. These solutions employ *linear control* techniques to limit the power consumption by controlling the CMP operating frequency. However, the relationship between operating frequency and power is often *nonlinear*. Figure 2 illustrates this by showing total power consumed by a 4-core ARM A15 cluster executing a CPU-intensive workload through its entire frequency range (200MHz–2GHz), along with the total power consumed by a 4-core ARM A7 cluster through its frequency range (200MHz–1400MHz). While the A7 cluster frequency-power relationship is almost linear, the A15 cluster's larger frequency range (and more voltage levels) results in a nonlinear relationship. Using a linear model to estimate the behavior of such a system leads to inaccuracies. Inaccurate models result in inefficient controllers, which defeats the very purpose of using control theoretic techniques for power management.

Ideally, control-theoretic solutions should provide formal guarantees, be simple enough for runtime implementation, and handle nonlinear system behavior. Static linear feedback controllers can provide robustness and stability guarantees with simple implementations, while adaptive controllers modify the controller at runtime to adapt to the discrepancies between the expected and the actual system behavior. However, modifying the controller at runtime is a costly operation that also invalidates the formal guarantees provided at design time.

Instead, consider integrating multiple linear models within a single controller implementation in order to estimate nonlinear behavior of DVFS for CMPs. This is a well-established and lightweight adaptive control theoretic technique called *gain scheduling*.

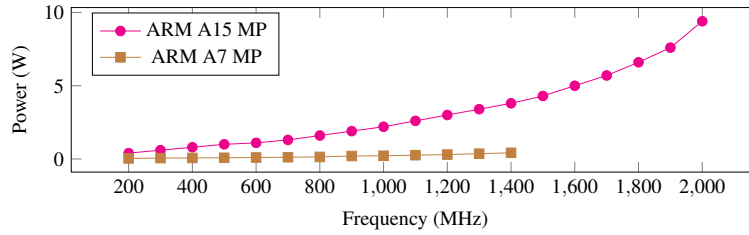


Fig. 2 Cluster power vs. operating frequency (from [20]).

2.1.1 Classical Control

Discrete-time control techniques are the most appropriate to implement control of computer systems. The proportional-integral-derivative (PID) controller is a simple and flexible classical feedback controller that computes control input $u(t)$ based on the error $e(t)$ between the measured output and reference output:

$$u(k) = K_p e(k) + K_i \sum_0^k e(k) \Delta t + K_d \frac{\Delta e(k)}{\Delta t} \quad (1)$$

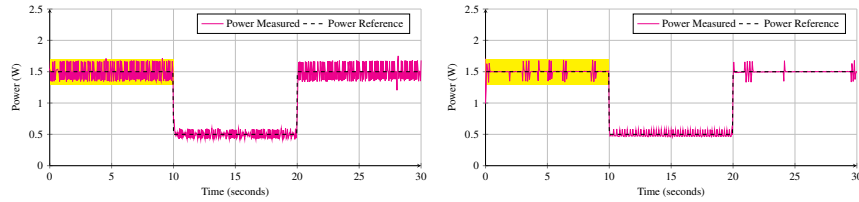
K_p , K_i , and K_d are control parameters for the proportional, integral, and derivative gains respectively.

PI controllers¹ have been successfully used to manage DVFS of CMPs [61, 99, 66, 56, 97]. Mishra et al. [61] propose the use of PID controllers for VF islands. The authors model power consumption based on the assumption that the difference relationship between power consumption in successive intervals can be approximated linearly as a function of frequency, which only holds for limited range. Similarly, Hoffman et al. [38] propose a feedback control technique for power management that includes DVFS, and their transfer function assumes a linear relationship between power and frequency. However, Figure 2 shows that $f \rightarrow P$ becomes non-linear at higher frequencies. Inaccuracies in linear estimation of nonlinear systems can negatively impact the steady-state error and transient response of the controller. Take for example a system operating under a power budget, or experiencing a thermal emergency – a DVFS controller designed from an inaccurate model could lead to wasted power or even unnecessary operation at an unsafe frequency.

Consider a DVFS controller for a 4-core CMP with a single frequency domain. The first steps in designing a controller are defining the system and identifying the model. The power consumption of our CMP is not linear across the entire range of supported operating frequencies (200MHz–2GHz), which makes it challenging to model the entire range with a single linear estimation. However, we can divide the measured output (power) for the entire range of frequencies into multiple *operating regions* that exhibit linear behavior. In this example, we identify a model for two different systems: (1) the CMP’s behavior through all operating frequencies; (2)

¹ Due to the significant stochastic component of computer systems, PI controllers are preferred over PID controllers [33].

the CMP's behavior through a sub-range of the operating frequencies. This specific operating region spans the frequency sub-range of 200MHz–1200MHz. Using these models, we can generate two different $f \rightarrow P$ Single-Input-Single-Output (SISO) PI controllers, and compare them using measured SASO analysis [33], focusing on *Accuracy* and *Settling time*. We refer to the full-range controller as Controller 1, and the sub-range controller as Controller 2. Figure 3 displays Controller 1 (Fig. 3a) and Controller 2's (Fig. 3b) ability to track a dynamic power reference over time for our CMP.



(a) Full range SISO controller (Controller 1). (b) Sub-range SISO controller (Controller 2).

Fig. 3 Time plots of two DVFS controllers tracking a dynamic power reference (from [20]).

Accuracy is defined by the steady-state error between the measured output and reference input, e.g., the yellow highlighted region in Figure 3b from 0-10 seconds. We calculate the steady-state error as the mean squared error (MSE) between the measured power and reference power. Both controllers are able to track within 1% of the target power. However, the MSE of Controller 2 is 0.003, while that of Controller 1 is 0.013 – an order of magnitude larger. This byproduct of model inaccuracy translates into wasted power and undesirable operating frequency, as well as unnecessary changes in the frequency control input (i.e., increased control effort cost).

Settling time is the time it takes to reach sufficiently close to the steady-state value after the reference values are specified, e.g., when the reference changes in Figure 3b at 10 seconds. The settling time of Controller 2 is 40ms on average, while Controller 1 is more than double on average at 100ms. Because our actuation periods are 50ms, this means that our sub-range controller often reaches steady state on its first actuation while the full range controller requires multiple actuation periods to respond to a change in reference.

Identifying operating regions at design time allows us to switch system models at runtime, improving the effectiveness of static controllers.

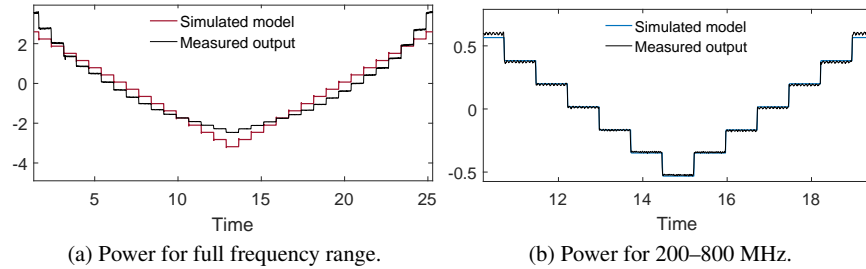


Fig. 4 Modeled and observed behavior of nonlinear full-range system (a) vs. linear operating region (b) (from [20]).

2.2 Case Study: Designing a Gain Scheduled Controller (GSC) for Power Management

As a demonstrative case study, we target the ODROID-XU3 platform [32] which contains an ARM big.LITTLE based Exynos 5422 Octa-core SoC that has heterogeneous multi-processing (HMP) cores. The Exynos platform contains an HMP with two 4-core clusters: the *big* cluster provides high-performance out-of-order cores, while the *little* cluster provides low-power in-order cores. For the purpose of our study, we disable the little cluster (due to its linear behavior) and use only the big cores to emulate a uniform nonlinear CMP².

2.2.1 Defining and Modeling Linear Subsystems

Selecting the control input and measured output of a DVFS controller is straightforward. Frequency is the knob available to the user in software, and power is the metric of interest. On our Exynos CMP, the operating frequency of cores is set at the cluster level, and power sensors measure power at the cluster level. A SISO controller is a natural solution, with the entire CMP composing the system under control.

For system identification we generate test waveforms from applications and use statistical black-box methods based on System Identification Theory [52, 53] for isolating the deterministic and stochastic components of the system to build the model.

Figure 4a shows a comparison of a simulated model output vs. the measured output over the entire frequency range of our CMP. It is evident that there are ranges for which the estimated behavior differs from that of the actual system behavior. We know that voltage has a nonlinear effect on dynamic power ($P = CV^2f$). The nonlinear relationship between frequency and voltage pairs through the range of operating frequencies amplifies this effect (Table 1). Table 1 lists all valid VF pairs

² We refer to this as the Exynos CMP or CMP throughout.

Table 1 VF Pairs for ARM A15 in Exynos 5422.

Region	Frequency Range (MHz)	Voltage (V)
1	1600 – 2000	1.25
2	1300 – 1500	1.10
3	900 – 1200	1.00
4	200 – 800	0.90

for the CMP, in which there are only four different voltage levels. Figure 4b shows the measured vs. modeled output when the system is defined by a single operating region grouped by frequencies that operate at the same voltage level.

2.2.2 Generating Linear Controllers

We generate a PI controller separately for each operating region using the system models and MATLAB’s Control System toolbox. This is a straightforward process for a simple off-the-shelf PI controller.

