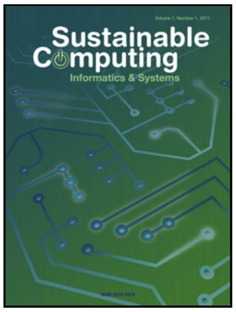

Title: Thermal-Aware Management Techniques for
Cyber-Physical Systems

Author: C.M. Krishna I. Koren

PII: S2210-5379(16)30177-9
DOI: http://dx.doi.org/doi:10.1016/j.suscom.2017.05.005
Reference: SUSCOM 170

To appear in:

Received date: 19-10-2016
Revised date: 4-5-2017
Accepted date: 22-5-2017

Please cite this article as: C.M. Krishna, I. Koren, Thermal-Aware Management Techniques for Cyber-Physical Systems, *<![CDATA[Sustainable Computing: Informatics and Systems]]>* (2017), http://dx.doi.org/10.1016/j.suscom.2017.05.005

Thermal-Aware Management Techniques for Cyber-Physical Systems

C.M. Krishna and I. Koren
Department of Electrical and Computer Engineering
University of Massachusetts, Amherst

***Abstract*—The power density of processors has increased greatly over time. Since elevated temperatures greatly shorten the lifetime of semiconductor devices, thermal management has emerged as a key topic in the design and control of computational platforms. In this paper, we provide a comprehensive yet compact survey of thermal management in cyber-physical systems. Such systems are constrained by the need to meet hard deadlines; this distinguishes them from general-purpose systems and motivates distinctive resource-management approaches.**

I. INTRODUCTION

Elevated temperatures rapidly accelerate chip death. With average chip temperature rising, heating-related failures are a serious concern. Just a few degrees' rise in temperature can halve the mean lifetime of a chip [1]. At the same time, high-speed processing that results in inceased temperatures is essential for many cyber-physical applications. The question of how to take thermal considerations into view while scheduling the real-time workload to meet all hard deadlines is therefore of considerable practical importance. The purpose of real-time scheduling is to meet all task deadlines, especially those of safety-critical tasks. A significant literature on the subject has grown over the past four decades. Introducing thermal constraints expands the problem into another dimension, requiring modification of traditional resource management algorithms.

The purpose of this paper is to provide an overview of the resource-management approaches currently taken in this field. While this paper is largely self-contained in its treatment of the major topics in thermal-aware scheduling, it largely complements two previously published surveys [2], [3]. We focus here on the particular requirements of real-time cyber-physical systems while [2], [3] dealt more broadly with applications as well as aspects of chip design and heat flow modeling.

This paper is organized as follows. In Section II, we provide some technical background; both the impact of high temperature on solid-state devices and the nature of cyber-physical systems are discussed. We then turn in Section III to the optimization criteria used in thermal-aware scheduling and discuss how well they capture the underlying vulnerability of devices to heating and how easy they are to compute. In Section IV, we review ways to measure or estimate the on-chip temperature. Since most thermal scheduling algorithms work by considering the current on-chip temperature, it is important that rapid, low-overhead, and accurate, temperature measurement/estimation techniques be available. Major challenges in doing so arise from the process variation that exists from chip to chip and also the wide variation in the thermal impact of processes from one execution to the next (due to the impact of input data on their execution path). In Section V, we present the heat flow equation that is at the heart of almost all contemporary thermal scheduling research. This is a linear equation and the convenience of linearity is explored at some depth. Also discussed is the issue of model granularity: how fine-grained is the heat flow model used in practice? Thermal control options are described in Section VI; almost all thermal scheduling approaches consist of using one or more of these. Given these options, Section VII covers *reactive* and *proactive* ways of deploying them. Real-time thermal-aware scheduling issues are covered in Section VIII and the paper concludes with a discussion in Section IX.

II. TECHNICAL BACKGROUND

A. Cyber-Physical Workloads

A cyber-physical system (CPS) consists of two major parts: (1) the cyber component, consisting of the controller which runs the control algorithms to compute control inputs, and (2) the physical component, in the form of a physical plant, which is being controlled. A very wide variety of applications are covered under this category: the physical plant may range in size and scope from an implanted medical device to a system for controlling a continent-wide power grid.

The cyber component – the controller – is in the feedback loop of the controlled plant. There are two aspects of its activity which affect the quality of the control it affords. One is the response time of its tasks; the other is the quality of the control algorithm it executes.

The response time affects feedback delay. We know from basic control theory (and, indeed, from common sense) that the greater the feedback delay, the worse the quality of control tends to be. In fact, beyond a certain delay, the controlled plant can actually become unstable. This response time is a function of two things: the capability of the computational platform and the intensity of its workload. Typically, control tasks are run either periodically or sporadically. Periodic tasks are released at regular intervals; sporadic tasks are run whenever triggered by a human operator or the occurrence of some event in the operating environment, under the condition that they will not be invoked more often than a specified number of times per unit time. Standard techniques from digital control theory are used to determine the appropriate rate at which periodic tasks must be dispatched.