On-Chip Dynamic Resource Management

Other titles in Foundations and Trends® in Electronic Design Automation

Error-Efficient Computing Systems
Phillip Stanley-Marbell and Martin Rinard
ISBN: 978-1-68083-358-4

Secure Processors Part II: Intel SGX Security Analysis and MIT Sanctum Architecture
Victor Costan, Ilia Lebedev and Srinivas Devadas
ISBN: 978-1-68083-302-7

Secure Processors Part I: Background, Taxonomy for Secure Enclaves and Intel SGX Architecture
Victor Costan, Ilia Lebedev and Srinivas Devadas
ISBN: 978-1-68083-300-3

On-Chip Dynamic Resource Management

Antonio Miele
Politecnico di Milano, Italy

Anil Kanduri
University of Turku, Finland

Kasra Moazzemi
University of California, Irvine, USA

Dávid Juhász
TU Wien, Vienna, Austria
Imsys AB, Stockholm, Sweden

Amir M. Rahmani
University of California, Irvine, USA
TU Wien, Vienna, Austria

Nikil Dutt
University of California, Irvine, USA

Pasi Liljeberg
University of Turku, Finland

Axel Jantsch
TU Wien, Vienna, Austria

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Electronic Design Automation

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

A. Miele, A. Kanduri, K. Moazzemi, D. Juhász, A. M. Rahmani, N. Dutt, P. Liljeberg and A. Jantsch. *On-Chip Dynamic Resource Management*. Foundations and Trends® in Electronic Design Automation, vol. 13, no. 1-2, pp. 1–144, 2019.

ISBN: 978-1-68083-579-3

Foundations and Trends® in Electronic Design Automation
Volume 13, Issue 1-2, 2019

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Electronic Design Automation publishes survey and tutorial articles in the following topics:

- System Level Design
- Behavioral Synthesis
- Logic Design
- Verification
- Test
- Physical Design
- Circuit Level Design
- Reconfigurable Systems
- Analog Design
- Embedded software and parallel programming
- Multicore, GPU, FPGA, and heterogeneous systems
- Distributed, networked embedded systems
- Real-time and cyberphysical systems

Information for Librarians

Foundations and Trends® in Electronic Design Automation, 2019, Volume 13, 4 issues. ISSN paper version 1551-3939. ISSN online version 1551-3947. Also available as a combined paper and online subscription.

Contents

On-Chip Dynamic Resource Management

Antonio Miele[1], Anil Kanduri[2], Kasra Moazzemi[3], Dávid Juhász[4,5], Amir M. Rahmani[3,4], Nikil Dutt[3], Pasi Liljeberg[2] and Axel Jantsch[4]

[1] *Politecnico di Milano, Italy;*
[2] *University of Turku, Finland;*
[3] *University of California, Irvine, USA;*
[4] *TU Wien, Vienna, Austria;*
[5] *Imsys AB, Stockholm, Sweden;*

ABSTRACT

The need for dynamic resource management has shadowed the exponential growth of on-chip transistor capacity, and the challenge is accentuated by the heterogeneity of resources, and the bewildering variety of constraints and requirements of applications, platforms and users. The field has started with a few research papers in the early 1990s but has grown today to over hundred yearly publications, leading to an accumulated body of literature presumable far above 1000 papers.

We focus on the *dynamic (run-time)* management of *on-chip* resources and mostly ignore design-time techniques and off-chip resources of larger electronic systems. Moreover, we do not attempt a complete review of all published work on the topic. Rather, this survey provides a structured review and discussion of the state of the art and is divided along the primary objectives of resource management techniques: performance, power, reliability and quality of service, each

Antonio Miele, Anil Kanduri, Kasra Moazzemi, Dávid Juhász, Amir M. Rahmani, Nikil Dutt, Pasi Liljeberg and Axel Jantsch (2019), "On-Chip Dynamic Resource Management", Foundations and Trends® in Electronic Design Automation: Vol. 13, No. 1-2, pp 1–144. DOI: 10.1561/1000000055.

of which has its dedicated chapter. We observe that many works describe dedicated techniques for point problems, like minimizing power or maximizing lifetime. In recent years more and more methods have started to appear that address two, three or more different goals of resource management, but work with a holistic scope that attempts to address and balance all relevant objectives is still rare. An exception is the area of reliability and life-time management. There we see that a large majority of approaches may be considered as holistic.

Observing current limitations and recent trends we assume that global and holistic approaches will make the most valuable contributions to this field in the years to come and therefore we expect that the focus of research will gradually shift to systematic methods combining point techniques in order to pursue and balance all relevant system goals of highly dynamic, adaptive systems.

Acronyms

ALU Arithmetic Logic Unit. 15

AVF Architectural Vulnerability Factor. 73, 74

BIST Built-In Self-Test. 79

BPIO Balanced Placement I/O. 37

CPU Central Processing Unit. 15, 28, 38, 57, 68, 69, 87, 98

DMA Direct Memory Access. 16

DPM Dynamic Power Management. 16, 19, 41, 43, 44, 45, 46, 47, 48, 49

DRAM Dynamic Random-Access Memory. 29, 93

DRM Dynamic Reliability Management. 59, 62, 63, 65, 68, 69, 70, 71

DSP Digital Signal Processor. 13, 15

DTM Dynamic Thermal Management. 41, 51, 52, 59

DVFS Dynamic Voltage and Frequency Scaling. 16, 40, 44, 46, 51, 53, 57, 62, 63, 64, 65, 67, 68, 71, 73, 83

DWC Duplication With Comparison. 76, 78, 81

EDF Earliest Deadline First. 48

EDP Energy Delay Product. 49

EM Electromigration. 57, 58, 60, 61, 65, 67, 68, 70

EPI Energy Per Instruction. 49

FeRAM Ferroelectric Random-Access Memory. 29

FinFET Fin Field-effect. 61, 64

FPGA Field Programmable Gate Array. 15, 52, 83, 93

FPU Floating-Point Unit. 63, 82

GPP General Purpose Processor. 13, 15

GPU Graphics Processing Unit. 15, 19, 28, 38, 59, 68, 69

HCI hot carrier injection. 58, 68, 69

HMP Heterogeneous Multi-core Processors. 47

HPC High Performance Computing. 49, 55, 78

I/O Input/Output. 13, 15, 16, 20, 21, 24, 36, 37, 38, 84, 85, 86, 88, 89

ILP Integer Linear Programming. 45

IPS Instruction Per Second. 26

ISA Instruction Set Architecture. 41, 93

MIMO Multiple-Input Mutiple-Output. 47

MRAM Magnetic Random-Access Memory. 29

MTTF Mean Time To Failure. 59, 60, 61, 62, 63, 64, 65

NBTI Negative Bias Temperature Instability. 57, 58, 60, 61, 66, 67, 68, 69, 70

NMR N-Modular Redundancy. 77

NoC Networks-on-Chip. 8, 16, 30, 31, 33, 42, 46, 53, 59, 60, 62, 65, 67, 77, 81

PCM Phase-Change Memory. 29

PE Processing Element. 13, 15, 16

PI Proportional-Integral. 64

PID Proportional-Integral-Derivative. 64

PLR Process-Level Redundancy. 77

QoR Quality of Result. 94

QoS Quality of Service. 8, 10, 11, 14, 17, 18, 81, 84, 85, 86, 87, 88, 89, 90, 91, 94, 95, 97, 99, 103

RRAM Resistive Random-Access Memory. 29

RTOS Real-Time Operating Systems. 50

SBST Software-Based Self-Test. 79, 80

SMT Simultaneous Multi-Threading. 51, 77

SoC Systems-on-Chip. 13, 42, 45

SRAM Static Random-Access Memory. 29, 50

SSD Solid State Disk. 36, 37, 38

STP System Through-Put. 26

STT-RAM Spin-Transfer Torque Random-Access Memory. 29

TDDB Time Dependent Dielectric Breakdown. 57, 58, 60, 70

TMR Triple Modular Redundancy. 77, 78, 79, 81

WCET Worst-Case Execution Time. 45

1

Introduction

Resource management has a long history in computing, from the early days of time-shared machines with pioneering fundamental work on run-time systems, distributed systems, real-time operating systems and middleware. The evolution in computing architectures – from single- to multi- and many-core platforms – has resulted in a Cambrian explosion in the diversity of computing architectures, applications and end-use cases, due to:

- Chips are getting more complex, with:
  - Increasing core count
  - Increasing core heterogeneity
  - Novel memory technologies and architectures
  - Disparate interconnects and I/O
- Workloads are getting more diverse and unpredictable (e.g., mobile to edge to data center)

- There is a renewed move towards programmable architectures tuned to certain domains (e.g., mobile, edge, cloud) and applications (neural networks, deep learning, security, cryptography, etc.)

- Computing systems must increasingly satisfy multi-dimensional constraints that straddle multiple metrics: performance, power/energy, temperature, reliability, lifetime, Quality of Service (QoS), etc. Furthermore, these constraints themselves may evolve over time (e.g., high-performance at times, throttling for impending thermal emergencies, and strategies for extending lifetime in the face of wear-out)

These trends have generated an incredibly large body of work for on-chip resource management in the past two decades, that have focused on many different combinations of architectures, workloads, constraints and use-cases. Indeed, due to large variations in the assumptions, use of different terminology, metrics, goals, and use-cases, anyone attempting to review the literature can easily get overwhelmed by the volume, diversity, and sometimes even seemingly contradictory approaches for on-chip resource management. Furthermore, advances in program analyses, system modeling, run-time monitoring, model building, and machine learning have generated new lines of research in dynamic and adaptive on-chip resource management that further complicate a global understanding of the current state-of-the-art and emerging directions for on-chip resource management.

A number of surveys offer systematic overviews of parts of the topics, challenges and techniques that we are concerned with. Some of them focus on specific subsystem of a many-core chip like Networks-on-Chip (NoC) (Bjerregaard and Mahadevan, 2006; Marculescu et al., 2009; Rahmani et al., 2010; Radetzki et al., 2013) or caches (Scolari et al., 2014), others deal systematically with particular metrics like power, energy (Maiterth et al., 2018), or aging (Khoshavi et al., 2017). Also, scheduling, mapping and task allocation have received a fair amount of attention in surveys (Singh et al., 2013a, 2017; Zhuravlev et al., 2012; Burns and Davis, 2017), with focus either on performance

Table 1.1: A selection of surveys.

Approximate computing	Mittal (2016); Xu et al. (2016a)
NoC	Bjerregaard and Mahadevan (2006)
NoC design	Marculescu et al. (2009)
3D NoC	Rahmani et al. (2010)
Fault tolerant NoC	Radetzki et al. (2013)
Cache management	Scolari et al. (2014)
Power management in high performance systems	Liu and Zhu (2010)
Energy and power aware job scheduling	Maiterth et al. (2018)
Thermal management	Kong et al. (2012)
Aging mitigation	Khoshavi et al. (2017)
Dark silicon	Shafique and Garg (2017)
Self-aware SoCs	Jantsch et al. (2017)
Energy and power aware job scheduling	Maiterth et al. (2018)
Mapping in many core systems	Singh et al. (2013a)
Resource allocation in hard- and soft- real-time systems	Singh et al. (2017)
Scheduling	Zhuravlev et al. (2012)
Mixed Criticality systems	Burns and Davis (2017)
Resource management in clouds	Jennings and Stadler (2015)

or real-time behavior. Current trends and opportunities like approximate computing (Mittal, 2016; Xu et al., 2016a), the dark silicon phenomenon (Shafique and Garg, 2017) and comprehensive self-monitoring on-chip (Jantsch et al., 2017) have recently inspired researchers to review, analyze and compare relevant work. But none of them, as Table 1.1 illustrates, covers all the aspects of on-chip resource management in a holistic way and many of them also include both design time and run time methods. It should be noted that any resource allocation

approach contributes to multiple metrics (e.g., energy, temperature, QoS, etc.) and cannot be adequately discussed in isolation. Therefore, a comprehensive review is needed to discuss the relation between different categories of on-chip resource management techniques and the metrics of interests in a holistic fashion.

This article attempts to cover all aspects of on-chip run-time resource management to facilitate understanding of recent trends in dynamic and adaptive strategies. We have organized the article into three major sections (also visualized in Figure 1.1):

1. Chapter 2 presents a taxonomy of on-chip resources, design metrics, objectives and constraints. This categorization helps the reader visualize the relationship between different resource categories (e.g., computing, storage, communication) and design metrics (e.g., performance, power/energy, temperature, QoS, lifetime, etc.). This chapter also relates the role of objectives and constraints that guide resource management policies.

2. Chapters 3 through 6 survey literature in dynamic on-chip resource management through the lens of the primary metrics that drive these bodies of work:

    - Chapter 3 reviews efforts focused on the traditional metric of optimizing system performance. Historically performance has been the major driving metric for computing platforms, and for many scenarios it continues to be an important design metric
    - Chapter 4 covers resource management techniques that address the metrics of power, energy and temperature. As the core count increased, the move to multi- and many-core architectures rapidly hit the power wall, resulting in a large body of work on power-aware resource management strategies via both hardware and software techniques. Concurrently, research on run-time energy efficiency gained traction via mapping and scheduling approaches. Thermal management techniques began to appear in the past decade, with the goal

of operating chips at a safe temperature via dynamic thermal management techniques.

- Chapter 5 surveys the large body of work on reliability, via the facets of lifetime management, soft-error resilience, and online fault management. Modern chips are increasingly susceptible to failures from a diverse set of causes, leading to premature lifetime failure (due to aging and wear-out) or intermittent/transient failures (due to soft errors). Numerous techniques have been developed to mitigate these failures. Additionally, there is a large body of work covering online fault management that dynamically detects, and proactively applies fault management strategies to prevent failures.
- Chapter 6 addresses QoS metrics that drive several important classes of applications. While QoS can be a nebulous "qualitative" term, we first give specific examples of QoS for different applications, and relate it to the specific metrics that drive the QoS. Here we survey resource management techniques from two angles: a) performance-bound QoS techniques that achieve desired QoS by controlling resources (e.g., compute, memory, I/O), and b) accuracy-bound QoS for applications that can tolerate some loss of accuracy in exchange for sacrificing resources and/or metrics via both static and dynamic strategies.

Note, that each chapter includes several tables listing all the approaches discussed in the various sections. The goal of these tables is to summarize the discussion and provide the taxonomy of the various approaches based on the specific aspects discussed in the text. Since the discussion in each section is specific to the aspects of the presented subtopic, each table has a custom organization of the rows and columns.

3. Chapter 7 addresses the limitations of existing work and outlines recent trends in enabling adaptive resource management in the face of dynamically varying workloads, system properties, and unforeseen environmental effects. Samples of emerging topics in

dynamic resource management are highlighted as examples of ongoing challenges facing researchers in this domain.

Chapters 3, 4, 5, and 6 can be read independently of each other and in any order. Hence, 2, 4, 7 would be reasonable flow of reading as would be 2, 6, 7, depending on the interest of the reader.

