Research Vision

My research is in the area of intelligent management of heterogeneous systems, reconciling tradeoffs between application goals and system constraints. I am interested in the way these challenges manifest in the real world, i.e., in the context of cyber-physical systems and internet-of-things networks that must satisfy multiple, often conflicting constraints. I address these challenges with solutions within and across different layers of the system stack, from architecture to software.

As mobile and embedded devices have become ubiquitous, the variety and sophistication of capabilities we expect from our devices has increased. As a result, the number and type of components integrated on a single chip are growing, the application workloads are becoming more diverse, but the systems are still resource-constrained. Often our workload demands and system constraints conflict. For example, we demand high-quality video in conjunction with extended battery life. Such complex interactions require intelligent management of on-chip resources that can coordinate conflicting constraints while adapting to dynamic workloads and operating conditions.

Adaptive Resource Management in Many-core Mobile Systems

In the past, I have explored adaptive policies for managing on-chip memory in systems with unpredictable workloads in order to maximize the energy efficiency of the system [1,2]. While they have produced good results, these heuristic approaches provide ad-hoc adaptivity without any formalism or guarantees. In my thesis, I expand on the need for intelligent adaptation by proposing resource management schemes that formalize adaptivity through the concept of computational self-awareness. My work focuses on three concepts related to computational self-awareness: 1) reflection, 2) self-adaptivity, and 3) self-optimization.

Ongoing Research: Computational Self-awareness for Heterogeneous Multiprocessors

For a system to be reflective, it must have a self-model. In my research, I provide reflection by generating a predictive self-model of a heterogeneous multiprocessor in order to perform energy-efficient task allocation [3]. Reflection is a tool that can be used to provide other self-*X* properties. For a system to be self-adaptive, it must be able to respond to changing goals based on environment or workload. In my research, I provide self-adaptivity by defining a supervisor that adjusts the parameters of distributed low-level resource controllers based on power constraints and quality of service requirements [4]. For a system to be self-optimizing, it must be able to identify changes in optimal operating points based on dynamic or unpredictable workloads. My initial research proposes self-optimization by switching between static DVFS controllers at runtime based on operating point [5]. However, this approach still relies on prior knowledge of the system dynamics to generate a fixed model of the system at design-time. In my latest research, I provide self-optimization in my supervisory resource management hierarchy by deploying hardware reinforcement learners on-device [6]. I am actively working on runtime failure prediction and proaction for the purpose of system life-cycle management, in the context of a larger collaborative effort between UC Irvine, TU Munich, and TU Braunschweig [7].

New Thrusts: Efficiency and Coordination in Cyber-physical Systems, Datacenters, and Internet-of-things Networks

To this point, my research has targeted mobile many-core systems. However, my proposed resource management approach and the principles of computational self-awareness are defined broadly such that they apply to any system that can be decomposed appropriately. The designer must simply provide interfaces to the necessary information and control knobs relevant to the goal. Systems of all scales and types have constraints and goals that require adaptation and coordination. I am broadening my current research by exploring the benefits of self-awareness in three particular contexts:

(1) Cross-layer model adaptivity in cyber-physical systems. High-performance battery-powered cyber-physical systems, e.g., autonomous vehicles, embody extreme divergence between resource constraints and application demands. Consider object detection in an autonomous car: neural networks must be deployed on-vehicle with high accuracy. Such workloads require flexibility to balance accuracy with energy, power, and thermal constraints. Many researchers address tradeoffs through reconfigurability in either the architecture or algorithm, but not both together.

(2) Energy-efficiency in datacenters. Energy-efficient execution of unpredictable and asymmetric workloads in datacenters presents a coordination challenge at true scale. Initial investigation has shown the promise of predictive models for making runtime decisions toward this end. However, existing models are rigid and limited in scope: contributors to power consumption of datacenters range from cooling infrastructure to core processor utilization. I plan to holistically provide energy-efficiency by combining predictive models and distributed decision-making agents. Typical approaches decompose systems into multiple components and distribute agents in order to make local decisions, but require the agents to share data with each other or a centralized entity for coordination. In this manner, existing research (including my own) seeks to manage emergent behavior. I believe this to be prohibitive at true scale, when tens or hundreds of agents make runtime decisions. I am interested in finding ways to guide emergent behavior within a large group of distributed resource management agents.

(3) Resource provisioning in internet-of-things networks. If we expand our view of cyber-physical systems to a network of such systems, scalability challenges manifest with communication cost and reliability as a first-order concern. If we consider internet-of-things networks, we have cross-layer resource provisioning challenges across network layers in addition to between edge devices. A significant amount of research has been done regarding resource provisioning and task mapping in IoT networks, but assumptions are typically made regarding prior knowledge of the expected workload. In line with my general research philosophy and approach, I seek to make such decisions at runtime in an efficient manner without a priori workload exposure using model-free reinforcement learning.

I continue to have positive collaborative experiences with various types of researchers. For me, collaboration with complementary systems and theory researchers is essential for inspiration, as well as maximizing the scope and impact of projects.

Continuing Contribution to Open-Source Projects

In addition to these thrusts, I plan to continue contributing to publicly available open-source code bases that enhance systems research. In the past, I have done significant integration in the gem5 architectural simulator in order to support software-programmable memories (SPMs) [8]. I continue to make contributions to a middleware framework for resource management called MARS [9] that was created and is maintained by my research group at UC Irvine.

My future research will absolutely be a product of my past experience, current ideas, as well as the inspirational researchers surrounding me. I look forward to the change induced by fresh environment and an ever-evolving research landscape. I am excited to continue finding opportunities for coordination of all types of systems and sharing them with colleagues and students.

Sample References

Adaptive Memory Management

[1] Hossein Tajik, Bryan Donyanavard, and Nikil Dutt. “On Detecting and Using Memory Phases in Multimedia Systems.” *ACM/IEEE Symposium on Embedded Systems for Real-Time Multimedia (ESTIMedia)*, 2016.

[2] Majid Shoushtari, Bryan Donyanavard *et al.*, “ShaVe-ICE: Sharing Distributed Virtualized SPMs in Many-Core Embedded Systems.” *ACM Transactions on Embedded Computing Systems (TECS)*, 2018.

Self-Aware Resource Management

[3] Bryan Donyanavard *et al.*, "SPARTA: Runtime task allocation for energy efficient heterogeneous manycores." *International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2016.

[4] Amir M. Rahmani, Bryan Donyanavard *et al.*, “SPECTR: Formal Supervisory Control and Coordination for Many-core Systems Resource Management.” *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2018.

[5] Bryan Donyanavard *et al.*, "Gain scheduled control for nonlinear power management in CMPs." *Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2018.

[6] Bryan Donyanavard *et al.*, “SOSA: Self-Optimizing Learning with Self-Adaptive Control for Hierarchical System-on-Chip Management.” *IEEE/ACM International Symposium on Microarchitecture (MICRO)*, 2019.

Information Processing Factory

[7] Eberle Rambo, Bryan Donyanavard *et al.*, "The Information Processing Factory – A Paradigm for Life Cycle Management of Dependable Systems." *International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, 2019.

Software Artifacts

[8] https://github.com/duttresearchgroup/gem5-spm

[9] https://github.com/duttresearchgroup/MARS

For a complete list of publications, see my CV (https://www.ics.uci.edu/~bdonyana/cv.pdf)