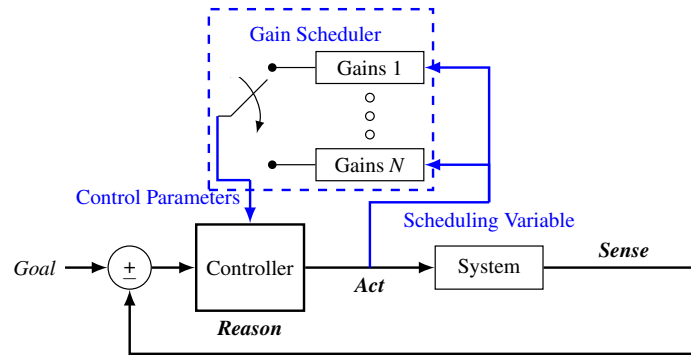


Fig. 5 Block diagram of GSC.

In the next step, the designed controller is evaluated against disturbance and uncertainties in order to ensure it remains stable at a defined confidence level. Unaccounted elements, modeling limitations, and environmental effects are estimated as model uncertainty in order to check the disturbance rejection of the controller. In our case, we can confirm our controller is robust enough to reject the disturbance from workload variation.

Each controller we design for an operating region is defined by its control parameters K_P and K_I which are stored (in memory) in the gain scheduler (Figure 5). In the gain scheduler, we incorporate logic to determine which gains to provide the controller when invoked.

2.2.3 Implementing Gain Scheduling

The gain scheduler enables us to adapt to nonlinear behavior (Figure 5) by combining multiple linear controllers. It stores predefined controller gains and is responsible for providing the most appropriate gains based on the operating region in which the system currently resides each time the controller is invoked.

Algorithm 1 Gain Scheduler Implementation

Input: f : frequency, scheduling variable;

Outputs: K_{P_n} , K_{I_n} , $offset_n$: updated controller parameters;

Variables: ref_{prev} , ref_{next} : power reference values for previous and next control periods;

Constants: $Region[N]$: operating regions, defined by mutually exclusive range of frequencies; $K_P[N]$, $K_I[N]$, $offset[N]$: stored controller parameters for each operating region; K_{P_G} , K_{I_G} , $offset_G$: controller parameters for full-range linear controller;

```

1: if  $ref_{next} \neq ref_{prev}$  then
2:    $K_{P_n} = K_{P_G}$ 
3:    $K_{I_n} = K_{I_G}$ 
4:    $offset_n = offset_G$ 
5:   return
6: else
7:   for  $i = 1$  to  $N$  do
8:     if  $Region[i].contains(f)$  then
9:        $K_{P_n} = K_P[i]$ 
10:       $K_{I_n} = K_I[i]$ 
11:       $offset_n = offset[i]$ 
12:      return
13:     end if
14:   end for
15: end if

```

The scheduling variable is the variable used to define operating regions. For our controller, the scheduling variable is frequency as it is simpler to implement in software and has a direct VF mapping (Table 1). Our gain scheduler implements lightweight logic that determines the set of gains based on the system's operating frequency (scheduling variable). Algorithm 1 shows the logic implemented in our gain scheduler with N operating regions where f is the scheduling variable and K_P and K_I are the controller parameters. In addition to the K_P and K_I controller parameters, there is also an *offset*. The *offset* is the mean actuation value for the operating region, and is necessary for providing the control input for the next control period. Algorithm 1 accounts for the transitions between operating regions (lines 1-6) by applying a full-range linear controller. This method is utilized as the sets of gains for a particular operating region perform poorly outside of that region.

2.2.4 Experiments

Our goal is to evaluate our nonlinear GSC with respect to the state-of-the-art linear controller in terms of both theoretical and observed ability to track power goals on a CMP. Our evaluation is done using the Exynos CMP running Ubuntu Linux.³ We consider a typical mobile scenario in which one or more multi-threaded applications execute concurrently across the CMP.

Controllers

We designed two DVFS controllers for power management of the CMP: 1) a **linear controller** that estimates the transfer function similarly to [38, 61]; and our proposed 2) **GSC**. The GSC contains three operating regions (Table 2). We combine the two smallest adjacent Regions, 1 and 2 (Table 1), to create Controller 2.1. Controllers are provided a single power reference for the whole system. The control input is frequency, and the measured output is power, applied to the entire CMP.

The controller is implemented as a Linux userspace process that executes in parallel with the applications. Power is calculated using the on-board current and voltage sensors present on the ODROID board. Power measurements and controller invocation are performed periodically every *50ms*.

Workloads

We developed a custom micro-benchmark used for system identification. The micro-benchmark consists of a sequence of independent multiply-accumulate operations yielding varied instruction-level parallelism. This allows us to model a wide range of behavior in system outputs given changes in the controllable inputs. We test our controllers using three PARSEC benchmarks: *bodytrack*, *streamcluster*, and *x264*. For each case, we execute one multithreaded application instance of the benchmark with four threads, resulting in a fully-loaded CMP. We empirically select three references that we alternate between during execution. ref_1 is 3.5W, the highest reference and a reasonable power envelope for a mobile SoC. This represents a high-performance mode that maximizes performance under a power budget. ref_2 is 0.5W, the lowest reference and represents a reduced budget in response to a thermal event. ref_3 is 1.5W, a middling reference that could represent the result of an optimizer that maximizes energy efficiency. These references are not necessarily trackable for all workloads, but should span at least three different operating regions for each workload. For each case, the applications run for a total of 65s. After the first 5s (warm-up period) the controllers are set to ref_1 for 20s, then changed to ref_2 for 20s, and to ref_3 for the remaining 20s.

³ Ubuntu 16.04.2 LTS and Linux kernel 3.10.105

2.2.5 Controller Design Evaluation

	Ctrl 1	Ctrl 2.1	Ctrl 2.2	Ctrl 2.3
Freq. Range	200 – 1800	1300 – 1800	900 – 1200	200 – 800
Stable	✓	✓	✓	✓
Accuracy (MSE)	0.1748	0.03089	0.0005382	0.0003701

Table 2 Accuracy of the full- (Ctrl 1) and sub-range (Ctrl 2.x) controllers.

We used a first-order system, with a target crossover frequency of 0.32. This resulted in a simple controller providing the fastest settling time with no overshoot. Models are generated with a stability focus and uncertainty guardbands of 30%.

All systems are stable according to Robust Stability Analysis. By design all overshoot values are 0. The settling times of Controllers 2.2 and 2.3 are comparably low at 5 control periods. Controller 2.1 (the most nonlinear operating region) and Controller 1 are slightly higher at 8-9 control periods. The ideal controllers are all very similar in terms of stability, settling time, and overshoot. The primary difference between them is in terms of accuracy. Controllers 2.1-2.3 achieve an order of magnitude better accuracy than Controller 1 (Table 2). This means that the region controllers are equally as responsive as the full-range model in achieving a target value while achieving the value more accurately.

2.2.6 Controller Implementation Evaluation

We now evaluate the effectiveness of our nonlinear control approach implemented in software on the Exynos CMP for multithreaded mobile workloads. Traditional SASO control analysis gives us a way to compare the controllers in theory, but the system-level effects of those metrics are not directly relatable. Therefore, we will compare the runtime behavior of the software controllers using a slightly modified set of metrics: power over target, power under target, number of actuations, and response time. These metrics are shown in Figure 6.

The *power over target* is the total amount of measured power exceeding the reference power throughout execution (Fig. 6a). This is the area under the output and above the reference. It represents the amount of power wasted due to inaccuracy, and can also represent unsafe execution above a power cap. Our GSC is able to achieve **12%** less power over target than the linear controller for `x264` and `streamcluster`. `bodytrack` is the most dynamic workload and results in the noisiest power output. In this case the GSC only improves the power over target by 1% compared to the linear controller.

The *power under target* is the total amount of measured power falling short of the reference power throughout execution (Fig. 6b). This is the area under the reference and above the output. A lower value translates to improved performance (i.e. lower

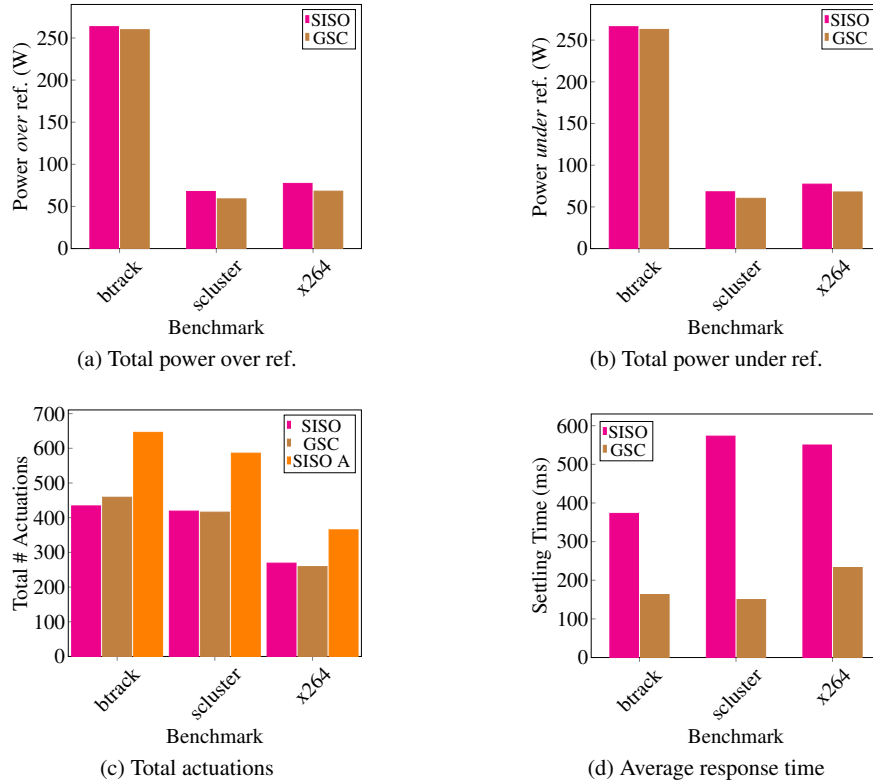


Fig. 6 Comparison of GSC with Controller 1 (from [20]).

is better). Similarly to the power over target, our GSC is able to reduce power under target by **12%** for `x264` and `streamcluster`, and **1%** for `bodytrack`.