While we have attempted to do a comprehensive survey, it is by no means complete, but should provide the reader with a framework within which to navigate both existing, as well as evolving research efforts in on-chip dynamic resource management.

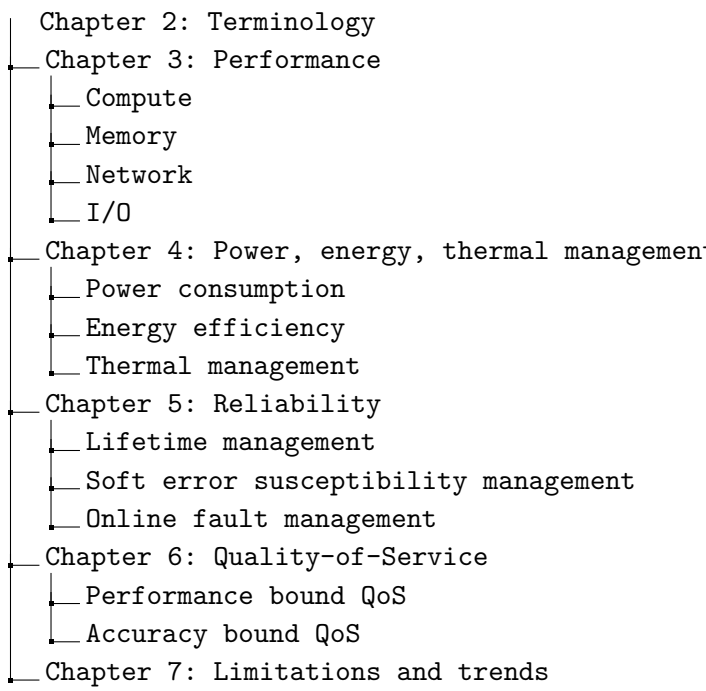

Figure 1.1: Article structure.

References

Acquaviva, A., L. Benini, and B. Riccó (2001), 'Energy Characterization of Embedded Real-time Operating Systems'. *SIGARCH Comput. Archit. News* pp. 13–18.

Agarwal, N., D. Nellans, M. Stephenson, M. O'Connor, and S. W. Keckler (2015), 'Page Placement Strategies for GPUs Within Heterogeneous Memory Systems'. *SIGPLAN Not.*

Aggarwal, N., P. Ranganathan, N. P. Jouppi, and J. E. Smith (2007), 'Configurable Isolation: Building High Availability Systems with Commodity Multi-core Processors'. In: *Proc. of Intl. Symp. on Computer Architecture (ISCA)*. pp. 470–481.

Ahn, J., S. Yoo, and K. Choi (2014), 'Dynamic power management of off-chip links for Hybrid Memory Cubes'. In: *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*. pp. 1–6.

Al Faruque, M. A., T. Ebi, and J. Henkel (2012), 'AdNoC: Runtime adaptive network-on-chip architecture'. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* pp. 257–269.

Annamalai, A., R. Rodrigues, I. Koren, and S. Kundu (2013), 'An Opportunistic Prediction-based Thread Scheduling to Maximize Throughput/Watt in AMPs'. In: *Proceedings of the 22Nd International Conference on Parallel Architectures and Compilation Techniques (PACT)*. pp. 63–72.

Apostolakis, A., D. Gizopoulos, M. Psarakis, and A. Paschalis (2009), 'Software-Based Self-Testing of Symmetric Shared-Memory Multiprocessors'. *IEEE Transactions on Computers* **58**(12), 1682–1694.

Ascia, G., V. Catania, M. Palesi, and D. Patti (2008), 'Implementation and analysis of a new selection strategy for adaptive routing in networks-on-chip'. *IEEE Transactions on Computers* pp. 809–820.

Attia, K. M., M. A. El-Hosseini, and H. A. Ali (2017), 'Dynamic power management techniques in multi-core architectures: A survey study'. *Ain Shams Engineering Journal* pp. 445–456.

Ausavarungnirun, R., K. K.-W. Chang, L. Subramanian, G. H. Loh, and O. Mutlu (2012), 'Staged Memory Scheduling: Achieving High Performance and Scalability in Heterogeneous Systems'. *SIGARCH Comput. Archit. News*.

Ayoub, R., K. R. Indukuri, and T. S. Rosing (2010), 'Energy efficient proactive thermal management in memory subsystem'. In: *2010 ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED)*. pp. 195–200.

Azevedo, A., I. Issenin, R. Cornea, R. Gupta, N. Dutt, A. Veidenbaum, and A. Nicolau (2002), 'Profile-based dynamic voltage scheduling using program checkpoints'. In: *Proceedings 2002 Design, Automation and Test in Europe Conference and Exhibition*. pp. 168–175.

Badr, H. G. and S. Podar (1989), 'An optimal shortest-path routing policy for network computers with regular mesh-connected topologies'. *IEEE transactions on computers* pp. 1362–1371.

Baek, W. and T. M. Chilimbi (2010), 'Green : A Framework for Supporting Energy-Conscious Programming using Controlled Approximation'. In: *Proceedings of the ACM SIGPLAN conference on Programming language design and implementation -PLDI '10*. pp. 198–209.

Bakhoda, A., J. Kim, and T. M. Aamodt (2010), 'Throughput-Effective On-Chip Networks for Manycore Accelerators'. In: *Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. pp. 421–432.

Baldassari, A., C. Bolchini, and A. Miele (2017), 'A dynamic reliability management framework for heterogeneous multicore systems'. In: *Proc. of Intl. Symp. on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*. pp. 1–6.

Balfour, J. and W. J. Dally (2014), 'Design Tradeoffs for Tiled CMP On-chip Networks'. In: *ACM International Conference on Supercomputing 25th Anniversary Volume*. pp. 390–401.

Banerjee, A., P. T. Wolkotte, R. D. Mullins, S. W. Moore, and G. J. M. Smit (2009), 'An Energy and Performance Exploration of Network-on-Chip Architectures'. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* pp. 319–329.

Baynes, K., C. Collins, E. Fiterman, B. Ganesh, P. Kohout, C. Smit, T. Zhang, and B. Jacob (2003), 'The performance and energy consumption of embedded real-time operating systems'. *IEEE Transactions on Computers* pp. 1454–1469.

Beckmann, N. and D. Sanchez (2013), 'Jigsaw: Scalable Software-defined Caches'. In: *Proceedings of the 22Nd International Conference on Parallel Architectures and Compilation Techniques*. Piscataway, NJ, USA, pp. 213–224, IEEE Press.

Beckmann, N., P.-A. Tsai, and D. Sanchez (2015), 'Scaling distributed cache hierarchies through computation and data co-scheduling'. In: *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on*. pp. 538–550.

Beloglazov, A. and R. Buyya (2010), 'Energy Efficient Resource Management in Virtualized Cloud Data Centers'. In: *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*. pp. 826–831.

Bender, M. A., D. P. Bunde, E. D. Demaine, S. P. Fekete, V. J. Leung, H. Meijer, and C. A. Phillips (2008), 'Communication-aware processor allocation for supercomputers: Finding point sets of small average distance'. *Springer Algorithmica* pp. 279–298.

Benini, L., A. Bogliolo, G. A. Paleologo, and G. D. Micheli (1999), 'Policy optimization for dynamic power management'. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* pp. 813–833.

Benini, L. and G. De Micheli (2002), 'Networks on Chips: A New SoC Paradigm'. *Computer* pp. 70–78.

Besta, M., S. M. Hassan, S. Yalamanchili, R. Ausavarungnirun, O. Mutlu, and T. Hoefler (2018), 'Slim NoC: A Low-Diameter On-Chip Network Topology for High Energy Efficiency and Scalability'. In: *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. pp. 43–55.

Besta, M. and T. Hoefler (2014), 'Slim Fly: A Cost Effective Low-diameter Network Topology'. In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. pp. 348–359.

Bhardwaj, K., K. Chakraborty, and S. Roy (2012), 'Towards graceful aging degradation in NoCs through an adaptive routing algorithm'. In: *Proc. of Design Automation Conf. (DAC)*. pp. 382–391.

Bhardwaj, S., W. Wang, R. Vattikonda, Y. Cao, and S. Vrudhula (2006), 'Predictive Modeling of the NBTI Effect for Reliable Design'. In: *Proc. of IEEE Custom Integrated Circuits Conf. (CICC)*. pp. 189–192.

Bitirgen, R., E. Ipek, and J. Martinez (2008), 'Coordinated management of multiple interacting resources in chip multiprocessors: A machine learning approach'. In: *41st annual IEEE/ACM International Symposium on Microarchitecture*. pp. 318–329.

Bjerregaard, T. and S. Mahadevan (2006), 'A Survey of Research and Practices of Network-on-chip'. *ACM Computing Surveys* **38**(1).

Blanton, R. D., X. Li, K. Mai, D. Marculescu, R. Marculescu, J. Paramesh, J. Schneider, and D. E. Thomas (2015), 'Statistical Learning in Chip (SLIC)'. In: *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design, ICCAD*. pp. 664–669.

Blome, J., S. Feng, S. Gupta, and S. Mahlke (2007), 'Self-calibrating Online Wearout Detection'. In: *Proc. of Intl. Symp. on Microarchitecture (MICRO)*. pp. 109–122.

Bogdan, P., R. Marculescu, and S. Jain (2013), 'Dynamic Power Management for Multidomain System-on-chip Platforms: An Optimal Control Approach'. *ACM Trans. Des. Autom. Electron. Syst.* pp. 46:1–46:20.

Bolchini, C., M. Carminati, and A. Miele (2013a), 'Self-Adaptive Fault Tolerance in Multi-/Many-Core Systems'. *Journal of Electronic Testing: Theory and Application* **29**(2), 159–175.

Bolchini, C., M. Carminati, A. Miele, A. Das, A. Kumar, and B. Veeravalli (2013b), 'Run-Time Mapping for Reliable Many-Cores Based on Energy/Performance Trade-offs'. In: *Proc. of Intl. Symp. on Defect and Fault Tolerance in VLSI and Nanotech. Systems (DFT)*. pp. 58–64.

Bolchini, C., L. Cassano, and A. Miele (2016), 'Lifetime-aware load distribution policies in multi-core systems: An in-depth analysis'. In: *Proc. of Conf. on Design, Automation & Test in Europe (DATE)*. pp. 804–809.

Bolchini, C., A. Miele, and D. Sciuto (2012), 'An adaptive approach for online fault management in many-core architectures'. In: *Proc. of Design, Automation Test in Europe Conf. Exhibition (DATE)*. pp. 1429–1432.

Bornholt, J., T. Mytkowicz, and K. S. McKinley (2014), 'Uncertain<T>: A First-Order Type for Uncertain Data'. In: *Proceedings of International conference on Architectural support for programming languages and operating systems - ASPLOS '14*. pp. 51–66.

Brooks, D. and M. Martonosi (2001), 'Dynamic thermal management for high-performance microprocessors'. In: *Proceedings HPCA Seventh International Symposium on High-Performance Computer Architecture*. pp. 171–182.

Bruestel, M. and A. Kumar (2017), 'Accounting for systematic errors in approximate computing'. In: *Proceedings of the Conference on Design, Automation & Test in Europe*. pp. 298–301.

Burns, A. and R. Davis (2013), 'Mixed criticality systems-a review'. *Department of Computer Science, University of York, Tech. Rep* pp. 1–69.

Burns, A. and R. I. Davis (2017), 'A Survey of Research into Mixed Criticality Systems'. *ACM Computing Surveys* **50**(6), 82:1–82:37.

Cai, E., D. Stamoulis, and D. Marculescu (2016), 'Exploring Aging Deceleration in FinFET-based Multi-core Systems'. In: *Proc of Intl. Conf. on Computer-Aided Design (ICCAD)*. pp. 111:1–111:8.

Carara, E. A. and F. G. Moraes (2010), 'Flow oriented routing for NOCS'. In: *SOC Conference (SOCC), 2010 IEEE International*. pp. 367–370.

Carbin, M., D. Kim, S. Misailovic, and M. C. Rinard (2012), 'Proving acceptability properties of relaxed nondeterministic approximate programs'. In: *Proceedings of {ACM} {SIGPLAN} Conference on Programming Language Design and Implementation - PLDI '12*, Vol. 47. pp. 169–180.

Carbin, M., S. Misailovic, M. Kling, and M. C. Rinard (2011), 'Detecting and escaping infinite loops with jolt'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 6813 LNCS. pp. 609–633.

Carbin, M., S. Misailovic, and M. C. Rinard (2013), 'Verifying quantitative reliability for programs that execute on unreliable hardware'. In: *Proceedings of Conference on Object-Oriented Programming Systems, Languages, and Applications - OOPSLA '13*. pp. 33–52.

Carbin, M. and M. C. M. Rinard (2010), 'Automatically identifying critical input regions and code in applications'. In: *Proceedings of international symposium on Software testing and analysis*. pp. 37–48.

Carvalho, E., N. Calazans, and F. Moraes (2007), 'Heuristics for dynamic task mapping in NoC-based heterogeneous MPSoCs'. In: *Proc. of IEEE/IFIP International Workshop on Rapid System Prototyping*. pp. 34–40.

Castrillon, J., A. Tretter, R. Leupers, and G. Ascheid (2012), 'Communication-aware mapping of KPN applications onto heterogeneous MPSoCs'. In: *Proceedings of the 49th Annual Design Automation Conference*. pp. 1266–1271.

Catania, V., R. Holsmark, S. Kumar, and M. Palesi (2006), 'A methodology for design of application specific deadlock-free routing algorithms for NoC systems'. In: *Hardware/Software Codesign and System Synthesis, 2006. CODES+ ISSS'06. Proceedings of the 4th International Conference*. pp. 142–147.

Ceratti, A., T. Copetti, L. Bolzani, and F. Vargas (2012), 'On-chip aging sensor to monitor NBTI effect in nano-scale SRAM'. In: *Proc. of Intl. Symp. on Design and Diagnostics of Electronic Circuits Systems (DDECS)*. pp. 354–359.

Chang, D.-W., H.-H. Chen, and W.-J. Su (2015a), 'VSSD: Performance Isolation in a Solid-State Drive'. *ACM Trans. Des. Autom. Electron. Syst.* pp. 51:1–51:33.

Chang, E.-J., H.-K. Hsin, C.-H. Chao, S.-Y. Lin, and A.-Y. A. Wu (2015b), 'Regional ACO-based cascaded adaptive routing for traffic balancing in mesh-based network-on-chip systems'. *IEEE Transactions on Computers* pp. 868–875.

Chang, E.-J., H.-K. Hsin, S.-Y. Lin, and A.-Y. Wu (2014), 'Path-congestion-aware adaptive routing with a contention prediction scheme for network-on-chip systems'. *IEEE Transactions on computer-aided design of Integrated circuits and systems* pp. 113–126.

Chang, K. K.-W., R. Ausavarungnirun, C. Fallin, and O. Mutlu (2012), 'HAT: Heterogeneous adaptive throttling for on-chip networks'. In: *2012 IEEE 24th International Symposium on Computer Architecture and High Performance Computing*. pp. 9–18.

Chantem, T., Y. Xiang, X. S. Hu, and R. P. Dick (2013), 'Enhancing Multicore Reliability through Wear Compensation in Online Assignment and Scheduling'. In: *Proc. of Conf. on Design, Automation & Test in Europe (DATE)*. pp. 1373–1378.

Chen, C. and A. Joshi (2013), 'Runtime Management of Laser Power in Silicon-Photonic Multibus NoC Architecture'. *IEEE Journal of Selected Topics in Quantum Electronics*.

Chen, C. O., S. Park, T. Krishna, S. Subramanian, A. P. Chandrakasan, and L. Peh (2013), 'SMART: A single-cycle reconfigurable NoC for SoC applications'. In: *2013 Design, Automation Test in Europe Conference Exhibition (DATE)*.

Chen, F., D. A. Koufaty, and X. Zhang (2011), 'Hystor: Making the Best Use of Solid State Drives in High Performance Storage Systems'. In: *Proceedings of the International Conference on Supercomputing*. pp. 22–32.

Chen, G., F. Li, S. W. Son, and M. Kandemir (2008), 'Application mapping for chip multiprocessors'. In: *Proceedings of the 45th annual design automation conference*. pp. 620–625.

Chen, K. H., J. J. Chen, F. Kriebel, S. Rehman, M. Shafique, and J. Henkel (2016), 'Task Mapping for Redundant Multithreading in Multi-Cores with Reliability and Performance Heterogeneity'. *IEEE Transactions on Computers* **65**(11), 3441–3455.

Chen, Q. and M. Guo (2014), 'Adaptive Workload-aware Task Scheduling for single-ISA Asymmetric Multicore Architectures'. *ACM Trans. Archit. Code Optim.* pp. 8:1–8:25.

Chen, T., Z. Du, N. Sun, J. Wang, C. Wu, Y. Chen, and O. Temam (2014a), 'DianNao: A Small-Footprint High-Throughput Accelerator for Ubiquitous Machine-Learning'. In: *Proceedings of International conference on Architectural support for programming languages and operating systems*. pp. 269–284.

Chen, X., Y. Wang, Y. Liang, Y. Xie, and H. Yang (2014b), 'Run-time technique for simultaneous aging and power optimization in GPGPUs'. In: *Proc. of Design Automation Conf. (DAC)*. pp. 1–6.

Chen, Y., T. Luo, S. Liu, S. Zhang, L. He, J. Wang, L. Li, T. Chen, Z. Xu, N. Sun, and O. Temam (2014c), 'DaDianNao: A Machine-Learning Supercomputer'. In: *IEEE/ACM International Symposium on Microarchitecture*. pp. 609–622.

Chen, Y.-Y., E.-J. Chang, H.-K. Hsin, K.-C. J. Chen, and A.-Y. A. Wu (2017), 'Path-Diversity-Aware Fault-Tolerant Routing Algorithm for Network-on-Chip Systems'. *IEEE Transactions on Parallel and Distributed Systems* pp. 838–849.

Chen, Z. and D. Marculescu (2015a), 'Distributed Reinforcement Learning for Power Limited Many-core System Performance Optimization'. In: *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*.

Chen, Z. and D. Marculescu (2015b), 'Distributed Reinforcement Learning for Power Limited Many-core System Performance Optimization'. In: *Proceedings of the 2015 Design, Automation & Test in Europe Conference (DATE)*. pp. 1521–1526.

Chen, Z., M. Yang, G. Francia, and J. Dongarra (2007), 'Self Adaptive Application Level Fault Tolerance for Parallel and Distributed Computing'. In: *Proc. of Intl. Parallel and Distributed Processing Symp.- (IPDPS)*. pp. 1–8.

Cheng, S.-T., C.-M. Chen, and J.-W. Hwang (1997), 'Low-power design for real-time systems'. In: *Proceedings of ICICS, 1997 International Conference on Information, Communications and Signal Processing. Theme: Trends in Information Systems Engineering and Wireless Multimedia Communications (Cat.* pp. 1746–1750 vol.3.

Cho, S. and L. Jin (2006), 'Managing distributed, shared L2 caches through OS-level page allocation'. In: *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*. pp. 455–468.

Chou, C., U. Y. Ogras, and R. Marculescu (2008), 'Energy- and Performance-Aware Incremental Mapping for Networks on Chip With Multiple Voltage Levels'. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

Chou, C. L. and R. Marculescu (2011), 'FARM: Fault-aware resource management in NoC-based multiprocessor platforms'. In: *Proc. of Design, Automation Test in Europe Conf. (DATE)*. pp. 1–6.

Chou, C.-L., U. Y. Ogras, and R. Marculescu (2008), 'Energy-and performance-aware incremental mapping for networks on chip with multiple voltage levels'. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **27**(10), 1866–1879.

Chou, C. T., Y. P. Lin, K. Y. Chiang, and K. C. Chen (2017), 'Dynamic Buffer Allocation for thermal-aware 3D network-on-chip systems'. In: *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*. pp. 65–66.

Christoforakis, I., O. Tomoutzoglou, D. Bakoyiannis, and G. Kornaros (2015), 'Dithering-Based Power and Thermal Management on FPGA-Based Multi-core Embedded Systems'. In: *2015 IEEE 13th International Conference on Embedded and Ubiquitous Computing*. pp. 173–177.

Chung, E.-Y., L. Benini, A. Bogliolo, and G. D. Micheli (1999), 'Dynamic power management for nonstationary service requests'. In: *Design, Automation and Test in Europe Conference and Exhibition, 1999. Proceedings (Cat. No. PR00078)*. pp. 77–81.

Cochran, R., C. Hankendi, A. K. Coskun, and S. Reda (2011), 'Pack & Cap: Adaptive DVFS and Thread Packing Under Power Caps'. In: *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*.

Coskun, A. K., T. S. Rosing, and K. C. Gross (2009), 'Utilizing Predictors for Efficient Thermal Management in Multiprocessor SoCs'. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **28**(10), 1503–1516.

Coskun, A. K., T. S. Rosing, and K. Whisnant (2007), 'Temperature Aware Task Scheduling in MPSoCs'. In: *2007 Design, Automation Test in Europe Conference Exhibition*.

Coskun, A. K., R. Strong, D. M. Tullsen, and T. S. Rosing (2009), 'Evaluating the Impact of Job Scheduling and Power Management on Processor Lifetime for Chip Multiprocessors'. In: *Proc. of Intl. Conf. Measurement and Modeling of Computer Systems (SIGMETRICS)*. pp. 169–180.

Cui, Z., S. A. Mckee, Z. Zha, Y. Bao, and M. Chen (2014), 'DTail : A Flexible Approach to DRAM Refresh Management'. In: *Proceedings of International Conference on Supercomputing - SC '14*. pp. 43–52.

Dai, J., W. Ma, X. Jiang, and T. Watanabe (2017), 'Hybrid path-diversity-dominant output selection method for Network-on-Chip systems'. In: *SoC Design Conference (ISOCC), 2017 International*. pp. 125–126.

Dally, W. J. and H. Aoki (1993), 'Deadlock-free adaptive routing in multicomputer networks using virtual channels'. *IEEE transactions on Parallel and Distributed Systems* pp. 466–475.

Dally, W. J. and B. Towles (2001), 'Route Packets, Not Wires: On-chip Interconnection Networks'. In: *Proceedings of the 38th Annual Design Automation Conference (DAC)*. pp. 684–689.

Das, A., R. Shafik, G. V. Merrett, B. M. Al-Hashimi, A. Kumar, and B. Veeravalli (2014), 'Reinforcement Learning-Based Inter- and Intra-Application Thermal Optimization for Lifetime Improvement of Multicore Systems'. In: *Proc. of Design Automation Conf. (DAC)*. pp. 170:1–170:6.

Das, A., M. J. Walker, A. Hansson, B. M. Al-Hashimi, and G. V. Merrett (2015), 'Hardware-software interaction for run-time power optimization: A case study of embedded Linux on multicore smartphones'. In: *2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*.

David, R., P. Bogdan, R. Marculescu, and U. Ogras (2011), 'Dynamic power management of voltage-frequency island partitioned Networks-on-Chip using Intel's Single-chip Cloud Computer'. In: *Proceedings of the Fifth ACM/IEEE International Symposium*. pp. 257–258.

de Souza Carvalho, E. L., N. L. V. Calazans, and F. G. Moraes (2010), 'Dynamic task mapping for MPSoCs'. *IEEE Design & Test of Computers* pp. 26–35.

Delimitrou, C. and C. Kozyrakis (2013), 'Paragon: QoS-aware Scheduling for Heterogeneous Datacenters'. In: *Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems*. pp. 77–88.

Delimitrou, C. and C. Kozyrakis (2014), 'Quasar: Resource-efficient and QoS-aware Cluster Management'. In: *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. pp. 127–144.

Deo, N. and C.-Y. Pang (1984), 'Shortest-path algorithms: Taxonomy and annotation'. *Networks* pp. 275–323.

Diemer, J. and R. Ernst (2010), 'Back Suction: Service Guarantees for Latency-Sensitive On-chip Networks'. In: *Proceedings of the 2010 Fourth ACM/IEEE International Symposium on Networks-on-Chip (NOCS)*. pp. 155–162.

Diemer, J., R. Ernst, and M. Kauschke (2010), 'Efficient throughput-guarantees for latency-sensitive networks-on-chip'. In: *Design Automation Conference (ASP-DAC), 2010 15th Asia and South Pacific*. pp. 529–534.

Donald, J. and M. Martonosi (2005), 'Leveraging Simultaneous Multithreading for Adaptive Thermal Control'. In: *Proc. of the Second Workshop on Temperature-Aware Computer Systems*.

Donyanavard, B., A. M. Rahmani, T. Muck, K. Moazemmi, and N. Dutt (2018), 'Gain scheduled control for nonlinear power management in CMPs'. In: *2018 Design, Automation Test in Europe Conference Exhibition (DATE)*. pp. 921–924.

Du, Z., R. Fasthuber, T. Chen, P. Ienne, L. Li, X. Feng, Y. Chen, and O. Temam (2015), 'ShiDianNao : Shifting Vision Processing Closer to the Sensor'. In: *Proceedings of IEEE/ACM International Symposium on Computer Architecture*.

Dutt, N., A. Jantsch, and S. Sarma (2016), 'Toward Smart Embedded Systems: A Self-aware System-on-Chip (SoC) Perspective'. *ACM Trans. Embed. Comput. Syst.* pp. 22:1–22:27.

Ebrahimi, E., C. J. Lee, O. Mutlu, and Y. N. Patt (2010), 'Fairness via Source Throttling: A Configurable and High-performance Fairness Substrate for Multi-core Memory Systems'. In: *Proceedings of the Fifteenth Edition of ASPLOS on Architectural Support for Programming Languages and Operating Systems*. pp. 335–346.

El-Harouni, W., S. Rehman, B. S. Prabakaran, A. Kumar, R. Hafiz, and M. Shafique (2017), 'Embracing approximate computing for energy-efficient motion estimation in high efficiency video coding'. In: *Proceedings of the Conference on Design, Automation & Test in Europe*. pp. 1388–1393.

Elyasi, N., M. Arjomand, A. Sivasubramaniam, M. T. Kandemir, C. R. Das, and M. Jung (2017), 'Exploiting Intra-Request Slack to Improve SSD Performance'. *SIGARCH Comput. Archit. News* pp. 375–388.

Esmaeilzadeh, H. (2015), 'Approximate Acceleration: A Path Through the Era of Dark Silicon and Big Data'. In: *Proceedings of the 2015 International Conference on Compilers, Architecture and Synthesis for Embedded Systems*. Piscataway, NJ, USA, pp. 31–32, IEEE Press.

Esmaeilzadeh, H., A. Sampson, L. Ceze, and D. Burger (2012a), 'Architecture support for disciplined approximate programming'. In: *ACM SIGARCH Computer Architecture News*, Vol. 40(1). p. 301.

Esmaeilzadeh, H., A. Sampson, L. Ceze, and D. Burger (2012b), 'Neural Acceleration for General-Purpose Approximate Programs'. In: *Proceedings of IEEE/ACM International Symposium on Microarchitecture*. pp. 449–460, Ieee.

Eyerman, S. and L. Eeckhout (2008), 'System-Level Performance Metrics for Multiprogram Workloads'. *IEEE Micro* pp. 42–53.

Faruque, A., M. Abdullah, R. Krist, and J. Henkel (2008), 'ADAM: run-time agent-based distributed application mapping for on-chip communication'. In: *Proceedings of the 45th annual Design Automation Conference*. pp. 760–765.

Fattah, M., M. Daneshtalab, P. Liljeberg, and J. Plosila (2013), 'Smart hill climbing for agile dynamic mapping in many-core systems'. In: *Proc. of IEEE/ACM Design Automation Conference*.

Fattah, M., P. Liljeberg, J. Plosila, and H. Tenhunen (2014), 'Adjustable contiguity of run-time task allocation in networked many-core systems'. In: *Proc. of IEEE Asia and South Pacific Design Automation Conference*. pp. 349–354.

Fattah, M., M. Ramirez, M. Daneshtalab, P. Liljeberg, and J. Plosila (2012), 'CoNA: Dynamic application mapping for congestion reduction in many-core systems'. In: *Proc. of IEEE International Conference on Computer Design*. pp. 364–370.

Feng, C., Z. Lu, A. Jantsch, J. Li, and M. Zhang (2010), 'A Reconfigurable Fault-tolerant Deflection Routing Algorithm Based on Reinforcement Learning for Networks-on-Chip'. In: *Proceedings of the International Workshop on Network on Chip Architectures (NoCArc)*.

Feng, C., Z. Lu, A. Jantsch, and M. Zhang (2012), 'A 1-cycle 1.25GHz Bufferless Router for 3D Network-on-Chip'. *IEICE Transactions on Information and Systems*.

Feng, W.-c. and K. G. Shin (1997), 'Impact of Selection Functions on Routing Algorithm Performance in Multicomputer Networks'. In: *Proceedings of the 11th International Conference on Supercomputing (ICS)*. pp. 132–139.

Foutris, N., M. Psarakis, D. Gizopoulos, A. Apostolakis, X. Vera, and A. Gonzalez (2010), 'MT-SBST: Self-test optimization in multithreaded multicore architectures'. In: *Proc. of IEEE Intl. Test Conf. (ITC)*. pp. 1–10.

Gaspar, F., A. Ilic, P. Tomás, and L. Sousa (2014), 'Performance-aware task management and frequency scaling in embedded systems'. In: *Computer Architecture and High Performance Computing (SBAC-PAD), 2014 IEEE 26th International Symposium on*. pp. 65–72.

Ge, R., X. Feng, S. Song, H. C. Chang, D. Li, and K. W. Cameron (2010), 'PowerPack: Energy Profiling and Analysis of High-Performance Systems and Applications'. *IEEE Transactions on Parallel and Distributed Systems* pp. 658–671.