The *number of actuations* is simply a count of how many times the frequency changes throughout execution, and is a measure of overhead (Fig. 6c). The GSC's actuation overhead is lower than the linear controller for `bodytrack`, `streamcluster`, and `x264` by **8%**, **1%**, and **4%** respectively. This is expected, as the controller's resistance to actuation is related to the crossover frequency specified at design time. For the same crossover frequency, the GSC benefits are primarily in the accuracy (power over/under target) and response (settling) time. To illustrate this tradeoff, we performed the same experiments for a full-range linear controller with a target crossover frequency of 0.8 (Controller 1b). We arrived at this value empirically: Controller 1b achieves comparable accuracy to the GSC. However, GSC reduces the actuation overhead by **29%** for all workloads compared to Controller 1b.

The *response time* is the average settling time when the target power changes, indicating the controller's ability to respond quickly to changes (Fig. 6d). Figure 6d shows the average response time for each workload for both controllers. The GSC

is able to improve the response time over Controller 1 by more than **50%** in each case. The GSC's overall average response time is $182ms$, which is less than 4 control periods.

The implementation overhead of the GSC w.r.t. the linear controller is negligible: it requires a single execution of Algorithm 1 upon each invocation, and storage for a K_P , K_I , and *offset* value for each operating region. Although workload disturbance plays a significant role in determining the magnitude in improvement of a nonlinear GSC over a state-of-the-art linear controller, a clear trend exists, and these advantages would increase with the modeled system's degree of nonlinearity.

2.3 Summary

Self-models are the core components of self-awareness. In computer systems, system dynamics can be complex. When utilizing a self-model at runtime for reflection, models must be simple and sufficiently accurate. The more accurate the self-model, the more effective the decisions made by a resource manager can be toward achieving a given goal. We propose a simple way to improve the accuracy of self-models for resource managers employing classical controllers: gain scheduling. Gain scheduled control generates multiple controllers based on optimized fixed models for different operating regions of the system, and can deploy the most accurate control at runtime based on the system state. This is an improvement over using a single controller based on a single fixed model with minimal overhead. In our case study, the gain scheduled controller more effectively provides dynamic power management of a single-core processor when compared to a single fixed SISO controller.

Using a static model for resource management may not be sufficient in complex mobile systems: system dynamics may change between applications or devices, and fixed models may not remain accurate over time. In the future, we plan to address such scenarios by identifying and continuously updating models during runtime based on observation, instead of identifying multiple fixed models at design-time to swap out at runtime.

3 Self-adaptivity

Self-adaptivity is the ability of a system to adjust to changes in goals due to external stimuli. For example, if a system experiences a thermal event during a computational sprint and enters an unsafe state, a self-adaptive manager will have the ability to modify the goal from maximizing performance to minimizing temperature.

3.1 Motivation and Background

To address self-optimization, we examined a relatively simple use-case in which we deployed a resource manager responsible for controlling a uniprocessor with only a single input and single output. However, modern computer systems incorporate up to hundreds of cores, from datacenters to mobile devices. Modern mobile devices commonly deploy architecturally differentiated cores on a single chip multiprocessor, known as heterogeneous multiprocessors (HMPs). In the case of mobile devices, systems are tasked with the challenge of balancing application goals with system constraints, e.g., a performance requirement within a power budget. Resource managers are required to configure the system at runtime to meet the goal. However, due to workload or operating condition variation, it is possible for goals to change unpredictably at runtime. In this section, we use self-adaptivity to enable a mobile HMP resource manager to adapt to a changing goal, coordinating and prioritizing multiple objectives.

3.1.1 Managing Dynamic System-wide Goals

Controllers may behave non-optimally, or even detrimentally, in meeting a shared goal without knowledge of the presence or behavior of seemingly orthogonal controllers [76, 7, 94, 23, 24, 11]. Consider the MIMO controller in Figure 7 that controls a single-core system with two control inputs ($u(t)$) and interdependent measured outputs ($y(t)$) [69]. The controller tracks two objectives (frames per second, or FPS, and power consumption) by controlling two actuators (operating frequency and cache size). We implement the MIMO using a Linear Quadratic Gaussian (LQG) controller [82] similarly to [69]:

$$x(t+1) = A \times x(t) + B \times u(t) \quad (2)$$

$$y(t) = C \times x(t) + D \times u(t) \quad (3)$$

where x is the system state, y is the measured output vector, and u is the control input vector.⁴

LQG control allows us to specify 1) the relative sensitivity of a system to control inputs, and 2) the relative priority of measured outputs. This is done using 1) a weighted Tracking Error Cost matrix (Q) and 2) a Control Effort Cost matrix (R). The weights are specified during the design of the controller. While this is convenient for achieving a fixed goal, it can be problematic for goals that change over time (e.g., minimizing power consumption before a predicted thermal emergency).

The controller must choose an appropriate trade-off when we cannot achieve both desirable performance and power concurrently. Unfortunately, classical MIMOs fix control weights at design time, and thus *cannot* perform *runtime* tradeoffs that re-

⁴ We interchangeably use the terms (*measured output* and *sensor*), as well as the terms (*control input* and *actuator*), as shown in Figure 7.

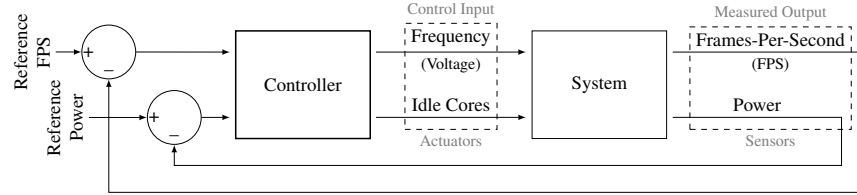


Fig. 7 Basic 2×2 MIMO for single-core system. Clock frequency and idle cores are used as control inputs. FPS and power are measured outputs that are compared with reference (i.e., target) values.

quire changing output priorities. Even with constant *reference values*, i.e., desired output values, unpredictable disturbances (e.g., changing workload and operating conditions) may cause the reference values to become unachievable. It is also plausible for the reference values themselves to change dynamically at runtime with system state and operating conditions (e.g., a thermal event).

Let us now consider a more complex scenario: a multi-threaded application running on Linux, executing on a mobile processor, where the system needs to track both the performance (FPS) and power simultaneously. Figure 7 shows the 2×2 MIMO model for this system with operating frequency and the number of active cores as control inputs, and FPS and power as measured outputs.

Both the FPS and power reference values are trackable individually, but not jointly. We implement and compare two different MIMO controllers in Linux to show the effect of competing objectives. One controller prioritizes FPS, and the other prioritizes power. Figure 8 shows the power and performance (in FPS) achieved by each MIMO controller using typical reference values for a mobile device: 60 FPS and 5 Watts. The application is $\times 264$, and the mobile processor consists of an ARM Cortex-A15 quad-core cluster. Each MIMO controller is designed with a different Q matrix to prioritize either FPS or power: Figure 8a's controller favors FPS over power by a ratio of 30:1 (i.e., only 1% deviation from the FPS reference is acceptable for a 30% deviation from the power reference), while Figure 8b uses a ratio of 1:30. We observe that neither controller is able to manage changing system goals. Thus, there is a need for a supervisor to autonomously orchestrate the system while considering the significance of competing objectives, user requirements, and operating conditions.

The use of supervisory control presents at least three additional advantages over conventional controllers. First, fully-distributed MIMO or SISO controllers *cannot* address system-wide goals such as power capping. Second, conventional controllers *cannot* model actuation effects that require system-wide perspective, such as task migration. Third, classical control theory *cannot* address problems requiring optimization (e.g., minimizing an objective function) alone [47, 69].

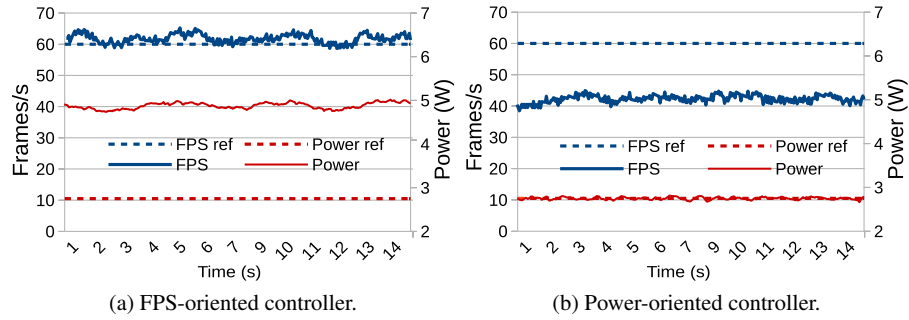


Fig. 8 $\times 264$ running on a quad-core cluster controlled by 2×2 MIMOs with different output priorities (from [77]).

3.1.2 Supervisory Control Theory

Supervisory control utilizes modular decomposition to mitigate the complexity of control problems, enabling automatic control of many individual controllers or control loops. Supervisory control theory (SCT) [78] benefits from formal synthesis methods to define principal control properties for *controllability* and *observability*. The emphasis on formal methods in addition to *modularity* leads to *hierarchical consistency* and *non-conflicting* properties.