Gebregiorgis, A., S. Kiamehr, and M. B. Tahoori (2017), 'Error Propagation Aware Timing Relaxation For Approximate Near Threshold Computing'. In: *Proceedings of the 54th Annual Design Automation Conference 2017*. New York, NY, USA, pp. 77:1–77:6, ACM.

Ghose, S., H. Lee, and J. F. Martínez (2013), 'Improving Memory Scheduling via Processor-side Load Criticality Information'. In: *Proceedings of the 40th Annual International Symposium on Computer Architecture*. pp. 84–95.

Glass, C. J. and L. M. Ni (1992), 'The Turn Model for Adaptive Routing'. In: *Proceedings of the 19th Annual International Symposium on Computer Architecture (ISCA)*. pp. 278–287.

Gnad, D., M. Shafique, F. Kriebel, S. Rehman, D. Sun, and J. Henkel (2015), 'Hayat: Harnessing Dark Silicon and Variability for Aging Deceleration and Balancing'. In: *Proc. of Design Automation Conf. (DAC)*. pp. 180:1–180:6.

Gratz, P., B. Grot, and S. W. Keckler (2008), 'Regional congestion awareness for load balance in networks-on-chip'. In: *High Performance Computer Architecture, 2008. HPCA 2008. IEEE 14th International Symposium on*. pp. 203–214.

Grigorian, B., N. Farahpour, and G. Reinman (2015), 'BRAINIAC: Bringing reliable accuracy into neurally-implemented approximate computing'. In: *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on*. pp. 615–626.

Grot, B., J. Hestness, S. W. Keckler, and O. Mutlu (2009a), 'Express cube topologies for on-chip interconnects'. In: *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on*. pp. 163–174.

Grot, B., J. Hestness, S. W. Keckler, and O. Mutlu (2011), 'Kilo-NOC: A Heterogeneous Network-on-chip Architecture for Scalability and Service Guarantees'. In: *Proceedings of the 38th Annual International Symposium on Computer Architecture*. New York, NY, USA, pp. 401–412, ACM.

Grot, B., S. W. Keckler, and O. Mutlu (2009b), 'Preemptive Virtual Clock: A Flexible, Efficient, and Cost-effective QOS Scheme for Networks-on-chip'. In: *Proceedings of the 42Nd Annual IEEE/ACM International Symposium on Microarchitecture*. New York, NY, USA, pp. 268–279, ACM.

Guang, L., P. Liljeberg, E. Nigussie, and H. Tenhunen (2009), 'A review of dynamic power management methods in NoC under emerging design considerations'. In: *2009 NORCHIP*. pp. 1–6.

Guo, F., H. Kannan, L. Zhao, R. Illikkal, R. Iyer, D. Newell, Y. Solihin, and C. Kozyrakis (2007a), 'From chaos to QoS: case studies in CMP resource management'. *ACM SIGARCH Computer Architecture News* pp. 21–30.

Guo, F., Y. Solihin, L. Zhao, and R. Iyer (2007b), 'A framework for providing quality of service in chip multi-processors'. In: *Microarchitecture, 2007. MICRO 2007. 40th Annual IEEE/ACM International Symposium on*. pp. 343–355.

Gupta, A., Y. Kim, and B. Urgaonkar (2009), 'DFTL: A Flash Translation Layer Employing Demand-based Selective Caching of Page-level Address Mappings'. *SIGPLAN Not.* pp. 229–240.

Gupta, S., S. Feng, A. Ansari, J. Blome, and S. Mahlke (2008), 'The StageNet fabric for constructing resilient multicore systems'. In: *Proc. of Intl. Symp. on Microarchitecture (MICRO)*. pp. 141–151.

Gupta, U., M. Babu, R. Ayoub, M. Kishinevsky, F. Paterna, and U. Y. Ogras (2018), 'STAFF: online learning with stabilized adaptive forgetting factor and feature selection algorithm'. In: *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*. pp. 1–6.

Gupta, U., J. Campbell, U. Y. Ogras, R. Ayoub, M. Kishinevsky, F. Paterna, and S. Gumussoy (2016), 'Adaptive performance prediction for integrated GPUs'. In: *2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*. pp. 1–8.

Gurumurthi, S., A. Sivasubramaniam, M. Kandemir, and H. Franke (2003), 'DRPM: dynamic speed control for power management in server class disks'. In: *30th Annual International Symposium on Computer Architecture, 2003. Proceedings*. pp. 169–179.

Ha, C. Y., Y. X. Wang, and C. W. Chang (2017), 'Dynamic Power Management for wearable devices with Non-Volatile Memory'. In: *2017 International Conference on Applied System Innovation (ICASI)*. pp. 37–39.

Haghbayan, M. H., A. Kanduri, A. M. Rahmani, P. Liljeberg, A. Jantsch, and H. Tenhunen (2015), 'MapPro: Proactive Runtime Mapping for Dynamic Workloads by Quantifying Ripple Effect of Applications on Networks-on-Chip'. In: *Proc. of Intl. Symp. on Networks-on-Chip (NOCS)*. pp. 1–8.

Haghbayan, M. H., A. Miele, A. M. Rahmani, P. Liljeberg, and H. Tenhunen (2016a), 'A lifetime-aware runtime mapping approach for many-core systems in the dark silicon era'. In: *Proc. of Conf. on Design, Automation & Test in Europe (DATE)*. pp. 854–857.

Haghbayan, M. H., A. Miele, A. M. Rahmani, P. Liljeberg, and H. Tenhunen (2017), 'Performance/Reliability-Aware Resource Management for Many-Cores in Dark Silicon Era'. *IEEE Transactions on Computers* **66**(9), 1599–1612.

Haghbayan, M. H., A. M. Rahmani, A. Miele, M. Fattah, J. Plosila, P. Liljeberg, and H. Tenhunen (2016b), 'A Power-Aware Approach for Online Test Scheduling in Many-Core Architectures'. *IEEE Transactions on Computers* **65**(3), 730–743.

Hajimiri, H., M. A. Qathrady, and P. Mishra (2013), 'Proactive thermal management using memory based computing'. In: *2013 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*. pp. 110–115.

Hamers, J. and L. Eeckhout (2010), 'Scenario-Based Resource Prediction for QoS-Aware Media Processing'. *Computer* pp. 56–63.

Hari, S. K. S., M. L. Li, P. Ramachandran, B. Choi, and S. V. Adve (2009), 'mSWAT: Low-cost hardware fault detection and diagnosis for multicore systems'. In: *Proc. of Intl. Symp. on Microarchitecture (MICRO)*. pp. 122–132.

Hartman, A. S. and D. E. Thomas (2012), 'Lifetime improvement through runtime wear-based task mapping'. In: *Proc. of Intl. Conf. Hardware/software codesign and system synthesis (CODES)*. pp. 13–22.

Heißwolf, J., R. König, and J. Becker (2012), 'A scalable noc router design providing qos support using weighted round robin scheduling'. In: *Parallel and Distributed Processing with Applications (ISPA), 2012 IEEE 10th International Symposium on*. pp. 625–632.

Hemani, A., A. Jantsch, S. Kumar, A. Postula, J. Öberg, M. Millberg, and D. Lindqvist (2000), 'Network on Chip: An architecture for billion transistor era'. In: *Proceeding of the IEEE NorChip Conference*.

Herbert, S. and D. Marculescu (2007), 'Analysis of dynamic voltage/frequency scaling in chip-multiprocessors'. In: *Proceedings of the 2007 international symposium on Low power electronics and design (ISLPED '07)*.

Herdrich, A., E. Verplanke, P. Autee, R. Illikkal, C. Gianos, R. Singhal, and R. Iyer (2016), 'Cache QoS: From concept to reality in the Intel® Xeon® processor E5-2600 v3 product family'. In: *High Performance Computer Architecture (HPCA), 2016 IEEE International Symposium on*. pp. 657–668.

Hetherington, G., T. Fryars, N. Tamarapalli, M. Kassab, A. Hassan, and J. Rajski (1999), 'Logic BIST for large industrial designs: real issues and case studies'. In: *Proc. of Intl. Test Conf. (ITC)*. pp. 358–367.

Ho, N.-M., E. Manogaran, W.-F. Wong, and A. Anoosheh (2017), 'Efficient floating point precision tuning for approximate computing'. In: *Design Automation Conference (ASP-DAC), 2017 22nd Asia and South Pacific*. pp. 63–68.

Hoffmann, H., J. Eastep, M. D. Santambrogio, J. E. Miller, and A. Agarwal (2010), 'Application heartbeats for software performance and health'. *SIGPLAN Not.* pp. 347–348.

Hoffmann, H., S. Sidiroglou, M. Carbin, S. Misailovic, A. Agarwal, and M. Rinard (2012), 'Dynamic knobs for responsive power-aware computing'. *ACM SIGPLAN Notices* **47**(4), 199.

Hong, S., S. H. K. Narayanan, M. Kandemir, and Ö. Özturk (2009), 'Process variation aware thread mapping for chip multiprocessors'. In: *Proceedings of the Conference on Design, Automation and Test in Europe*. pp. 821–826.

Horowitz, M., T. Indermaur, and R. Gonzalez (1994), 'Low-power digital design'. In: *Proceedings of 1994 IEEE Symposium on Low Power Electronics*. pp. 8–11.

Howard, J., S. Dighe, Y. Hoskote, S. Vangal, D. Finan, G. Ruhl, D. Jenkins, H. Wilson, N. Borkar, G. Schrom, et al. (2010), 'A 48-core IA-32 message-passing processor with DVFS in 45nm CMOS'. In: *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*. pp. 108–109.

Hsu, C.-H. and W.-C. Feng (2005), 'A Power-Aware Run-Time System for High-Performance Computing'. In: *Proceedings of the ACM/IEEE Supercomputing Conference*.

Hu, J. and R. Marculescu (2003), 'Energy-aware mapping for tile-based NoC architectures under performance constraints'. In: *Proceedings of the 2003 Asia and South Pacific Design Automation Conference*. pp. 233–239.

Hu, J. and R. Marculescu (2003), 'Exploiting the routing flexibility for energy/performance aware mapping of regular NoC architectures'. In: *2003 Design, Automation and Test in Europe Conference and Exhibition*.

Hu, J. and R. Marculescu (2004), 'DyAD: smart routing for networks-on-chip'. In: *Proceedings of the 41st annual Design Automation Conference*. pp. 260–263.

Hu, J. and R. Marculescu (2005), 'Energy-and performance-aware mapping for regular NoC architectures'. *IEEE Transactions on computer-aided design of integrated circuits and systems* **24**(4), 551–562.

Huang, J., J. Lach, and G. Robins (2012), 'A Methodology for Energy-quality Tradeoff Using Imprecise Hardware'. In: *Proceedings of the 49th Annual Design Automation Conference*. New York, NY, USA, pp. 504–509, ACM.

Huang, K., L. Santinelli, J. J. Chen, L. Thiele, and G. C. Buttazzo (2009), 'Adaptive Dynamic Power Management for Hard Real-Time Systems'. In: *2009 30th IEEE Real-Time Systems Symposium*. pp. 23–32.

Huang, L. and Q. Xu (2010), 'Performance Yield-driven Task Allocation and Scheduling for MPSoCs Under Process Variation'. In: *Proceedings of the 47th Design Automation Conference*. pp. 326–331.

Huang, L., R. Ye, and Q. Xu (2011), 'Customer-aware task allocation and scheduling for multi-mode MPSoCs'. In: *Proc. of Design Automation Conf. (DAC)*. pp. 387–392.

Hughes, C. J., J. Srinivasan, and S. V. Adve (2001), 'Saving energy with architectural and frequency adaptations for multimedia applications'. In: *Proceedings. 34th ACM/IEEE International Symposium on Microarchitecture. MICRO-34*. pp. 250–261.

Hwang, W., S. Yoo, H. Ko, and B. Park (2010), 'An area efficient temperature sensor with software calibration for mobile application'. In: *2010 International SoC Design Conference*. pp. 349–352.

Isci, C., A. Buyuktosunoglu, C. y. Cher, P. Bose, and M. Martonosi (2006a), 'An Analysis of Efficient Multi-Core Global Power Management Policies: Maximizing Performance for a Given Power Budget'. In: *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06)*. pp. 347–358.

Isci, C., G. Contreras, and M. Martonosi (2006b), 'Live, Runtime Phase Monitoring and Prediction on Real Systems with Application to Dynamic Power Management'. In: *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06)*. pp. 359–370.

Ishihara, T. and H. Yasuura (1998), 'Voltage scheduling problem for dynamically variable voltage processors'. In: *Proceedings. 1998 International Symposium on Low Power Electronics and Design (IEEE Cat. No.98TH8379)*. pp. 197–202.

Iyer, R., L. Zhao, F. Guo, R. Illikkal, S. Makineni, D. Newell, Y. Solihin, L. Hsu, and S. Reinhardt (2007), 'QoS Policies and Architecture for Cache/Memory in CMP Platforms'. In: *Proceedings of the 2007 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. pp. 25–36.

Jantsch, A., A. Anzanpour, H. Kolerdi, I. Azimi, L. C. Siafara, A. M. Rahmani, N. TaheriNejad, P. Liljeberg, and N. Dutt (2018), 'Hierarchical Dynamic Goal Management for IoT Systems'. In: *Proceedings of the IEEE International Symposium on Quality Electronic Design (ISQED 2018)*. USA.

Jantsch, A., N. Dutt, and A. M. Rahmani (2017), 'Self-Awareness in Systems on Chip – A Survey'. *IEEE Design Test* **34**(6), 1–19.

JEDEC Solid State Tech. Ass. (2010), 'Failure mechanisms and models for semiconductor devices'. *JEDEC Publication JEP122G*.

Jennings, B. and R. Stadler (2015), 'Resource Management in Clouds: Survey and Research Challenges'. *Journal of Network and Systems Management* **23**(3), 567–619.

Joao, J. A., M. A. Suleman, O. Mutlu, and Y. N. Patt (2012), 'Bottleneck Identification and Scheduling in Multithreaded Applications'. In: *Proceedings of the Seventeenth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. pp. 223–234.

Joao, J. A., M. A. Suleman, O. Mutlu, and Y. N. Patt (2013), 'Utility-based Acceleration of Multithreaded Applications on Asymmetric CMPs'. In: *Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA)*. pp. 154–165.

Jung, H. and M. Pedram (2006), 'Stochastic Dynamic Thermal Management: A Markovian Decision-based Approach'. In: *2006 International Conference on Computer Design*. pp. 452–457.

Kadjo, D., U. Ogras, R. Ayoub, M. Kishinevsky, and P. Gratz (2014), 'Towards platform level power management in mobile systems'. In: *2014 27th IEEE International System-on-Chip Conference (SOCC)*. pp. 146–151.

Kaggle Inc. (2017), 'The State of Data Science and Machine Learning'. https://www.kaggle.com/surveys/2017. Accessed: 2018-08-08.

Kaliorakis, M., M. Psarakis, N. Foutris, and D. Gizopoulos (2014), 'Accelerated online error detection in many-core microprocessor architectures'. In: *Proc. of IEEE VLSI Test Symp. (VTS)*. pp. 1–6.

Kanduri, A., M.-H. Haghbayan, A. M. Rahmani, P. Liljeberg, A. Jantsch, N. Dutt, and H. Tenhunen (2016), 'Approximation Knob: Power Capping Meets Energy Efficiency'. In: *Proceedings of the 35th International Conference on Computer-Aided Design*. New York, NY, USA, pp. 122:1–122:8, ACM.

Kanduri, A., M. H. Haghbayan, A. M. Rahmani, P. Liljeberg, A. Jantsch, and H. Tenhunen (2015), 'Dark silicon aware runtime mapping for many-core systems: A patterning approach'. In: *2015 33rd IEEE International Conference on Computer Design (ICCD)*.

Kanduri, A., M. H. Haghbayan, A. M. Rahmani, M. Shafique, A. Jantsch, and P. Liljeberg (2018a), 'adBoost: Thermal Aware Performance Boosting through Dark Silicon Patterning'. *IEEE Transactions on Computers*.

Kanduri, A., A. Miele, A. M. Rahmani, P. Liljeberg, C. Bolchini, and N. Dutt (2018b), 'Approximation-Aware Coordinated Power/Performance Management for Heterogeneous Multi-cores'. In: *Proceedings of the 55th Annual Design Automation Conference*. p. 39.

Kao, Y.-H., M. Yang, N. S. Artan, and H. J. Chao (2011), 'CNoC: high-radix clos network-on-chip'. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* pp. 1897–1910.

Kapadia, N. and S. Pasricha (2015), 'VARSHA: Variation and reliability-aware application scheduling with adaptive parallelism in the dark-silicon era'. In: *Proc. of Design, Automation Test in Europe Conf. Exhibition (DATE)*. pp. 1060–1065.

Karl, E., D. Blaauw, D. Sylvester, and T. Mudge (2008), 'Multi-Mechanism Reliability Modeling and Management in Dynamic Systems'. *IEEE Transactions on VLSI Systems* **16**(4), 476–487.

Karpuzcu, U. R., B. Greskamp, and J. Torrellas (2009), 'The BubbleWrap Many-core: Popping Cores for Sequential Acceleration'. In: *Proc. of Intl. Symp. on Microarchitecture (MICRO)*. pp. 447–458.

Kasture, H. and D. Sanchez (2014), 'Ubik: Efficient Cache Sharing with Strict Qos for Latency-critical Workloads'. In: *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. pp. 729–742.

Khoshavi, N., R. A. Ashraf, R. F. DeMara, S. Kiamehr, F. Oboril, and M. B. Tahoori (2017), 'Contemporary CMOS aging mitigation techniques: Survey, taxonomy, and methods'. *Integration, the VLSI Journal* **59**, 10 – 22.

Khudia, D. S., B. Zamirai, M. Samadi, and S. Mahlke (2015), 'Rumba: An online quality management system for approximate computing'. In: *Computer Architecture (ISCA), 2015 ACM/IEEE 42nd Annual International Symposium on*. pp. 554–566.

Kim, J., J. Balfour, and W. Dally (2007), 'Flattened butterfly topology for on-chip networks'. In: *Proceedings of the 40th Annual IEEE/ACM International Symposium on Microarchitecture*. pp. 172–182.

Kim, J., W. J. Dally, S. Scott, and D. Abts (2008), 'Technology-Driven, Highly-Scalable Dragonfly Topology'. In: *Proceedings of the 35th Annual International Symposium on Computer Architecture (ISCA)*. pp. 77–88.

Kim, R. G., W. Choi, Z. Chen, J. R. Doppa, P. P. Pande, D. Marculescu, and R. Marculescu (2017), 'Imitation Learning for Dynamic VFI Control in Large-Scale Manycore Systems'. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*.

Kim, S., D. Chandra, and Y. Solihin (2004), 'Fair cache sharing and partitioning in a chip multiprocessor architecture'. In: *Proceedings of the 13th International Conference on Parallel Architectures and Compilation Techniques*. pp. 111–122.

Kim, T., X. Huang, H. B. Chen, V. Sukharev, and S. X.-D. Tan (2016), 'Learning-based dynamic reliability management for dark silicon processor considering EM effects'. In: *Proc. of Conf. on Design, Automation & Test in Europe (DATE)*. pp. 463–468.

Kim, Y. (2010), 'ATLAS: A scalable and high-performance scheduling algorithm for multiple memory controllers'. In: *High Performance Computer Architecture (HPCA)*.

Kim, Y., M. Papamichael, O. Mutlu, and M. Harchol-Balter (2010), 'Thread Cluster Memory Scheduling: Exploiting Differences in Memory Access Behavior'. In: *2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*. pp. 65–76.

Kirovski, D. and M. Potkonjak (1997), 'System-level Synthesis of Low-power Hard Real-time Systems'. In: *Proceedings of the 34th Annual Design Automation Conference*. pp. 697–702.

Kling, M., S. Misailovic, M. Carbin, and M. Rinard (2012), 'Bolt: on-demand infinite loop escape in unmodified binaries'. *Proceedings of the ACM ...* pp. 431–450.

Kong, J., S. W. Chung, and K. Skadron (2012), 'Recent Thermal Management Techniques for Microprocessors'. *ACM Computing Surveys* **44**(3), 13:1–13:42.

Koren, I. and C. M. Krishna (2007), *Fault-Tolerant Systems*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1st edition.

Koufaty, D., D. Reddy, and S. Hahn (2010), 'Bias Scheduling in Heterogeneous Multi-core Architectures'. In: *Proceedings of the 5th European Conference on Computer Systems*. pp. 125–138.

Krishna, C. M. and Y. H. Lee (2000), 'Voltage-clock-scaling adaptive scheduling techniques for low power in hard real-time systems'. In: *Proceedings Sixth IEEE Real-Time Technology and Applications Symposium. RTAS 2000*. pp. 156–165.

Kulkarni, C., F. Catthoor, and H. D. Man (1998), 'Code transformations for low power caching in embedded multimedia processors'. In: *Proceedings of the First Merged International Parallel Processing Symposium and Symposium on Parallel and Distributed Processing*. pp. 292–297.

Kumar, R., D. M. Tullsen, N. P. Jouppi, and P. Ranganathan (2005), 'Heterogeneous Chip Multiprocessors'. *Computer* pp. 32–38.

Lackey, D. E., P. S. Zuchowski, T. R. Bednar, D. W. Stout, S. W. Gould, and J. M. Cohn (2002), 'Managing power and performance for system-on-chip designs using Voltage Islands'. In: *IEEE/ACM International Conference on Computer Aided Design, 2002. ICCAD 2002.*

LaFrieda, C., E. Ipek, J. F. Martinez, and R. Manohar (2007), 'Utilizing Dynamically Coupled Cores to Form a Resilient Chip Multiprocessor'. In: *Proc. of Intl. Confl. on Dependable Systems and Networks (DSN)*. pp. 317–326.

Laurenzano, M. A., P. Hill, M. Samadi, S. Mahlke, J. Mars, and L. Tang (2016), 'Input Responsiveness: Using Canary Inputs to Dynamically Steer Approximation'. In: *Proceedings of the 37th ACM SIGPLAN Conference on Programming Language Design and Implementation*. New York, NY, USA, pp. 161–176, ACM.

Lee, H., M. Shafique, and M. A. A. Faruque (2018), 'Aging-aware Workload Management on Embedded GPU Under Process Variation'. *IEEE Transactions on Computers* **67**(7), 920–933.

Lee, J. W., M. C. Ng, and K. Asanovic (2008), 'Globally-Synchronized Frames for Guaranteed Quality-of-Service in On-Chip Networks'. In: *Proceedings of the 35th Annual International Symposium on Computer Architecture*. pp. 89–100.

Lee, S., K. Kang, and C. M. Kyung (2015), 'Runtime Thermal Management for 3-D Chip-Multiprocessors With Hybrid SRAM/MRAM L2 Cache'. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* pp. 520–533.

Lee, S. and T. Sakurai (2000), 'Run-time voltage hopping for low-power real-time systems'. In: *Proceedings 37th Design Automation Conference*. pp. 806–809.

Li, B. and K. Nahrstedt (1999), 'A control-based middleware framework for quality-of-service adaptations'. *IEEE journal on selected areas in communications* pp. 1632–1650.

Li, B., L.-S. Peh, L. Zhao, and R. Iyer (2012), 'Dynamic QoS Management for Chip Multiprocessors'. *ACM Trans. Archit. Code Optim.* pp. 17:1–17:29.

Li, B., L. Zhao, R. Iyer, L.-S. Peh, M. Leddige, M. Espig, S. E. Lee, and D. Newell (2011), 'CoQoS: Coordinating QoS-aware shared resources in NoC-based SoCs'. *Journal of Parallel and Distributed Computing* pp. 700–713.

Li, C., W. Luo, S. S. Sapatnekar, and J. Hu (2015), 'Joint Precision Optimization and High Level Synthesis for Approximate Computing'. In: *Proceedings of Design Automation Conference - DAC '15*. New York, NY, USA, pp. 104:1–104:6, ACM.

Li, M., Q.-A. Zeng, and W.-B. Jone (2006), 'DyXY: A Proximity Congestion-aware Deadlock-free Dynamic Routing Method for Network on Chip'. In: *Proceedings of the 43rd Annual Design Automation Conference (DAC)*. pp. 849–852.

Li, Y., S. Ghose, J. Choi, J. Sun, H. Wang, and O. Mutlu (2017), 'Utility-Based Hybrid Memory Management'. In: *2017 IEEE International Conference on Cluster Computing (CLUSTER)*.

Li, Y. and W. Wolf (1997), 'A Task-level Hierarchical Memory Model for System Synthesis of Multiprocessors'. In: *Proceedings of the 34th Annual Design Automation Conference*. pp. 153–156.

Liao, W., L. He, and K. M. Lepak (2005), 'Temperature and supply Voltage aware performance and power modeling at microarchitecture level'. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* pp. 1042–1053.

Liu, D., T. Chen, S. Liu, J. Zhou, S. Zhou, O. Teman, X. Feng, Z. Xuehai, and Y. Chen (2015), 'PuDianNao : A Polyvalent Machine Learning Accelerator'. In: *Proceedings of International Conference on Architectural Support for Programming Languages and Operating Systems*. pp. 369–381.

Liu, J., P. H. Chou, N. Bagherzadeh, and F. Kurdahi (2001), 'Power-aware scheduling under timing constraints for mission-critical embedded systems'. In: *Proceedings of the 38th Design Automation Conference (IEEE Cat. No.01CH37232)*. pp. 840–845.

Liu, Y. and H. Zhu (2010), 'A survey of the research on power management techniques for high-performance systems'. *Software: Practice and Experience* **40**(11), 943–964.

Lo, D., L. Cheng, R. Govindaraju, L. A. Barroso, and C. Kozyrakis (2014), 'Towards Energy Proportionality for Large-scale Latency-critical Workloads'. In: *Proceeding of the 41st Annual International Symposium on Computer Architecuture (ISCA)*. pp. 301–312.

Lo, D., L. Cheng, R. Govindaraju, P. Ranganathan, and C. Kozyrakis (2015), 'Heracles: Improving Resource Efficiency at Scale'. In: *Proceedings of the 42Nd Annual International Symposium on Computer Architecture (ISCA)*. pp. 450–462.

Lo, W. H., K. z. Liang, and T. Hwang (2016), 'Thermal-aware dynamic page allocation policy by future access patterns for Hybrid Memory Cube (HMC)'. In: *2016 Design, Automation Test in Europe Conference Exhibition (DATE)*. pp. 1084–1089.

Lorch, J. R. and A. J. Smith (2004), 'PACE: a new approach to dynamic voltage scaling'. *IEEE Transactions on Computers* pp. 856–869.

Luo, J. and N. K. Jha (2001), 'Battery-aware static scheduling for distributed real-time embedded systems'. In: *Proceedings of the 38th Design Automation Conference (IEEE Cat. No.01CH37232)*. pp. 444–449.

Lysne, O., T. Skeie, S.-A. Reinemo, and I. Theiss (2006), 'Layered routing in irregular networks'. *IEEE Transactions on Parallel and Distributed Systems* **17**(1), 51–65.

Ma, K. and X. Wang (2012), 'PGCapping: Exploiting Power Gating for Power Capping and Core Lifetime Balancing in CMPs'. In: *Proc. of Intl. Conf. on Parallel Architectures and Compilation Techniques (PACT)*. pp. 13–22.

Ma, S., N. Enright Jerger, and Z. Wang (2011), 'DBAR: An Efficient Routing Algorithm to Support Multiple Concurrent Applications in Networks-on-chip'. In: *Proceedings of the 38th Annual International Symposium on Computer Architecture (ISCA)*. pp. 413–424.

Ma, S., N. E. Jerger, Z. Wang, M. Lai, and L. Huang (2014), 'Holistic routing algorithm design to support workload consolidation in NoCs'. *IEEE Transactions on Computers* pp. 529–542.

Ma, T. C. L. and K. G. Shin (2000), 'A user-customizable energy-adaptive combined static/dynamic scheduler for mobile applications'. In: *Proceedings 21st IEEE Real-Time Systems Symposium*. pp. 227–236.

Ma, Y., T. Chantem, R. P. Dick, and X. S. Hu (2017a), 'Improving System-Level Lifetime Reliability of Multicore Soft Real-Time Systems'. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **25**(6), 1895–1905.

Ma, Y., T. Chantem, R. P. Dick, S. Wang, and X. S. Hu (2017b), 'An on-line framework for improving reliability of real-time systems on "big-little" type MPSoCs'. In: *Proc. of Design, Automation Test in Europe Conf. Exhibition (DATE)*. pp. 446–451.

Maggio, M., H. Hoffmann, M. D. Santambrogio, A. Agarwal, and A. Leva (2011), 'Decision Making in Autonomic Computing Systems: Comparison of Approaches and Techniques'. In: *Proceedings of the 8th ACM International Conference on Autonomic Computing*. pp. 201–204.

Mahajan, D., A. Yazdanbakhsh, J. Park, B. Thwaites, and H. Esmaeilzadeh (2016), 'Towards Statistical Guarantees in Controlling Quality Tradeoffs for Approximate Acceleration'. In: *Proceedings of the 43rd International Symposium on Computer Architecture*. Piscataway, NJ, USA, pp. 66–77, IEEE Press.

Mahmood, A. and E. J. McCluskey (1988), 'Concurrent error detection using watchdog processors-a survey'. *IEEE Transactions on Computers* **37**(2), 160–174.

Maiterth, M., G. Koenig, K. Pedretti, S. Jana, N. Bates, A. Borghesi, D. Montoya, A. Bartolini, and M. Puzovic (2018), 'Energy and Power Aware Job Scheduling and Resource Management: Global Survey — Initial Analysis'. In: *2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. pp. 685–693.

Mak, T., P. Y. Cheung, K.-P. Lam, and W. Luk (2011), 'Adaptive routing in network-on-chips using a dynamic-programming network'. *IEEE Transactions on industrial electronics* pp. 3701–3716.