SCT solves complex synthesis problems by breaking them into small-scale sub-problems, known as modular synthesis. The results of modular synthesis characterize the conditions under which decomposition is effective. In particular, results identify whether a valid decomposition exists. A decomposition is valid if the solutions to sub-problems combine to solve the original problem, and the resulting composite supervisors are *non-blocking* and *minimally restrictive*. Decomposition also adds robustness to the design because nonlinearities in the supervisor do *not* directly affect the system dynamics.

Figure 9 illustrates how a supervisory control structure can hierarchically manage control loops. As shown in the figure, supervision is vertically decomposed into tasks performed at different levels of abstraction [92]. The supervisory controller is designed to control the high-level *system model*, which represents an abstraction of the system. The *subsystems* compose the pre-existing *system* that does *not* meet the given specifications without the aid of a controller or a supervisor. The *information* channel provides information about the updates in the high-level model to the supervisory controller. Due to the fact that the system model is an abstract model, the controlling channel is an indirect *virtual control* channel. In other words, the control decisions of the supervisory controller will be implemented by controlling the *low-level controller(s)* through *control parameters*. Consequently, the low-level controller(s) can control one or multiple subsystems using the *control* channel and gather information via *feedback*. The changes in the subsystems can trigger *model*

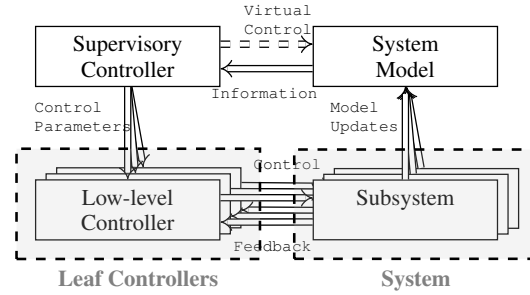


Fig. 9 Supervisory Control structure. Low level control loops are guided by the Supervisory Controller that achieves system-wide goals based on the high-level system model.

updates in the state of the high-level system model. These updates reflect the results of low-level controllers' controlling actions.

The scheme of Figure 9 describes the division of supervision into high-level management and low-level operational supervision. Virtual control exercised via the high-level control channel can be implemented by modifying control parameters to adaptively coordinate the low-level controllers, e.g., by adjusting their objective functions according to the system goal. The combination of horizontal and vertical decomposition enables us to not only physical divide the system into subsystems, but also to logically divide the sub-problems in any appropriate way, e.g., due to varying epochs (control invocation period) or scope. The important requirement of this hierarchical control scheme is control consistency and hierarchical consistency between the high-level model and the low-level system, as defined in the standard Ramadge-Wonham control mechanism [92]. For a detailed description of SCT, we refer the reader to [78, 79, 5, 92].

3.1.3 Self-Adaptivity via Supervisory Control

Supervisory controllers are preferable to *adaptive (self-tuning) controllers* for complex system control due to their ability to integrate **logic** with **continuous dynamics**. Specifically, supervisory control has two key properties: i) rapid adaptation in response to abrupt changes in management policy [35], and ii) low computational complexity by computing control parameters for different policies **offline**. New policies and their corresponding parameters can be added to the supervisor on demand (e.g., by upgrading the firmware or OS), rendering online learning-based self-tuning methods, e.g., least-squares estimation [3], unnecessary.

Figure 10 depicts the two mechanisms that enable SCT-based management via low-level controllers: **gain scheduling** and **dynamic references**. Gain scheduling is a nonlinear control technique that uses a set of linear controllers predesigned for different operating regions. Gain scheduling enables the appropriate linear con-

troller based on runtime observations [49]. Scheduling is implemented by switching between sets of control parameters, i.e., $A_1 \rightarrow A_2$, $B_1 \rightarrow B_2$, $C_1 \rightarrow C_2$, and $D_1 \rightarrow D_2$ in Equations 2 and 3. In this case, the *controller gains* are the values of the control parameters A , B , C , and D . Gains are useful to change objectives at runtime in response to abrupt and sudden changes in management policy. In LQG controllers, this is done by changing priorities of outputs using the Q and R matrices (Section 3.1.1). This is what we call the Hierarchical Control structure, in which local controllers solve specified tasks while the higher-level supervisory controller coordinates the global objective function. In this structure, the supervisory controller receives information from the plant (e.g., the presence of a thermal emergency) or the user/application (e.g., new QoS reference value), and steers the system towards the desired policy using its design logic and high-level model. Thanks to its top-level perspective, the supervisor can update reference values for each low-level controller to either optimize for a certain goal (e.g., getting to the optimum energy-efficient point) or manage resource allocation (e.g., allocating power budget to different cores).

3.2 Case Study: On-chip Resource Management

In this section, we design and evaluate a supervisor used to implement a hierarchical resource manager. The use-case requires management of QoS under a power budget on a HMP. The resource manager (SPECTR) consists of a supervisor that guides low-level classical controllers to configure core operating frequency and number of active cores for each core cluster.

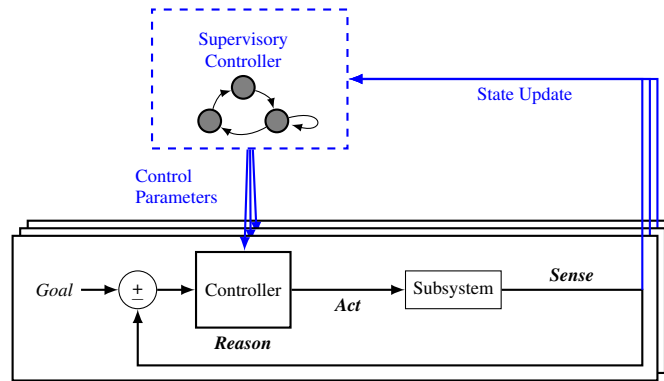


Fig. 10 Self-adaptivity via gain scheduling in SCT.

3.2.1 Hierarchical System Architecture

Figure 11 depicts a high-level view of SPECTR for many-core system resource management. Either the user or the system software may specify *Variable Goals and Policies*. The *Supervisory Controller* aims to meet system goals by managing the low-level controllers. High-level decisions are made based on the feedback given by the *High-level Plant Model*, which provides an abstraction of the entire system. Various types of *Classic Controllers*, such as PID or state-space controllers, can be used to implement each low-level controller based on the target of each subsystem. The flexibility to incorporate any pre-verified off-the-shelf controllers without the need for system-wide verification is essential for the modularity of this approach. The supervisor provides parameters such as output references or gain values to each low-level controller during runtime according to the system policy. Low-level controller subsystems update the high-level model to maintain global system state, and potentially trigger the supervisory controller to take action. The high-level model can be designed in various fashions (e.g., rule-based or estimator-based [79][35][63]) to track the system state and provide the supervisor with guidelines. We illustrate the steps for designing a supervisory controller using the following experimental case study in which SCT is deployed on a real HMP platform, and we then outline the entire design flow from modeling of the high-level plant to generating the supervisory controller.

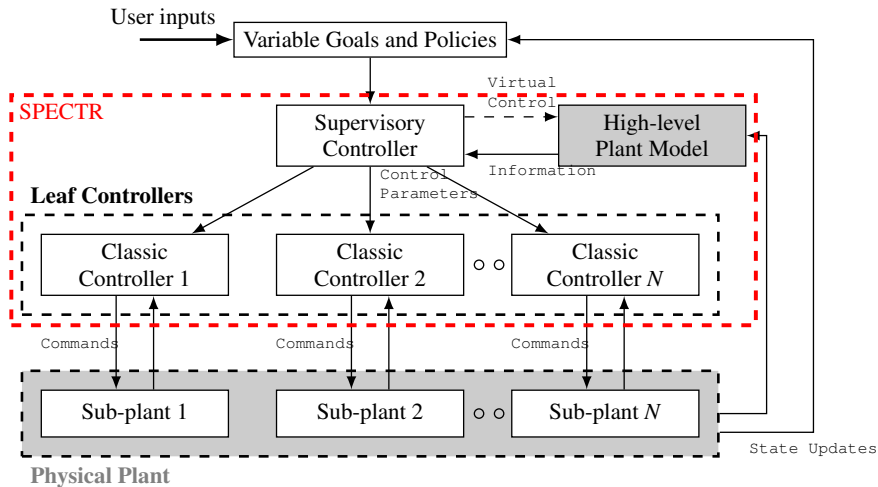


Fig. 11 SPECTR overview.

3.2.2 SPECTR Resource Manager

Figure 12 shows an overview of our experimental setup. We target the Exynos platform [32], which contains an HMP with two quad-core clusters: the **Big** core cluster provides high-performance out-of-order cores, while the **Little** core cluster provides low-power in-order cores. Memory is shared across all cores, so application threads can transparently execute on any core in any cluster. We consider a typical mobile scenario in which a single foreground application (the *QoS application*) is running concurrently with many background applications (the *Non-QoS applications*). This mimics a typical mobile use-case in which gaming or media processing is performed in the foreground in conjunction with background email or social media syncs.

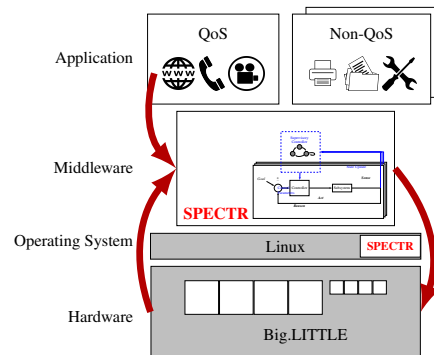


Fig. 12 SPECTR implementation on the Exynos HMP with two heterogeneous quad-core clusters. Representing a typical mobile scenario with a single foreground application running concurrently with many background applications.

The system goals are twofold: i) meet the QoS requirement of the foreground application while minimizing its energy consumption; and ii) ensure the total system power always remains below the Thermal Design Power (TDP).