Manzak, A. and C. Chakrabarti (2000), 'Variable voltage task scheduling for minimizing energy or minimizing power'. In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. pp. 3239–3242 vol.6.

Manzak, A. and C. Chakrabarti (2001), 'Variable voltage task scheduling algorithms for minimizing energy'. In: *Low Power Electronics and Design, International Symposium on, 2001.* pp. 279–282.

Marculescu, R., U. Y. Ogras, L. Peh, N. E. Jerger, and Y. Hoskote (2009), 'Outstanding Research Problems in NoC Design: System, Microarchitecture, and Circuit Perspectives'. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*.

Mars, J., L. Tang, R. Hundt, K. Skadron, and M. L. Soffa (2011), 'Bubble-Up: Increasing Utilization in Modern Warehouse Scale Computers via Sensible Co-locations'. In: *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. pp. 248–259.

Martinez, J. and E. Ipek (2009), 'Dynamic multicore resource management: A machine learning approach'. *IEEE Micro, 29:8–17*.

Meena, J. S., S. M. Sze, U. Chand, and T.-Y. Tseng (2014), 'Overview of emerging nonvolatile memory technologies'. *Nanoscale Research Letters* **9**(526).

Mercati, P., F. Paterna, A. Bartolini, L. Benini, and T. S. Rosing (2017), 'WARM: Workload-Aware Reliability Management in Linux/Android'. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **36**(9), 1557–1570.

Meterelliyoz, M., H. Mahmoodi, and K. Roy (2005), 'A leakage control system for thermal stability during burn-in test'. In: *ITC*.

Millberg, M., E. Nilsson, R. Thid, and A. Jantsch (2004), 'Guaranteed Bandwidth using Looped Containers in Temporally Disjoint Networks within the Nostrum Network on Chip'. In: *Proceedings of the Design Automation and Test Europe Conference (DATE)*.

Misailovic, S., S. Sidiroglou, H. Hoffmann, and M. Rinard (2010), 'Quality of service profiling'. In: *ACM/IEEE 32nd International Conference on Software Engineering*, Vol. 1. pp. 25–34.

Mishra, N., H. Zhang, J. D. Lafferty, and H. Hoffmann (2015), 'A Probabilistic Graphical Model-based Approach for Minimizing Energy Under Performance Constraints'. In: *Proceedings of the Twentieth International Conference on Architectural Support for Programming Languages and Operating Systems*.

Mitra, S., M. Das, A. Banerjee, K. Datta, and T.-Y. Ho (2016), 'A Verification Guided Approach for Selective Program Transformations for Approximate Computing'. In: *Asian Test Symposium (ATS), 2016 IEEE 25th*. pp. 37–42.

Mittal, S. (2016), 'A Survey of Techniques for Approximate Computing'. *ACM Comput. Surv.* **48**(4), 62:1–62:33.

Moazzemi, K., C. Y. Hsieh, and N. Dutt (2016), 'HAMEX: heterogeneous architecture and memory exploration framework'. In: *2016 International Symposium on Rapid System Prototyping (RSP)*.

Moreau, T., F. Augusto, P. Howe, A. Alaghi, and L. Ceze (2017), 'Exploiting quality-energy tradeoffs with arbitrary quantization: special session paper'. In: *Proceedings of the Twelfth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis Companion*. p. 30.

Moreau, T., W. M, J. Nelson, A. Sampson, H. Esmaeilzadeh, L. Ceze, and M. Oskin (2015), 'SNNAP : Approximate Computing on Programmable SoCs via Neural Acceleration'. In: *International Symposium on High-Performance Computer Architecture (HPCA)*.

Moscibroda, T. and B. G. Zorn (2011), 'Flikker : Saving DRAM Refresh-power through Critical Data Partitioning'. In: *Proceedings of International conference on Architectural support for programming languages and operating systems - ASPLOS '11*. pp. 213–224.

Muck, T. R., B. Donyanavard, K. Moazzemi, A. M. Rahmani, A. Jantsch, and N. D. Dutt (2018), 'Design Methodology for Responsive and Robust MIMO Control of Heterogeneous Multicores'. *IEEE Transactions on Multi-Scale Computing Systems*.

Mück, T. R., Z. Ghaderi, N. D. Dutt, and E. Bozorgzadeh (2017), 'Exploiting Heterogeneity for Aging-Aware Load Balancing in Mobile Platforms'. *IEEE Transactions on Multi-Scale Computing Systems* **3**(1), 25–35.

Mukherjee, S. S., M. Kontz, and S. K. Reinhardt (2002), 'Detailed design and evaluation of redundant multi-threading alternatives'. In: *Proc. of Intl. Symp. on Computer Architecture (ISCA)*. pp. 99–110.

Murali, S., G. De Micheli, G. De Micheli, and G. De Micheli (2004), 'SUNMAP: a tool for automatic topology selection and generation for NoCs'. In: *Proceedings of the 41st annual Design Automation Conference*. pp. 914–919.

Muralidhara, S. P., L. Subramanian, O. Mutlu, M. Kandemir, and T. Moscibroda (2011), 'Reducing Memory Interference in Multicore Systems via Application-aware Memory Channel Partitioning'. In: *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*. pp. 374–385.

Murray, J., R. Kim, P. Wettin, P. P. Pande, and B. Shirazi (2014), 'Performance Evaluation of Congestion-Aware Routing with DVFS on a Millimeter-Wave Small-World Wireless NoC'. *J. Emerg. Technol. Comput. Syst.* **11**(2).

Mushtaq, H., Z. Al-Ars, and K. Bertels (2011), 'Survey of fault tolerance techniques for shared memory multicore/multiprocessor systems'. In: *Proc. of Intl. Design and Test Workshop (IDT)*. pp. 12–17.

Nair, R. (2015), 'Big data needs approximate computing: technical perspective'. *Communications of the ACM* **58**(1), 104–104.

Naithani, A., S. Eyerman, and L. Eeckhout (2017), 'Reliability-Aware Scheduling on Heterogeneous Multicore Processors'. In: *Proc. of IEEE Intl. Symp. on High Performance Computer Architecture (HPCA)*. pp. 397–408.

Naithani, A., S. Eyerman, and L. Eeckhout (2018), 'Optimizing Soft Error Reliability Through Scheduling on Heterogeneous Multicore Processors'. *IEEE Transactions on Computers* **67**(6), 830–846.

Nakai, M., S. Akui, K. Seno, T. Meguro, T. Seki, T. Kondo, A. Hashiguchi, H. Kawahara, K. Kumano, and M. Shimura (2005), 'Dynamic voltage and frequency management for a low-power embedded microprocessor'. *IEEE Journal of Solid-State Circuits* pp. 28–35.

Navada, S., N. K. Choudhary, S. V. Wadhavkar, and E. Rotenberg (2013), 'A Unified View of Non-monotonic Core Selection and Application Steering in Heterogeneous Chip Multiprocessors'. In: *Proceedings of the 22Nd International Conference on Parallel Architectures and Compilation Techniques (PACT)*. pp. 133–144.

Nazari, A., N. Sehatbakhsh, M. Alam, A. Zajic, and M. Prvulovic (2017), 'EDDIE: EM-based detection of deviations in program execution'. In: *2017 ACM/IEEE 44th Annual International Symposium on Computer Architecture (ISCA)*.

Neuwirth, S., F. Wang, S. Oral, and U. Bruening (2017), 'Automatic and Transparent Resource Contention Mitigation for Improving Large-Scale Parallel File System Performance'. In: *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*. pp. 604–613.

Neuwirth, S., F. Wang, S. Oral, S. Vazhkudai, J. Rogers, and U. Bruening (2016), 'Using Balanced Data Placement to Address I/O Contention in Production Environments'. In: *2016 28th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*. pp. 9–17.

Nielsen, L. S., C. Niessen, J. Sparso, and K. van Berkel (1994), 'Low-power operation using self-timed circuits and adaptive scaling of the supply voltage'. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* pp. 391–397.

Nilsson, E., M. Millberg, J. Öberg, and A. Jantsch (2003), 'Load distribution with the Proximity Congestion Awareness in a Network on Chip'. In: *Proceedings of the Design Automation and Test Europe (DATE)*. pp. 1126–1127.

Ogras, U. Y. and R. Marculescu (2008), 'Analysis and Optimization of Prediction-based Flow Control in Networks-on-chip'. *ACM Trans. Des. Autom. Electron. Syst.* **13**(1), 11:1–11:28.

Ogras, U. Y., R. Marculescu, and D. Marculescu (2008), 'Variation-adaptive feedback control for networks-on-chip with multiple clock domains'. In: *2008 45th ACM/IEEE Design Automation Conference*.

Ogras, U. Y., R. Marculescu, D. Marculescu, and E. G. Jung (2009), 'Design and Management of Voltage-Frequency Island Partitioned Networks-on-Chip'. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*.

Ouyang, J. and Y. Xie (2010), 'LOFT: A high performance network-on-chip providing quality-of-service support'. In: *Microarchitecture (MICRO), 2010 43rd Annual IEEE/ACM International Symposium on*. pp. 409–420.

Palomino, D., M. Shafique, A. Susin, and J. Henkel (2016), 'Thermal optimization using adaptive approximate computing for video coding'. In: *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2016*. pp. 1207–1212.

Papazoglou, M. P. and D. Georgakopoulos (2003), 'Introduction: Service-oriented Computing'. *Commun. ACM* pp. 24–28.

Park, J., E. Amaro, D. Mahajan, B. Thwaites, and H. Esmaeilzadeh (2016), 'AxGames: Towards Crowdsourcing Quality Target Determination in Approximate Computing'. In: *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*. New York, NY, USA, pp. 623–636, ACM.

Parloff, R. (2016), 'WHY DEEP LEARNING IS SUDDENLY CHANGING YOUR LIFE'. *Fortune*.

Passos, R. M., J. A. Nacif, R. A. F. Mini, A. A. F. Loureiro, A. O. Fernandes, and C. N. Coelho (2006), 'System-level Dynamic Power Management Techniques for Communication Intensive Devices'. In: *2006 IFIP International Conference on Very Large Scale Integration*. pp. 373–378.

Pathania, A., Q. Jiao, A. Prakash, and T. Mitra (2014), 'Integrated CPU-GPU Power Management for 3D Mobile Games'. In: *Proceedings of the 51st Annual Design Automation Conference*. New York, NY, USA, pp. 40:1–40:6, ACM.

Petrucci, V., M. A. Laurenzano, J. Doherty, Y. Zhang, D. Mosse, J. Mars, and L. Tang (2015), 'Octopus-man: Qos-driven task management for heterogeneous multicores in warehouse-scale computers'. In: *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on*. pp. 246–258.

Pillai, P. and K. G. Shin (2001), 'Real-time Dynamic Voltage Scaling for Low-power Embedded Operating Systems'. In: *Proceedings of the Eighteenth ACM Symposium on Operating Systems Principles*. pp. 89–102.

Pritchett, T. and M. Thottethodi (2010), 'SieveStore: A Highly-selective, Ensemble-level Disk Cache for Cost-performance'. *SIGARCH Comput. Archit. News* pp. 163–174.

Psarakis, M., D. Gizopoulos, E. Sanchez, and M. S. Reorda (2010), 'Microprocessor Software-Based Self-Testing'. *IEEE Design & Test of Computers* **27**(3), 4–19.

Psarakis, M., A. Vavousis, C. Bolchini, and A. Miele (2014), 'Design and implementation of a self-healing processor on SRAM-based FPGAs'. In: *Proc. of IEEE Intl. Symp. on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT)*. pp. 165–170.

Pumma, S., M. Si, W. Feng, and P. Balaji (2017), 'Parallel I/O Optimizations for Scalable Deep Learning'. In: *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*.

Qiao, F., N. Zhou, Y. Chen, and H. Yang (2015), 'Approximate Computing in Chrominance Cache for Image/Video Processing'. *2015 IEEE International Conference on Multimedia Big Data* pp. 180–183.

Qiu, Q. and M. Pedram (1999), 'Dynamic power management based on continuous-time Markov decision processes'. In: *Proceedings 1999 Design Automation Conference (Cat. No. 99CH36361)*. pp. 555–561.

Qu, G., D. Kirovski, M. Potkonjak, and M. B. Srivastava (1999), 'Energy minimization of system pipelines using multiple voltages'. In: *Circuits and Systems, 1999. ISCAS '99. Proceedings of the 1999 IEEE International Symposium on*. pp. 362–365 vol.1.

Radetzki, M., C. Feng, X. Zhao, and A. Jantsch (2013), 'Methods for Fault Tolerance in Networks-on-Chip '. *ACM Computing Surveys* **46**(1), 8:1–8:38.

Raha, A., S. Venkataramani, V. Raghunathan, and A. Raghunathan (2015), 'Quality Configurable Reduce-and-rank for Energy Efficient Approximate Computing'. In: *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*. San Jose, CA, USA, pp. 665–670, EDA Consortium.

Rahmani, A., P. Liljeberg, A. Hemani, A. Jantsch, and H. Tenhunen (2017a), *The Dark Side of Silicon*. Springer.

Rahmani, A. M., B. Donyanavard, T. Mück, K. Moazzemi, A. Jantsch, O. Mutlu, and N. Dutt (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*. pp. 169–183.

Rahmani, A. M., M. H. Haghbayan, A. Kanduri, A. Y. Weldezion, P. Liljeberg, J. Plosila, A. Jantsch, and H. Tenhunen (2015), 'Dynamic power management for many-core platforms in the dark silicon era: A multi-objective control approach'. In: *2015 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*.

Rahmani, A. M., A. Jantsch, and N. Dutt (2017b), 'HDGM: Hierarchical Dynamic Goal Management for Many-Core Resource Allocation'. *IEEE Embedded Systems letters*.

Rahmani, A. M., K. Latif, P. Liljeberg, J. Plosila, and H. Tenhunen (2010), 'Research and practices on 3D networks-on-chip architectures'. In: *NORCHIP 2010*. pp. 1–6.

Ranganathan, P. and N. Jouppi (2005), 'Enterprise IT trends and implications for architecture research'. In: *High-Performance Computer Architecture, 2005. HPCA-11. 11th International Symposium on*. pp. 253–256.

Rantala, V., T. Lehtonen, P. Liljeberg, and J. Plosila (2008), 'Hybrid NoC with traffic monitoring and adaptive routing for future 3D integrated chips'. *Diagnostic Services in Network-on-Chips* p. 11.

Rehman, S., K. H. Chen, F. Kriebel, A. Toma, M. Shafique, J. J. Chen, and J. Henkel (2016), 'Cross-Layer Software Dependability on Unreliable Hardware'. *IEEE Transactions on Computers* **65**(1), 80–94.

Rinard, M. (2006), 'Probabilistic Accuracy Bounds for Fault-Tolerant Computations that Discard Tasks'. In: *Proceedings of International Conference on Supercomputing - ICS '06*. pp. 324–334.