The subsystems are the two heterogeneous quad-core (*Big* and *Little*) clusters. Each cluster has two actuators: one actuator to set the operating frequency (F_{next}) and associated voltage of the cluster; and one to set the number of active cores (AC_{next}) on the cluster. We measure the power consumption (P_{curr}) of each cluster, and simultaneously monitor the QoS performance (QoS_{curr}) of the designated application to compare it to the required QoS (QoS_{ref}).⁵

Supervisory control commands guide the **low-level MIMO controllers** in Figure 12 to determine the number of active cores and the core operating frequency within each cluster.

Supervisory control minimizes the system-wide power consumption while maintaining QoS. In our scenario, the QoS application runs only on the Big cluster, and

⁵ The Exynos platform provides only per-cluster power sensors and DVFS; hence our use of cluster-level sensors and actuators.

the supervisor determines whether and how to adjust the cluster’s power budget based on QoS measurements.

Gain scheduling is used to switch the priority objective of the low-level controllers. We define two sets of gains for this case-study: 1) *QoS-based* gains are tuned to ensure that the QoS application can meet the performance reference value, and 2) *Power-based* gains are tuned to limit the power consumption while possibly sacrificing some performance if the system is exceeding the power budget threshold.

3.2.3 Experimental Evaluation

We compare SPECTR with three alternative resource managers. The first two managers use two uncoordinated 2×2 MIMOs, one for each cluster: *MM-Pow* uses power-oriented gains, and *MM-Perf* uses performance-oriented gains. These fixed MIMO controllers act as representatives of a state-of-the-art solution, as presented in [69], one prioritizing power and the other prioritizing performance. The third manager consists of a single full-system controller (*FS*): a system-wide 4×2 MIMO with individual control inputs for each cluster. *FS* uses power-oriented gains and its measured outputs are chip power and QoS. This single system-wide MIMO acts as a representative for [103], maximizing performance under a power cap.

We analyze an execution scenario that consists of three different phases of execution:

1. *Safe Phase*: In this phase, only the QoS application executes (with an achievable QoS reference within the TDP). The goal is to meet QoS and minimize power consumption.
2. *Emergency Phase*: In this phase, the QoS reference remains the same as that in the Safe Phase while the power envelope is reduced (emulating a thermal emergency). The goal is to adapt to the change in reference power while maintaining QoS (if possible).
3. *Workload Disturbance Phase*: In this phase, the power envelope returns to TDP and background tasks are added (to induce interference from other tasks). The goal is to meet the QoS reference value without exceeding the power envelope.

This execution scenario with three different phases allows us to evaluate how SPECTR compares with state-of-the-art resource managers when facing workload variation and system-wide changes in state (e.g., thermal emergency) and goals.

Evaluated resource manager configurations

We generate stable low-level controllers for each resource manager using the Matlab System Identification Toolbox [59].⁶ We use the Control Effort Cost matrix (R) to prioritize changing clock frequency over number of cores at a ratio of 2:1, as

⁶ We generate the models with a stability focus. All systems are stable according to Robust Stability Analysis. We use Uncertainty Guardbands of 50% for QoS and 30% for power, as in [69].

frequency is a finer-grained and lower-overhead actuator than core count. We generate training data by executing an in-house microbenchmark and varying control inputs in the format of a staircase test (i.e., a sine wave), both with single-input variation and all-input variation. The micro-benchmark consists of a sequence of independent multiply-accumulate operations performed over both sequentially and randomly accessed memory locations, thus yielding various levels of instruction-level and memory-level parallelism. The range of exercised behavior resembles or exceeds the variation we expect to see in typical mobile workloads, which is the target application domain of our case studies.

Experimental setup

We perform our evaluations on the ARM big.LITTLE [2] based Exynos SoC (ODROID-XU3 board [32]) as described in our case study (Figure 12). We implement a Linux userspace daemon process that invokes the low-level controllers every $50ms$. When evaluating SPECTR, the daemon invokes the supervisor every $100ms$. We use ARM's Performance Monitor Unit (PMU) and per-cluster power sensors for the performance and power measurements required by the resource managers. The userspace daemon also implements the Heartbeats API [39] monitor to measure QoS. By periodically issuing *heartbeats*, the application informs the system about its current performance. The user provides a performance reference value using the Heartbeats API.

To evaluate the resource managers, we use the following benchmarks from the PARSEC benchmark suite [6] as QoS applications (i.e., the applications that issue heartbeats to the controller): `x264`, `bodytrack`, `canneal`, and `streamcluster`. The selected applications consist of the most CPU-bound along with the most cache-bound PARSEC benchmarks, providing varied responses to change in resource allocation. Speedups from $3.2X$ (`streamcluster`) to $4.5X$ (`x264`) are observed with the maximum resource allocation values compared to the minimum. We also use one of four machine-learning workloads as our QoS application: `k-means`, `KNN`, `least squares`, and `linear regression`. These four workloads provide a wide range of data-intensive use cases. For all experiments, each QoS application uses four threads. The background (non-QoS) tasks used in the third execution phase are single-threaded microbenchmarks, and have no runtime restrictions, i.e., the Linux scheduler can freely migrate them between and within clusters.

3.2.4 Effectiveness of Self-Adaptivity through Supervision

We focus our discussion on the `x264` benchmark results. Other results are summarized at the end of this section. We use heartbeats to measure the frames per second (FPS) as our QoS metric. Figure 13 shows the measured FPS and power for `x264` with respect to their reference values over the course of execution for all of the resource management controllers.

×264 Benchmark

To show the energy efficiency of SPECTR, we study the Safe Phase. The Safe Phase consists of the first 5 seconds of execution during which only the QoS application executes on the Big cluster. In this phase, all controllers are able to achieve the FPS reference value within the power envelope. Figures 14a and 14b show the average steady-state error (%) of QoS and power respectively for each resource manager in Phase 1. Steady-state error is used to define *accuracy* in feedback control systems [33]. Steady-state error values are calculated as $reference - measured\ output$. Negative values indicate that the power/QoS **exceeds** the reference value, positive values indicate power savings or failure to meet QoS. We make two key observations. First, both MM-Perf and SPECTR reduce power consumption by 25% (Fig. 14b) while maintaining FPS within 10% (Fig. 14a) of the reference value. The MM-Perf controller operates efficiently because the reference FPS value is achievable within the TDP threshold. The SPECTR controller similarly operates efficiently: it is able to recognize that the FPS is achievable within TDP and, as a result, lower the reference power. Second, the FS and MM-Pow controllers unnecessarily exceed the reference FPS value and, as a result, consume excessive power. This is because these controllers prioritize meeting the power reference value, consuming the entire available power budget to maximize performance.

To show SPECTR's ability to adapt to a sudden change in operating constraints, we study the Emergency Phase. The Emergency Phase of execution emulates a thermal emergency, during which, the TDP is lowered to ensure that the system operates in a safe state. This occurs during the second 5-second period of execution in Figure 13. We observe that all controllers are able to react to the change in power reference value and maintain QoS. However, compared to the other controllers, FS has a sluggish reaction (Figure 13f) to the change in power reference, despite the fact that it is designed to prioritize tracking the power output. *Settling time* is a property used to quantify responsiveness of feedback control systems [33]. *Settling time* is the time it takes to reach sufficiently close to the steady-state value after the reference values are set. The average settling time for the power output of FS is 2.07 seconds, while SPECTR has an average settling time of 1.28 seconds. The larger size of the state-space ($x(t)$ matrix in Equation 2 and 3) and the higher number of control inputs in the 4×2 FS compared to those of 2×2 controllers in SPECTR is the reason for the slow settling time of FS. This is also the reason why SISO controllers are generally faster than MIMOs [33].

To show SPECTR's ability to adapt to workload disturbance and changing system goals, we study the Workload Disturbance Phase. The Workload Disturbance Phase occurs in seconds 10-15 of execution in Figure 13. In this phase, 1) the QoS reference value and the power envelope return to the same values as in Phase 1, and 2) we introduce disturbance in the form of background tasks. As a result of the workload disturbance, the QoS reference is *not* achievable within the TDP. We make two observations regarding the steady-state error in Figures 14e and 14f. First, SPECTR behaves similarly to MM-Pow, even though in Phase 1 it behaved similarly to MM-Perf. The SPECTR supervisor is able to recognize the change in execution scenario

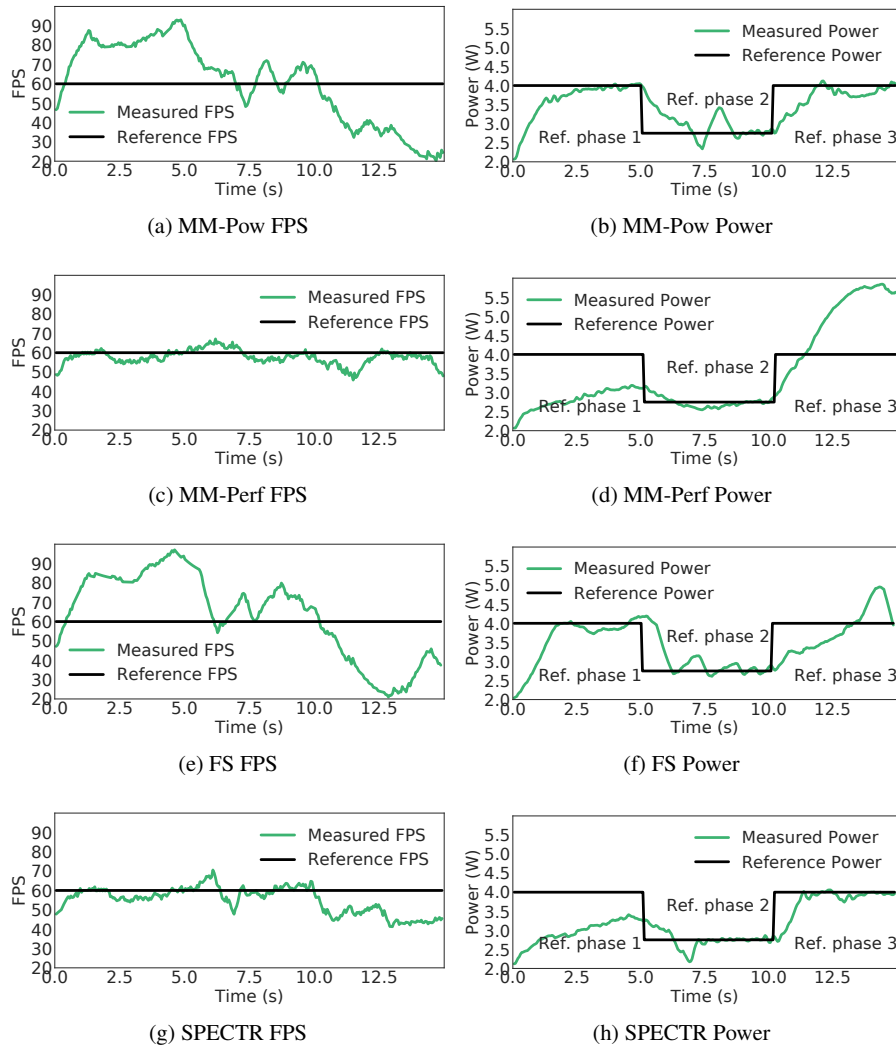
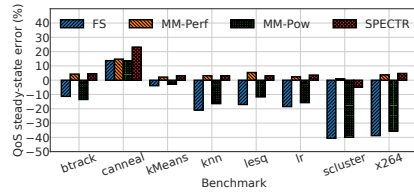
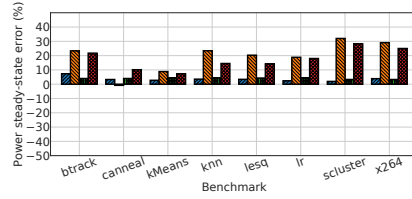


Fig. 13 Measured FPS and Power of all four resource managers for three Phases of 5 seconds each, for the $x264$ benchmark (from [77]).

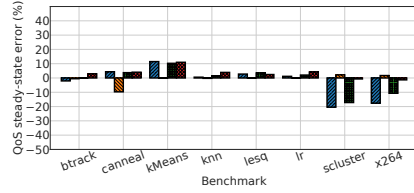
and constraints, and adapt its priorities appropriately. In this case, SPECTR achieves much higher FPS than all controllers except MM-Perf (Fig. 14e), while obeying the TDP limit (Fig. 14f). Second, both FS and MM-Pow operate at the TDP limit, but achieve a significantly lower FPS than the reference value. MM-Perf comes within $\sim 5\%$ of the reference FPS (Fig. 14e) while exceeding the TDP by more than 30% (Fig. 14f), which is undesirable.



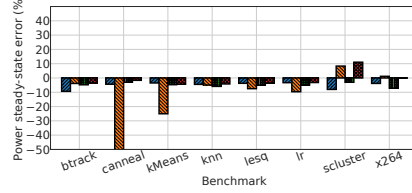
(a) QoS steady-state error in Phase 1.



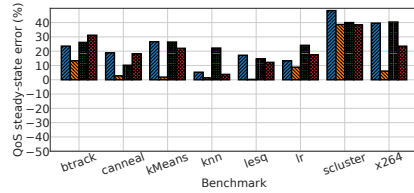
(b) Power steady-state error in Phase 1.



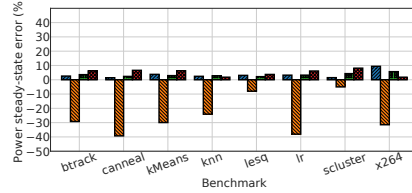
(c) QoS steady-state error in Phase 2.



(d) Power steady-state error in Phase 2.



(e) QoS steady-state error in Phase 3.



(f) Power steady-state error in Phase 3.

Fig. 14 Steady-state error for all benchmarks, grouped by phase. A negative value indicates the amount of power/QoS **exceeding** the reference value (bad), a positive value indicates the amount of power saved (good) or QoS degradation (bad) (from [77]).

Other Benchmarks

We perform the same experiments for PARSEC benchmarks `bodytrack`, `canneal`, `streamcluster`, as well as machine-learning benchmarks `k-means`, `KNN`, `least squares`, and `linear regression`. For these workloads, we use the generic *heartbeat rate* (HB) directly as the QoS metric, as FPS is not an appropriate metric. Figures 14a, 14c, and 14e show the average steady-state error (%) of QoS for Phases 1, 2, and 3 respectively. Figures 14b, 14d, and 14f show the average steady-state error (%) of power for Phases 1, 2, and 3 respectively. We summarize the observations for the additional experiments with respect to `x264` for the three phases. In the Safe Phase, the behavior of `bodytrack`, `streamcluster`, `k-means`, `KNN`, `least squares`, and `linear regression` is similar to that of `x264` (Figures 14a and 14b). `canneal` follows the same pattern with respect to power as all other benchmarks (Fig. 14b). `canneal`'s QoS steady-state error is the only difference in behavior we observe in Phase 1. None of the managers are able to meet

the QoS reference value for `canneal` in Phase 1 (Fig. 14a). This is due to the fact that the phase of `canneal` captured in the experiment primarily consists of serialized input processing, so the number of idle cores has reduced affect on QoS. In the Emergency Phase, our observations from `x264` hold for nearly all benchmarks regarding response to change in power reference value, achieving less than 10% power steady-state error (Fig. 14d). The only exceptions are `canneal` and `k-means`: the MM-Perf manager is unable to react to change in TDP for `canneal` and `k-means`. The MM-Perf manager lacks a supervisory coordinator and prioritizes performance, and was unable to find a configuration for `canneal` and `k-means` that satisfied the QoS reference value within TDP. In the Workload Disturbance Phase, SPECTR, FS, and MM-Pow all achieve near-reference power (Fig. 14f). As expected, MM-Perf violates the TDP in all cases, but always achieves the highest QoS (Fig. 14e).

We conclude that SPECTR is effective at (1) efficiently meeting multiple system objectives when it is possible to do so, (2) appropriately balancing multiple conflicting objectives, and (3) quickly responding to sudden and unpredictable changes in constraints due to workload or system state.

3.2.5 Overhead Evaluation

To show the overhead of the low-level MIMO controllers, we study their execution time. We measure the MIMO controller execution time to be $2.5ms$, on average, over 30 seconds. The MIMO controller is invoked every $50ms$ resulting in a 5% overhead, which is experienced by all evaluated controllers. We measure the runtime of the supervisor to be $30\mu s$, which is negligible even with respect to the MIMO controller execution time. The supervisor is invoked less frequently than the MIMO controllers ($2\times$ the period in our case), executes in parallel to the workload and MIMO controllers, and simply evaluates the system state in order to determine if the MIMO controller gains need changing. State changes that result in interventions on the low-level controllers occur only due to system-wide changes in the state (e.g., thermal emergency) or goals (e.g., change in performance reference value or execution mode), which are infrequent. When the supervisor needs to change the MIMO gains, it simply points the coefficient matrices to a different set of stored values. In our case study, we have two sets of gains (QoS and power oriented) that are generated when the controllers are designed and stored during system initialization. Changing the coefficient arrays at runtime takes effect immediately, and has no additional overhead.

To show the overhead of SPECTR's supervisory controller, we compare the total execution time of identical workloads with and without SPECTR. With respect to the preemption overhead due to globally managing resources, Linux's HMP scheduler typically maps SCT threads to a core on the low-power Little cluster. Therefore, the SCT threads are executed without preempting the QoS application, which always executes on the Big cluster. We verify the overall impact of the control system overhead by running the benchmarks on two different systems: i) a vanilla Linux

setup⁷ and ii) vanilla Linux with SPECTR running in the background. For (ii), SPECTR controllers perform all the required computations but do *not* change the system knobs (thus only the SPECTR overhead affects the system). When comparing the QoS of the applications across multiple runs, we verify a negligible average difference of 0.1% between the two systems.

We conclude that the benefits of SPECTR come at a negligible performance overhead.

3.3 Summary

Modern mobile systems require intelligent management to balance user demands and system constraints. At any given time, the relative priority of demands and constraints may change based on uncontrollable context, such as dynamic workload or operating condition. A resource manager must be able to autonomously detect such context changes and adapt appropriately. This property is known as self-adaptivity. We demonstrate one way to design a self-adaptive resource manager: using supervisory control theory. Supervisory control theory lends itself well to this challenge due to its high level of abstraction and lightweight implementation. The proposed supervisor successfully adapts to changes when managing quality of service under a power budget for chip multiprocessors. The hierarchy using supervisory control theory represents early exploration of self-adaptivity in the resource management domain, and a slight degree of self-awareness. This approach can be enhanced in one way through the definition and generation of goals. Initial work based on goal-driven autonomy has been done toward this end [80].

4 Conclusion

We use two forms of computational self-awareness to implement resource managers with simple but effective self-aware components. Systems can be self-aware to varying degrees, and the degree to which self-*X* properties are utilized is case-specific. We demonstrate the use of self-optimization in implementing a DVFS governor for managing power in a processor core. We demonstrate the use of self-adaptivity in implementing a multi-goal resource manager for managing QoS within a power budget in an HMP. Moving forward, as systems scale and configuration spaces grow, computational self-awareness provides a useful abstract tool for tackling various challenges in resource management.