Rinard, M. C. (2007), 'Using Early Phase Termination to Eliminate Load Imbalances at Barrier Synchronization Points'. *SIGPLAN Not.* **42**(10), 369–386.

Rotem, E., A. Naveh, A. Ananthakrishnan, E. Weissmann, and D. Rajwan (2012), 'Power-Management Architecture of the Intel Microarchitecture Code-Named Sandy Bridge'. *IEEE Micro* **32**(2), 20–27.

Rozo, L., A. M. Landwehr, Y. Zheng, C. Yang, and G. Gao (2018), 'Reliability-Aware Runtime Adaption Through a Statically Generated Task Schedule'. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **26**(1), 11–22.

Rusu, C., A. Ferreira, C. Scordino, and A. Watson (2006), 'Energy-Efficient Real-Time Heterogeneous Server Clusters'. In: *12th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS'06)*. pp. 418–428.

Saez, J. C., A. Fedorova, D. Koufaty, and M. Prieto (2012), 'Leveraging Core Specialization via OS Scheduling to Improve Performance on Asymmetric Multicore Systems'. *ACM Trans. Comput. Syst.* pp. 6:1–6:38.

Saez, J. C., M. Prieto, A. Fedorova, and S. Blagodurov (2010), 'A Comprehensive Scheduler for Asymmetric Multicore Systems'. In: *Proceedings of the 5th European Conference on Computer Systems (EuroSys)*. pp. 139–152.

Samadi, M., D. A. Jamshidi, J. Lee, and S. Mahlke (2014), 'Paraprox : Pattern-Based Approximation for Data Parallel Applications'. In: *Proceedings of International conference on Architectural support for programming languages and operating systems - ASPLOS '14*. pp. 35–50.

Samadi, M., J. Lee, and D. Jamshidi (2013), 'Sage: Self-tuning approximation for graphics engines'. In: *Proceedings of IEEE/ACM International Symposium on Microarchitecture - MICRO '13*.

Samman, F. A., T. Hollstein, and M. Glesner (2013), 'Runtime contention and bandwidth-aware adaptive routing selection strategies for networks-on-chip'. *IEEE Transactions on Parallel and Distributed Systems* pp. 1411–1421.

Sampson, A., W. Dietl, E. Fortuna, D. Gnanapragasam, L. Ceze, and D. Grossman (2011), '{EnerJ}: Approximate Data Types for Safe and General Low-power Computation'. *Proceedings of the 32nd {ACM} {SIGPLAN} Conference on Programming Language Design and Implementation* pp. 164–174.

San Miguel, J. and M. Badr (2014), 'Load Value Approximation'. In: *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*.

Sanchez, D. and C. Kozyrakis (2011), 'Vantage: Scalable and Efficient Fine-grain Cache Partitioning'. In: *Proceedings of the 38th Annual International Symposium on Computer Architecture*. New York, NY, USA, pp. 57–68, ACM.

Saputra, H., M. Kandemir, N. Vijaykrishnan, M. J. Irwin, J. S. Hu, C.-H. Hsu, and U. Kremer (2002), 'Energy-conscious compilation based on voltage scaling'. *ACM SIGPLAN Notices* **37**(7), 2.

Scolari, A., F. Sironi, D. Sciuto, and M. D. Santambrogio (2014), 'A Survey on Recent Hardware and Software-Level Cache Management Techniques'. In: *2014 IEEE International Symposium on Parallel and Distributed Processing with Applications*. pp. 242–247.

Scott, S., D. Abts, J. Kim, and W. J. Dally (2006), 'The blackwidow high-radix clos network'. *ACM SIGARCH Computer Architecture News* pp. 16–28.

Sehatbakhsh, N., A. Nazari, A. Zajic, and M. Prvulovic (2016), 'Spectral profiling: Observer-effect-free profiling by monitoring EM emanations'. In: *2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*.

Semiconductor Industry Association et al. (2011), 'International Technology Roadmap for Semiconductors'. http://www.itrs2.net/.

Seo, D., A. Ali, W.-T. Lim, N. Rafique, and M. Thottethodi (2005), 'Near-Optimal Worst-Case Throughput Routing for Two-Dimensional Mesh Networks'. In: *Proceedings of the 32nd Annual International Symposium on Computer Architecture (ISCA)*. pp. 432–443.

Shafique, M. and S. Garg (2017), 'Computing in the Dark Silicon Era: Current Trends and Research Challenges'. *IEEE Design Test* **34**(2), 8–23.

Shafique, M., S. Garg, T. Mitra, S. Parameswaran, and J. Henkel (2014), 'Dark silicon as a challenge for hardware/software co-design: Invited special session paper'. In: *Proc. of ACM International Conference on Hardware/Software Codesign and System Synthesis*. p. 13.

Shafique, M., B. Vogel, and J. Henkel (2013), 'Self-adaptive hybrid Dynamic Power Management for many-core systems'. In: *2013 Design, Automation Test in Europe Conference Exhibition (DATE)*. pp. 51–56.

Shahosseini, S., K. Moazzemi, A. M. Rahmani, and N. Dutt (2017), 'Dependability evaluation of SISO control-theoretic power managers for processor architectures'. In: *2017 IEEE Nordic Circuits and Systems Conference (NORCAS): NORCHIP and International Symposium of System-on-Chip (SoC)*.

Shamsa, E., A. Kanduri, A. M. Rahmani, P. Liljeberg, A. Jantsch, and N. Dutt (2018), 'Goal Formulation: Abstracting Dynamic Objectives for Efficient On-chip Resource Allocation'. In: *2018 IEEE Nordic Circuits and Systems Conference (NORCAS): NORCHIP and International Symposium of System-on-Chip (SoC)*. pp. 1–4.

Shamsa, E., A. Kanduri, A. M. Rahmani, P. Liljeberg, A. Jantsch, and N. Dutt (2019), 'Goal-Driven Autonomy for Efficient On-chip Resource Management: Transforming Objectives to Goals'. In: *Proc. of Conf. on Design, Automation Test in Europe (DATE)*.

Shang, L., L.-S. Peh, and N. K. Jha (2003), 'Dynamic voltage scaling with links for power optimization of interconnection networks'. In: *The Ninth International Symposium on High-Performance Computer Architecture, 2003. HPCA-9 2003. Proceedings*. pp. 91–102.

Sharifi, A., S. Srikantaiah, A. K. Mishra, M. Kandemir, and C. R. Das (2011), 'METE: Meeting End-to-end QoS in Multicores Through System-wide Resource Management'. *SIGMETRICS Perform. Eval. Rev.* pp. 13–24.

Shelepov, D., J. C. Saez Alcaide, S. Jeffery, A. Fedorova, N. Perez, Z. F. Huang, S. Blagodurov, and V. Kumar (2009), 'HASS: a scheduler for heterogeneous multicore systems'. *ACM SIGOPS Operating Systems Review* pp. 66–75.

Shin, Y. and K. Choi (1999), 'Power conscious fixed priority scheduling for hard real-time systems'. In: *Proceedings 1999 Design Automation Conference (Cat. No. 99CH36361)*. pp. 134–139.

Shojaei, H., A. Ghamarian, T. Basten, M. Geilen, S. Stuijk, and R. Hoes (2009), 'A parameterized compositional multi-dimensional multiple-choice knapsack heuristic for CMP run-time management'. In: *Design Automation Conference, 2009. DAC'09. 46th ACM/IEEE*. pp. 917–922.

Shoushtari, M., A. BanaiyanMofrad, and N. Dutt (2015), 'Exploiting Partially-Forgetful Memories for Approximate Computing'. *IEEE Embedded Systems Letters* **7**(1), 19–22.

Shye, A., J. Blomstedt, T. Moseley, V. J. Reddi, and D. A. Connors (2009), 'PLR: A Software Approach to Transient Fault Tolerance for Multicore Architectures'. *IEEE Transactions on Dependable and Secure Computing* **6**(2), 135–148.

Sidiroglou, S., S. Misailovic, H. Hoffmann, and M. Rinard (2011), 'Managing performance vs. accuracy trade-offs with loop perforation'. In: *Proceedings of ACM SIGSOFT Symposium and European Conference on Foundations of Software Engineering - SIGSOFT/FSE '11*. pp. 124–134.

Simevski, A., R. Kraemer, and M. Krstic (2014), 'Increasing multiprocessor lifetime by Youngest-First Round-Robin core gating patterns'. In: *Proc. of NASA/ESA Conf. on Adaptive Hardware and Systems (AHS)*. pp. 233–239.

Singh, A., W. J. Dally, A. K. Gupta, and B. Towles (2003), 'GOAL: a load-balanced adaptive routing algorithm for torus networks'. In: *ACM SIGARCH Computer Architecture News*. pp. 194–205.

Singh, A. et al. (2013a), 'Mapping on multi/many-core systems: survey of current and emerging trends'. In: *Proc. of DAC, 2013*. pp. 1:1–1:10.

Singh, A. K., P. Dziurzanski, H. R. Mendis, and L. S. Indrusiak (2017), 'A Survey and Comparative Study of Hard and Soft Real-Time Dynamic Resource Allocation Strategies for Multi-/Many-Core Systems'. *ACM Computing Surveys* **50**(2), 24:1–24:40.

Singh, A. K., A. Kumar, and T. Srikanthan (2011), 'A Hybrid Strategy for Mapping Multiple Throughput-constrained Applications on MPSoCs'. In: *Proceedings of the 14th International Conference on Compilers, Architectures and Synthesis for Embedded Systems*. pp. 175–184.

Singh, A. K., A. Kumar, and T. Srikanthan (2013b), 'Accelerating Throughput-aware Runtime Mapping for Heterogeneous MPSoCs'. *ACM Trans. Des. Autom. Electron. Syst.* pp. 9:1–9:29.

Singh, N. and S. Rao (2014), 'Ensemble Learning for Large-Scale Workload Prediction'. *IEEE Transactions on Emerging Topics in Computing* pp. 149–165.

Singla, G., G. Kaur, A. K. Unver, and U. Y. Ogras (2015), 'Predictive dynamic thermal and power management for heterogeneous mobile platforms'. In: *2015 Design, Automation Test in Europe Conference Exhibition (DATE)*. pp. 960–965.

Skadron, K., M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan (2003), 'Temperature-aware microarchitecture'. In: *30th Annual International Symposium on Computer Architecture, 2003. Proceedings.* pp. 2–13.

Skitsas, M. A., C. A. Nicopoulos, and M. K. Michael (2016), 'DaemonGuard: Enabling O/S-Orchestrated Fine-Grained Software-Based Selective-Testing in Multi-/Many-Core Microprocessors'. *IEEE Transactions on Computers* **65**(5), 1453–1466.

Skitsas, M. A., C. A. Nicopoulos, and M. K. Michael (2018), 'Exploring System Availability During Software-Based Self-Testing of Multi-core CPU'. *Journal of Electronic Testing: Theory and Application* **34**(1), 67–81.

Smolens, J. C., B. T. Gold, B. Falsafi, and J. C. Hoe (2006), 'Reunion: Complexity-Effective Multicore Redundancy'. In: *Proc. of the Intl. Symp. on Microarchitecture (MICRO)*. pp. 223–234.

Song, W. J., S. Mukhopadhyay, and S. Yalamanchili (2015), 'Managing performance-reliability tradeoffs in multicore processors'. In: *Proc. of IEEE Intl. Reliability Physics Symp.* pp. 3C.1.1–3C.1.7.

Soteriou, V. and L. Peh (2007), 'Exploring the Design Space of Self-Regulating Power-Aware On/Off Interconnection Networks'. *IEEE Transactions on Parallel and Distributed Systems* pp. 393–408.

Srinivasan, J., S. V. Adve, P. Bose, and J. A. Rivers (2004), 'The Case for Lifetime Reliability-Aware Microprocessors'. In: *Proc. of Intl. Symp. on Computer Architecture (ISCA)*. pp. 276–287.

Srinivasan, S. T. and A. R. Lebeck (1998), 'Load latency tolerance in dynamically scheduled processors'. In: *Proceedings. 31st Annual ACM/IEEE International Symposium on Microarchitecture*. pp. 148–159.

St Amant, R., A. Yazdanbakhsh, J. Park, B. Thwaites, H. Esmaeilzadeh, A. Hassibi, L. Ceze, and D. Burger (2014), 'General-purpose code acceleration with limited-precision analog computation'. In: *Proceeding of International Symposium on Computer Architecuture - ISCA '14*. pp. 505–516.

Subasi, O., G. Yalcin, F. Zyulkyarov, O. Unsal, and J. Labarta (2017), 'Designing and Modelling Selective Replication for Fault-Tolerant HPC Applications'. In: *Proc. of Intl. Symp. on Cluster, Cloud and Grid Computing (CCGRID)*. pp. 452–457.

Subramanian, L., D. Lee, V. Seshadri, H. Rastogi, and O. Mutlu (2016), 'BLISS: Balancing Performance, Fairness and Complexity in Memory Access Scheduling'. *IEEE Transactions on Parallel and Distributed Systems* pp. 3071–3087.

Subramanian, L., V. Seshadri, A. Ghosh, S. Khan, and O. Mutlu (2015), 'The Application Slowdown Model: Quantifying and Controlling the Impact of Inter-application Interference at Shared Caches and Main Memory'. In: *Proceedings of the 48th International Symposium on Microarchitecture (MICRO)*. pp. 62–75.

Sui, X., A. Lenharth, D. S. Fussell, and K. Pingali (2016), 'Proactive Control of Approximate Programs'. In: *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*. New York, NY, USA, pp. 607–621, ACM.

Suleman, M. A., O. Mutlu, M. K. Qureshi, and Y. N. Patt (2009), 'Accelerating Critical Section Execution with Asymmetric Multi-core Architectures'. In: *Proceedings of the 14th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. pp. 253–264.

Sun, G., C.-W. Chang, and B. Lin (2013), 'A New Worst-Case Throughput Bound for Oblivious Routing in Odd Radix Mesh Network'. *IEEE Computer Architecture Letters* **12**(1), 9–12.

Sun, J., R. Lysecky, K. Shankar, A. Kodi, A. Louri, and J. Roveda (2014), 'Workload Assignment Considering NBTI Degradation in Multicore Systems'. *Journal Emerg. Technol. Comput. Syst.* **10**(1), 4:1–4:22.

Sung, H., J. Min, S. Ha, and H. Eom (2017), 'OMBM: Optimized Memory Bandwidth Management for Ensuring QoS and High Server Utilization'. In: *Foundations and Applications of Self* Systems (FAS* W), 2017 IEEE 2nd International Workshops on*. pp. 269–276.

Tai, J., D. Liu, Z. Yang, X. Zhu, J. Lo, and N. Mi (2017), 'Improving Flash Resource Utilization at Minimal Management Cost in Virtualized Flash-Based Storage Systems'. *IEEE Transactions on Cloud Computing* pp. 537–549.

Tan, C. et al. (2015), 'Approximation-aware scheduling on heterogeneous multi-core architectures'. In: *In Proc. of ASP-DAC,*. pp. 618–623.