Acknowledgements This work was partially supported by NSF grant CCF-1704859.

⁷ Ubuntu 16.04.2 LTS and Linux kernel 3.10.105 (<https://dn.odroid.com/5422/ODROID-XU3/Ubuntu/>).

References

- [1] Annamalai A, Rodrigues R, Koren I, Kundu S (2013) An opportunistic prediction-based thread scheduling to maximize throughput/watt in AMPs. In: Parallel Architectures and Compilation Techniques - Conference Proceedings, PACT, IEEE, pp 63–72, DOI 10.1109/PACT.2013.6618804
- [2] ARM (2013) big.LITTLE Technology: The Future of Mobile. Tech. rep., URL https://www.arm.com/files/pdf/big_LITTLE_Technology_the_Futue_of_Mobile.pdf
- [3] Astrom KJ, Wittenmark B (1995) Adaptive Control. Addison-Wesley
- [4] Bartolini A, Cacciari M, Tilli A, Benini L (2011) A distributed and self-calibrating model-predictive controller for energy and thermal management of high-performance multicores. In: DATE
- [5] Bertil A Brandin MW, Benhabib B (1991) Discrete Event System Supervisory Control Applied to the Management of Manufacturing Workcells. In: Computer-Aided Production Engineering, C. Venkatesh and J.A. McGeough, eds. (Amsterdam: Elsevier)
- [6] Bienia C (2011) Benchmarking modern multiprocessors. PhD thesis, Princeton University
- [7] Bitirgen R, Ipek E, Martinez JF (2008) Coordinated Management of Multiple Interacting Resources in Chip Multiprocessors: A Machine Learning Approach. In: MICRO
- [8] Chang KK, Yağlıkçı AG, Ghose S, Agrawal A, Chatterjee N, Kashyap A, Lee D, O'Connor M, Hassan H, Mutlu O (2017) Understanding Reduced-Voltage Operation in Modern DRAM Devices: Experimental Characterization, Analysis, and Mechanisms. Proc ACM Meas Anal Comput Syst
- [9] Choi S, Yeung D (2006) Learning-Based SMT Processor Resource Distribution via Hill-Climbing. In: ISCA
- [10] Cochran R, Hankendi C, Coskun AK, Reda S (2011) Pack & Cap: Adaptive DVFS and Thread Packing Under Power Caps. In: MICRO
- [11] Das R, Mutlu O, Moscibroda T, Das CR (2009) Application-aware prioritization mechanisms for on-chip networks. In: MICRO
- [12] Das R, Mutlu O, Moscibroda T, Das CR (2010) AéRgia: Exploiting Packet Latency Slack in On-chip Networks. In: ISCA
- [13] Das R, Ausavarungnirun R, Mutlu O, Kumar A, Azimi M (2013) Application-to-core mapping policies to reduce memory system interference in multi-core systems. In: HPCA
- [14] David H, Fallin C, Gorbatov E, Hanebutte UR, Mutlu O (2011) Memory power management via dynamic voltage/frequency scaling. In: ICAC
- [15] Delimitrou C, Kozyrakis C (2014) Quasar: Resource-efficient and QoS-aware Cluster Management. In: ASPLOS
- [16] Deng Q, Meisner D, Bhattacharjee A, Wenisch TF, Bianchini R (2012) CoScale: Coordinating CPU and Memory System DVFS in Server Systems. In: MICRO

- [17] Dhodapkar AS, Smith JE (2002) Managing Multi-configuration Hardware via Dynamic Working Set Analysis. In: ISCA
- [18] Donyanavard B, Mück T, Sarma S, Dutt N (2016) SPARTA: Runtime Task Allocation for Energy Efficient Heterogeneous Many-cores. In: Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis - CODES '16, ACM Press, New York, New York, USA, DOI 10.1145/2968456.2968459
- [19] Donyanavard B, Mück T, Sarma S, Dutt N (2016) Sparta: Runtime task allocation for energy efficient heterogeneous many-cores. In: CODES+ISSS
- [20] Donyanavard B, Rahmani AM, Mück T, Moazemmi K, Dutt ND (2018) Gain scheduled control for nonlinear power management in cmps. In: 2018 Design, Automation & Test in Europe Conference & Exhibition, DATE 2018, Dresden, Germany, March 19-23, 2018, DOI 10.23919/DATE.2018.8342141
- [21] Dubach C, Jones TM, Bonilla EV, O'Boyle MFP (2010) A Predictive Model for Dynamic Microarchitectural Adaptivity Control. In: MICRO
- [22] Dubach C, Jones TM, Bonilla EV (2013) Dynamic Microarchitectural Adaptation Using Machine Learning. In: TACO
- [23] Ebrahimi E, Mutlu O, Lee CJ, Patt YN (2009) Coordinated Control of Multiple Prefetchers in Multi-core Systems. In: MICRO
- [24] Ebrahimi E, Lee CJ, Mutlu O, Patt YN (2010) Fairness via source throttling: A configurable and high-performance fairness substrate for multi-core memory systems. In: ASPLOS
- [25] Ebrahimi E, Lee CJ, Mutlu O, Patt YN (2011) Prefetch-aware shared-resource management for multi-core systems. In: ISCA
- [26] Fan S, Zahedi SM, Lee BC (2016) The Computational Sprinting Game. In: ASPLOS
- [27] Fu X, Kabir K, Wang X (2011) Cache-Aware Utilization Control for Energy Efficiency in Multi-Core Real-Time Systems. In: ECRTS
- [28] Gupta U, Campbell J, Ogras UY, Ayoub R, Kishinevsky M, Paterna F, Gummusoy S (2016) Adaptive performance prediction for integrated GPUs. In: ICCAD
- [29] Gupta U, Ayoub R, Kishinevsky M, Kadjo D, Soundararajan N, Tursun U, Ogras U (2017) Dynamic Power Budgeting for Mobile Systems Running Graphics Workloads. In: TMSCS
- [30] Haghbayan MH, Miele A, Rahmani AM, Liljeberg P, Tenhunen H (2017) Performance/Reliability-Aware Resource Management for Many-Cores in Dark Silicon Era. IEEE Transactions on Computers
- [31] Hanumaiah V, Desai D, Gaudette B, Wu CJ, Vrudhula S (2014) STEAM: A Smart Temperature and Energy Aware Multicore Controller. In: TECS
- [32] Hardkernel (2016) ODROID-XU. Tech. rep., URL <http://www.hardkernel.com/main/main.php>
- [33] Hellerstein JL, Diao Y, Parekh S, Tilbury DM (2004) Feedback Control of Computing Systems. John Wiley & Sons
- [34] Herbert S, Marculescu D (2007) Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In: ISLPED, DOI 10.1145/1283780.1283790

- [35] Hespanha JP (2011) Tutorial on supervisory control. In: Lecture Notes for the workshop *Control using Logic and Switching* for the 40th Conference on Decision and Control
- [36] Hoffmann H (2014) Coadapt: Predictable behavior for accuracy-aware applications running on power-aware systems. In: ECRTS
- [37] Hoffmann H, Sidiroglou S, Carbin M, Misailovic S, Agarwal A, Rinard M (2011) Dynamic Knobs for Responsive Power-aware Computing. In: ASPLOS
- [38] Hoffmann H, Sidiroglou S, Carbin M, Misailovic S, Agarwal A, Rinard M (2011) Dynamic knobs for responsive power-aware computing. In: ASPLOS, DOI 10.1145/1950365.1950390
- [39] Hoffmann H, Maggio M, Santambrogio MD, Leva A, Agarwal A (2013) A generalized software framework for accurate and efficient management of performance goals. In: EMSOFT
- [40] Ipek E, Mutlu O, Martínez JF, Caruana R (2008) Self-Optimizing Memory Controllers: A Reinforcement Learning Approach. In: ISCA
- [41] Isci C, Buyuktosunoglu A, Cher CY, Bose P, Martonosi M (2006) An Analysis of Efficient Multi-Core Global Power Management Policies: Maximizing Performance for a Given Power Budget. In: MICRO
- [42] Jantsch A, Dutt N, Rahmani AM (2017) Self-awareness in systems on chip—a survey. *IEEE Design & Test* 34(6)
- [43] Juang P, Wu Q, Peh LS, Martonosi M, Clark DW (2005) Coordinated, distributed, formal energy management of chip multiprocessors. In: ISLPED
- [44] Jung H, Rong P, Pedram M (2008) Stochastic modeling of a thermally-managed multi-core system. In: DAC
- [45] Kadjo D, Ayoub R, Kishinevsky M, Gratz PV (2015) A Control-theoretic Approach for Energy Efficient CPU-GPU Subsystem in Mobile Platforms. In: DAC
- [46] Kanduri A, Haghbayan MH, Rahmani AM, Liljeberg P, Jantsch A, Dutt N, Tenhunen H (2016) Approximation knob: Power Capping meets energy efficiency. In: ICCAD
- [47] Karamanolis C, Karlsson M, Zhu X (2005) Designing Controllable Computer Systems. In: HoTOS
- [48] Lee CJ, Narasiman V, Ebrahimi E, Mutlu O, Patt YN (2010) DRAM-aware last-level cache writeback: Reducing write-caused interference in memory systems. Tech. rep., UT Austin
- [49] Leith D, Leithead W (2000) Survey of gain-scheduling analysis and design. In: *International Journal of Control*
- [50] Lewis PR, Platzner M, Rinner B, Tørresen J, Yao X (2016) *Self-Aware Computing Systems*. Springer
- [51] Liu G, Park J, Marculescu D (2013) Dynamic thread mapping for high-performance, power-efficient heterogeneous many-core systems. In: ICCD, IEEE, DOI 10.1109/ICCD.2013.6657025
- [52] Ljung L (1999) *System Identification: Theory for the User*. Prentice Hall PTR

- [53] Ljung L (2001) Black-box models from input-output measurements. In: I2MTC
- [54] Lo D, Song T, Suh GE (2015) Prediction-guided Performance-energy Trade-off for Interactive Applications. In: MICRO
- [55] Ma K, Li X, Chen M, Wang X (2011) Scalable power control for many-core architectures running multi-threaded applications. In: ISCA
- [56] Ma K, Li X, Chen M, Wang X (2011) Scalable power control for many-core architectures running multi-threaded applications. In: ACM SIGARCH Comp. Arc. News
- [57] Maggio M, Hoffmann H, Santambrogio MD, Agarwal A, Leva A (2010) Controlling software applications via resource allocation within the heartbeats framework. In: CDC
- [58] Mahajan D, Yazdanbakhsh A, Park J, Thwaites B, Esmailzadeh H (2016) Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration. In: ISCA
- [59] MathWorks (2017) System Identification Toolbox. Tech. rep., URL <https://www.mathworks.com/products/sysid.html>
- [60] Mishra AK, Srikantaiah S, Kandemir M, Das CR (2010) CPM in CMPs: Coordinated Power Management in Chip-Multiprocessors. In: SC
- [61] Mishra AK, Srikantaiah S, Kandemir M, Das CR (2010) Cpm in cmps: Coordinated power management in chip-multiprocessors. In: SC, DOI 10.1109/SC.2010.15
- [62] Mishra N, Imes C, Lafferty JD, Hoffmann H (2018) CALOREE: Learning Control for Predictable Latency and Low Energy. In: Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS '18, ACM Press, New York, New York, USA, pp 184–198, DOI 10.1145/3173162.3173184
- [63] Morse S (1997) Control using logic-based switching, Springer
- [64] Mück T, Sarma S, Dutt N (2015) Run-DMC: Runtime dynamic heterogeneous multicore performance and power estimation for energy efficiency. In: 2015 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS), IEEE, DOI 10.1109/CODESISSS.2015.7331380
- [65] Muthukaruppan TS, Pricopi M, Venkataramani V, Mitra T, Vishin S (2013) Hierarchical Power Management for Asymmetric Multi-core in Dark Silicon Era. In: DAC
- [66] Muthukaruppan TS, Pricopi M, Venkataramani V, Mitra T, Vishin S (2013) Hierarchical power management for asymmetric multi-core in dark silicon era. In: DAC
- [67] Muthukaruppan TS, Pathania A, Mitra T (2014) Price Theory Based Power Management for Heterogeneous Multi-cores. In: ASPLOS
- [68] Petrica P, Izraelevitz AM, Albonesei DH, Shoemaker CA (2013) Flicker: A Dynamically Adaptive Architecture for Power Limited Multicore Systems. In: ISCA

- [69] Pothukuchi RP, Ansari A, Voulgaris P, Torrellas J (2016) Using Multiple Input, Multiple Output Formal Control to Maximize Resource Efficiency in Architectures. In: ISCA
- [70] Pothukuchi RP, Pothukuchi SY, Voulgaris P, Torrellas J (2018) Yukta: Multi-layer resource controllers to maximize efficiency. In: Proceedings of the 45th Annual International Symposium on Computer Architecture, IEEE Press, Piscataway, NJ, USA, ISCA '18, pp 505–518, DOI 10.1109/ISCA.2018.00049, URL <https://doi.org/10.1109/ISCA.2018.00049>
- [71] Pricopi M, Muthukaruppan TS, Venkataramani V, Mitra T, Vishin S (2013) Power-performance modeling on asymmetric multi-cores. In: 2013 International Conference on Compilers, Architecture and Synthesis for Embedded Systems (CASES), IEEE, DOI 10.1109/CASES.2013.6662519
- [72] Q Wu, P Juang, M Martonosi, D W Clark (2004) Formal Online Methods for Voltage/Frequency Control in Multiple Clock Domain Microprocessors. In: ASPLOS
- [73] Raghavendra R, Ranganathan P, Talwar V, Wang Z, Zhu X (2008) No "Power" Struggles: Coordinated Multi-level Power Management for the Data Center. In: ISCA
- [74] Rahmani AM, Haghbayan MH, Kanduri A, Weldezion AY, Liljeberg P, Plosila J, Jantsch A, Tenhunen H (2015) Dynamic power management for many-core platforms in the dark silicon era: A multi-objective control approach. In: ISLPED
- [75] Rahmani AM, Haghbayan MH, Miele A, Liljeberg P, Jantsch A, Tenhunen H (2017) Reliability-Aware Runtime Power Management for Many-Core Systems in the Dark Silicon Era. In: TVLSI
- [76] Rahmani AM, Jantsch A, Dutt N (2017) HDGM: Hierarchical Dynamic Goal Management for Many-Core Resource Allocation. In: ESL
- [77] Rahmani AM, Donyanavard B, Mück T, Moazzemi K, Jantsch A, Mutlu O, Dutt N (2018) SPECTR: Formal Supervisory Control and Coordination for Many-core Systems Resource Management. In: Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems - ASPLOS '18, ACM Press, New York, New York, USA, DOI 10.1145/3173162.3173199
- [78] Ramadge PJ, Wonham WM (1989) The control of discrete event systems. In: Proceedings of the IEEE
- [79] Safanov MH (1997) Focusing on the knowable: Controller invalidation and learning, Springer
- [80] Shamsa E, Kanduri A, Rahmani AM, Liljeberg P, Jantsch A, Dutt N (2019) Goal-driven autonomy for efficient on-chip resource management: Transforming objectives to goals. In: 2019 Design, Automation Test in Europe Conference Exhibition (DATE), pp 1397–1402, DOI 10.23919/DATE.2019.8715134
- [81] Singh K, Bhadauria M, McKee Sa (2009) Real time power estimation and thread scheduling via performance counters. ACM SIGARCH Computer Architecture News 37(2):46, DOI 10.1145/1577129.1577137

- [82] Skogestad S, Postlethwaite I (2005) *Multivariable Feedback Control: Analysis and Design*. John Wiley & Sons
- [83] Smith BC (1982) *Reflection and Semantics in a Procedural Programming Language*. Phd, MIT
- [84] Srikantaiah S, Kandemir M, Wang Q (2009) SHARP control: Controlled shared cache management in chip multiprocessors. In: MICRO
- [85] Stuecheli J, Kaseridis D, Daly D, Hunter HC, John LK (2010) The virtual write queue: Coordinating dram and last-level cache policies. In: ISCA
- [86] Su B, Gu J, Shen L, Huang W, Greathouse JL, Wang Z (2014) PPEP: Online Performance, Power, and Energy Prediction Framework and DVFS Space Exploration. In: MICRO
- [87] Subramanian L, Seshadri V, Kim Y, Jaiyen B, Mutlu O (2013) Mise: Providing performance predictability and improving fairness in shared main memory systems. In: HPCA
- [88] Subramanian L, Seshadri V, Ghosh A, Khan S, Mutlu O (2015) The Application Slowdown Model: Quantifying and Controlling the Impact of Inter-application Interference at Shared Caches and Main Memory. In: MICRO
- [89] Sui X, Lenharth A, Fussell DS, Pingali K (2016) Proactive Control of Approximate Programs. In: ASPLOS
- [90] Tembey P, Gavrilovska A, Schwan K (2012) A Case for Coordinated Resource Management in Heterogeneous Multicore Platforms. In: ISCA
- [91] Teodorescu R, Torrellas J (2008) Variation-Aware Application Scheduling and Power Management for Chip Multiprocessors. In: ISCA
- [92] Thistle J (1996) Supervisory control of discrete event systems. In: *Mathematical and Computer Modelling*
- [93] Vardhan V, Yuan W, Harris AF, Adve SV, Kravets R, Nahrstedt K, Sachs D, Jones D (2009) GRACE-2: integrating fine-grained application adaptation with global adaptation for saving energy. In: IJES
- [94] Vega A, Buyuktosunoglu A, Hanson H, Bose P, Ramani S (2013) Crank It Up or Dial It Down: Coordinated Multiprocessor Frequency and Folding Control. In: ISCA
- [95] Wang X, Martínez JF (2016) ReBudget: Trading Off Efficiency vs. Fairness in Market-Based Multicore Resource Allocation via Runtime Budget Reassignment. In: ASPLOS
- [96] Wang X, Ma K, Wang Y (2011) Adaptive Power Control with Online Model Estimation for Chip Multiprocessors. In: TPDS
- [97] Wang X, Ma K, Wang Y (2011) Adaptive power control with online model estimation for chip multiprocessors. *IEEE TPDS*
- [98] Wang Y, Ma K, Wang X (2009) Temperature-constrained Power Control for Chip Multiprocessors with Online Model Estimation. In: ISCA
- [99] Wu Q, Juang P, Martonosi M, Clark DW (2004) Formal online methods for voltage/frequency control in multiple clock domain microprocessors. In: *ACM SIGARCH Comp. Arc. News*
- [100] Wu Q, Juang P, Martonosi M, Peh LS, Clark DW (2005) Formal control techniques for power-performance management. In: *IEEE Micro*

- [101] Wu Q, Deng Q, Ganesh L, Hsu CH, Jin Y, Kumar S, Li B, Meza J, Song YJ (2016) Dynamo: Facebook's Data Center-Wide Power Management System. In: ISCA
- [102] Yan K, Zhang X, Tan J, Fu X (2016) Redefining QoS and customizing the power management policy to satisfy individual mobile users. In: MICRO
- [103] Zhang H, Hoffmann H (2016) Maximizing Performance Under a Power Cap: A Comparison of Hardware, Software, and Hybrid Techniques. In: ASPLOS