Tang, L., J. Mars, and M. L. Soffa (2012), 'Compiling for Niceness: Mitigating Contention for QoS in Warehouse Scale Computers'. In: *Proceedings of the Tenth International Symposium on Code Generation and Optimization (CGO)*. pp. 1–12.

Tang, L., J. Mars, N. Vachharajani, R. Hundt, and M. L. Soffa (2011), 'The impact of memory subsystem resource sharing on datacenter applications'. In: *ACM SIGARCH Computer Architecture News*, Vol. 39(3). pp. 283–294.

Tavakkol, A., M. Sadrosadati, S. Ghose, J. Kim, Y. Luo, Y. Wang, N. M. Ghiasi, L. Orosa, J. Gómez-Luna, and O. Mutlu (2018), 'FLIN: Enabling Fairness and Enhancing Performance in Modern NVMe Solid State Drives'. In: *Proc. of Intl. Symp. on Computer Architecture (ISCA)*. pp. 397–410.

Tedesco, L. P., T. Rosa, F. Clermidy, N. Calazans, and F. G. Moraes (2010), 'Implementation and Evaluation of a Congestion Aware Routing Algorithm for Networks-on-chip'. In: *Proceedings of the 23rd Symposium on Integrated Circuits and System Design (SBCCI)*.

Teodorescu, R. and J. Torrellas (2008), 'Variation-Aware Application Scheduling and Power Management for Chip Multiprocessors'. In: *2008 International Symposium on Computer Architecture*. pp. 363–374.

Tesauro, G. and et.al. (2007), 'Managing Power Consumption and Performance of Computing Systems Using Reinforcement Learning'. In: *Int. Conf. on Neural Information Processing Systems*.

Tesauro, G., N. Jong, R. Das, and M. Bennani (2006), 'A hybrid reinforcement learning approach to autonomic resource allocation'. In: *3rd IEEE International Conference on Autonomic Computing*. pp. 65–73.

Thwaites, B., G. Pekhimenko, A. Yazdanbakhsh, J. Park, G. Mururu, and T. Mowry (2014), 'Rollback-Free Value Prediction with Approximate Loads'. In: *Proceedings of IEEE/ACM International Conference on Parallel Architectures and Compilation Techniques - PACT '14)*.

Tilli, A., A. Bartolini, M. Cacciari, and L. Benini (2015), 'Guaranteed Computational Resprinting via Model-Predictive Control'. *ACM Trans. Embed. Comput. Syst.*

Torng, C., M. Wang, and C. Batten (2016), 'Asymmetry-aware work-stealing runtimes'. In: *Computer Architecture (ISCA), 2016 ACM/IEEE 43rd Annual International Symposium on*. pp. 40–52.

Tziantzioulis, G., A. Gok, S. Faisal, N. Hardavellas, S. Ogrenci-Memik, and S. Parthsarathy (2016), 'Lazy pipelines: Enhancing quality in approximate computing'. In: *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2016*. pp. 1381–1386.

Tzilis, S., I. Sourdis, V. Vasilikos, D. Rodopoulos, and D. Soudris (2016), 'Runtime Management of Adaptive MPSoCs for Graceful Degradation'. In: *Proc. of Intl. Conf. on Compilers, Architectures and Synthesis for Embedded Systems (CASES)*. pp. 5:1–5:10.

Unsal, O. S., R. Ashok, I. Koren, C. M. Krishna, and C. A. Moritz (2001), 'Cool-cache for hot multimedia'. In: *Proceedings. 34th ACM/IEEE International Symposium on Microarchitecture. MICRO-34*. pp. 274–283.

Unsal, O. S. and I. Koren (2003), 'System-level power-aware design techniques in real-time systems'. *Proceedings of the IEEE* pp. 1055–1069.

Unsal, O. S., I. Koren, C. M. Krishna, and C. A. Moritz (2002), 'The minimax cache: an energy-efficient framework for media processors'. In: *Proceedings Eighth International Symposium on High Performance Computer Architecture*. pp. 131–140.

Valiant, L. G. and G. J. Brebner (1981), 'Universal Schemes for Parallel Communication'. In: *Proceedings of the Thirteenth Annual ACM Symposium on Theory of Computing*. New York, NY, USA, pp. 263–277, ACM.

Van Craeynest, K., S. Akram, W. Heirman, A. Jaleel, and L. Eeckhout (2013), 'Fairness-aware Scheduling on single-ISA Heterogeneous Multi-cores'. In: *Proceedings of the 22Nd International Conference on Parallel Architectures and Compilation Techniques (PACT)*. pp. 177–188.

Van Craeynest, K., A. Jaleel, L. Eeckhout, P. Narvaez, and J. Emer (2012), 'Scheduling Heterogeneous Multi-cores Through Performance Impact Estimation (PIE)'. In: *Proceedings of the 39th Annual International Symposium on Computer Architecture*. pp. 213–224.

van den Brand, J. W., C. Ciordas, K. Goossens, and T. Basten (2007), 'Congestion-controlled best-effort communication for networks-on-chip'. In: *Proceedings of the conference on Design, automation and test in Europe*. pp. 948–953.

Vargas, V., P. Ramos, J.-F. Méhaut, and R. Velazco (2018), 'NMR-MPar: A Fault-Tolerance Approach for Multi-Core and Many-Core Processors'. *Applied Sciences* **8**(3).

Vassighi, A. and M. Sachdev (2006), 'Thermal runaway in integrated circuits'. *IEEE Transactions on Device and Materials Reliability*.

Venkataramani, S., V. K. Chippa, S. T. Chakradhar, K. Roy, and A. Raghunathan (2013), 'Quality programmable vector processors for approximate computing'. *46th Annual IEEE/ACM International Symposium* pp. 1–12.

Vogt, L., Y. Chara, H. Ouannani, and M. Nazih (2007), 'Integrated temperature sensor with digital output for SoC power management'. In: *2007 International Conference on Design Technology of Integrated Systems in Nanoscale Era*. pp. 7–12.

Wang, F., S. Oral, S. Gupta, D. Tiwari, and S. S. Vazhkudai (2014), 'Improving large-scale storage system performance via topology-aware and balanced data placement'. In: *2014 20th IEEE International Conference on Parallel and Distributed Systems (ICPADS)*. pp. 656–663.

Wang, L., X. Wang, and T. Mak (2016), 'Adaptive Routing Algorithms for Lifetime Reliability Optimization in Network-on-Chip'. *IEEE Transactions on Computers* **65**(9), 2896–2902.

Wang, T. and Q. Xu (2014), 'On the Simulation of NBTI-Induced Performance Degradation Considering Arbitrary Temperature and Voltage Variations'. In: *Proc. of Design Automation Conf. (DAC)*. pp. 169:1–169:6.

Wang, T., Q. Zhang, and Q. Xu (2017a), 'ApproxQA: a unified quality assurance framework for approximate computing'. In: *Proceedings of the Conference on Design, Automation & Test in Europe*. pp. 254–257.

Wang, X. and J. F. Martínez (2015), 'XChange: A market-based approach to scalable dynamic multi-resource allocation in multicore architectures'. In: *High Performance Computer Architecture (HPCA), 2015 IEEE 21st International Symposium on*. pp. 113–125.

Wang, X. and J. F. Martínez (2016), 'ReBudget: Trading Off Efficiency vs. Fairness in Market-Based Multicore Resource Allocation via Runtime Budget Reassignment'. In: *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems (ISCA)*. pp. 19–32.

Wang, Y., H. Li, and X. Li (2017b), 'Real-Time Meets Approximate Computing: An Elastic CNN Inference Accelerator with Adaptive Trade-off Between QoS and QoR'. In: *Proceedings of the 54th Annual Design Automation Conference 2017*. New York, NY, USA, pp. 33:1–33:6, ACM.

Wells, P. M., K. Chakraborty, and G. S. Sohi (2009), 'Mixed-mode Multicore Reliability'. *SIGARCH Comput. Archit. News* **37**(1), 169–180.

Wentzlaff, D., P. Griffin, H. Hoffmann, L. Bao, B. Edwards, C. Ramey, M. Mattina, C.-C. Miao, J. F. Brown III, and A. Agarwal (2007), 'On-chip interconnection architecture of the tile processor'. *IEEE micro* pp. 15–31.

Winter, J. A., D. H. Albonesi, and C. A. Shoemaker (2010), 'Scalable thread scheduling and global power management for heterogeneous many-core architectures'. In: *2010 19th International Conference on Parallel Architectures and Compilation Techniques (PACT)*.

Wolf, M., S. Bhattacharyya, J. Florence, and A. E. Sapio (2016), 'Power and Thermal Modeling for Communication Systems'. In: *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*. pp. 136–141.

Wu, D., B. M. Al-Hashimi, and M. T. Schmitz (2006), 'Improving Routing Efficiency for Network-on-chip Through Contention-aware Input Selection'. In: *Proceedings of the 2006 Asia and South Pacific Design Automation Conference (ASP-DAC)*. pp. 36–41.

Wu, W. and A. Louri (2016), 'A Methodology for Cognitive NoC Design'. *IEEE Computer Architecture Letters* **15**(1), 1–4.

Wu, Y., C. Lu, and Y. Chen (2016), 'A Survey of Routing Algorithm for Mesh Network-on-Chip'. *Front. Comput. Sci.* pp. 591–601.

Xiang, Y., T. Chantem, R. P. Dick, X. S. Hu, and L. Shang (2010), 'System-level reliability modeling for MPSoCs'. In: *Proc. of Conf. on Hardware/Software Codesign and System Synthesis (CODES)*. pp. 297–306.

Xiang, Y. and S. Pasricha (2015), 'Soft and Hard Reliability-Aware Scheduling for Multicore Embedded Systems with Energy Harvesting'. *IEEE Transactions on Multi-Scale Computing Systems* **1**(4), 220–235.

Xu, C., X. Wu, W. Yin, Q. Xu, N. Jing, X. Liang, and L. Jiang (2017), 'On Quality Trade-off Control for Approximate Computing Using Iterative Training'. In: *Proceedings of the 54th Annual Design Automation Conference 2017*. New York, NY, USA, pp. 52:1–52:6, ACM.

Xu, Q., T. Mytkowicz, and N. S. Kim (2016a), 'Approximate Computing: A Survey'. *IEEE Design & Test* **33**(1), 8–22.

Xu, S., B. Fu, M. Chen, and L. Zhang (2016b), 'Congestion-Aware Adaptive Routing with Quantitative Congestion Information'. In: *High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS), 2016 IEEE 18th International Conference on*. pp. 216–223.

Xue, L., F. Li, M. Kandemir, I. Kolcu, et al. (2006), 'Dynamic partitioning of processing and memory resources in embedded MPSoC architectures'. In: *Proceedings of the conference on Design, automation and test in Europe: Proceedings*. pp. 690–695.

Yamamoto, A. Y. and C. Ababei (2014), 'Unified reliability estimation and management of NoC based chip multiprocessors'. *Microprocessors and Microsystems* **38**(1), 53–63.

Yang, H., A. Breslow, J. Mars, and L. Tang (2013), 'Bubble-flux: Precise Online QoS Management for Increased Utilization in Warehouse Scale Computers'. In: *Proceedings of the 40th Annual International Symposium on Computer Architecture (ISCA)*. pp. 607–618.

Yang, Z., J. Tai, J. Bhimani, J. Wang, N. Mi, and B. Sheng (2016), 'GReM: Dynamic SSD resource allocation in virtualized storage systems with heterogeneous IO workloads'. In: *2016 IEEE 35th International Performance Computing and Communications Conference (IPCCC)*. pp. 1–8.

Yao, F., A. Demers, and S. Shenker (1995), 'A scheduling model for reduced CPU energy'. In: *Proceedings of IEEE 36th Annual Foundations of Computer Science*. pp. 374–382.

Ye, Y., R. West, Z. Cheng, and Y. Li (2014), 'COLORIS: A Dynamic Cache Partitioning System Using Page Coloring'. In: *Proceedings of the 23rd International Conference on Parallel Architectures and Compilation*. New York, NY, USA, pp. 381–392, ACM.

You, D. and K. S. Chung (2014), 'Dynamic power management for embedded processors in system-on-chip designs'. *Electronics Letters* pp. 1309–1310.

Zahedi, S. M. and B. C. Lee (2014), 'REF: Resource Elasticity Fairness with Sharing Incentives for Multiprocessors'. In: *Proceedings of the 19th International Conference on Architectural Support for Programming Languages and Operating Systems (ISCA)*. pp. 145–160.

Zhang, T., J. L. Abellán, A. Joshi, and A. K. Coskun (2014), 'Thermal management of manycore systems with silicon-photonic networks'. In: *2014 Design, Automation Test in Europe Conference Exhibition (DATE)*.

Zhang, X., S. Dwarkadas, and K. Shen (2009), 'Towards practical page coloring-based multicore cache management'. In: *Proceedings of the 4th ACM European conference on Computer systems*. pp. 89–102.

Zhang, X., Y. Zhang, B. R. Childers, and J. Yang (2017), 'DrMP: Mixed Precision-aware DRAM for High Performance Approximate and Precise Computing'. In: *Parallel Architectures and Compilation Techniques (PACT), 2017 26th International Conference on*. pp. 53–63.

Zhang, Y., M. A. Laurenzano, J. Mars, and L. Tang (2014), 'SMiTe: Precise QoS Prediction on Real-System SMT Processors to Improve Utilization in Warehouse Scale Computers'. In: *Proceedings of the 47th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. pp. 406–418.

Zhao, B., H. Aydin, and D. Zhu (2008), 'Reliability-aware Dynamic Voltage Scaling for energy-constrained real-time embedded systems'. In: *Proc. of IEEE Intl. Conf. on Computer Design (ICCD)*. pp. 633–639.

Zhao, Y., J. Rao, and Q. Yi (2016), 'Characterizing and Optimizing the Performance of Multithreaded Programs Under Interference'. In: *Proceedings of the 2016 International Conference on Parallel Architectures and Compilation (PACT)*. pp. 287–297.

Zhou, Y., H. Hoffmann, and D. Wentzlaff (2016), 'CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture'. In: *Proceedings of the 43rd International Symposium on Computer Architecture (ISCA)*. pp. 682–694.

Zhuravlev, S., J. C. Saez, S. Blagodurov, A. Fedorova, and M. Prieto (2012), 'Survey of Scheduling Techniques for Addressing Shared Resources in Multicore Processors'. *ACM Computing Surveys* **45**(1), 4:1–4:28.

Zipf, P., G. Sassatelli, N. Utlu, N. Saint-Jean, P. Benoit, and M. Glesner (2009), 'A Decentralised Task Mapping Approach for Homogeneous Multiprocessor Network-on-chips'. *Int. J. Reconfig. Comput.* pp. 3:1–3:14.

Zong, W., M. O. Agyemen, X. Wang, and T. Maky (2015), 'Unbiased Regional Congestion Aware Selection Function for NoCs'. In: *Proceedings of the 9th International Symposium on Networks-on-Chip (NOCS)*. pp. 19:1–19:8.