Diss. ETH No. 16285

Modeling Multimedia Workloads for Embedded System Design

A dissertation submitted to the

SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZÜRICH

for the degree of

Doctor of Sciences

presented by

ALEXANDER MAKSYAGIN
Dipl. Radio-Eng. MTUCI, Russia

born 15.03.1973
citizen of Russia

accepted on the recommendation of

Prof. Dr. Lothar Thiele, examiner
Prof. Dr. Petru Eles, co-examiner

2005

Abstract

To design a successful computer system, designers need to know characteristics of the computational workload that this system is supposed to process. This knowledge forms the necessary basis for optimizations of the system. In order to use this knowledge in the design process, designers need to characterize the workload using a formal workload model. This model represents an abstraction of the concrete workload and serves as an input to a number of critical design tasks, such as system performance analysis. The quality of the workload model largely determines the quality of the design decisions made based on it.

Coming up with a proper workload model represents a difficult problem in many computer system design contexts. One such context, addressed in this thesis, is system-level design of embedded computers whose main functionality involves real-time processing of media streams (e.g. streams of audio-video data). Of late, there is a growing demand for such computers because they are increasingly being embedded into many electronic products, especially those found in consumer electronics domain, e.g., digital TVs, audio and video players, digital video cameras, advanced set-top boxes, multimedia-enabled mobile phones and a myriad of other electronic devices supporting multimedia applications. To meet high performance requirements and stringent constraints pertaining to cost, size and energy consumption, these embedded computers tend to have complex, heterogeneous multiprocessor architectures. This architectural complexity, coupled with the ever-growing complexity of the multimedia applications themselves, results in a very complex workload behavior and by that poses many challenges to the workload modeling.

In this thesis, we argue that the variability of various parameters of the multimedia workloads is the key property to be captured in a workload model for the embedded systems design. We show that conventional workload models fail to accurately characterize the dynamic nature of the multimedia workloads and, as a result, return overly pessimistic estimations of system performance (especially, if worst-case performance bounds are of interest). As a solution, we propose a novel workload model capable of accurately capturing the workload's dynamic nature. We demonstrate the advantages of the proposed workload model over conventional ways to characterize the workload and develop a number of system-level design methods which use this model. These methods include system-level performance analysis, automatic identification of representative workload sce-

narios for system simulation, design and optimization of resource management policies and a run-time processor rate adaptation strategy for energy-efficient processing of media streams on heterogeneous multiprocessor embedded architectures with stringent memory constraints. We demonstrate the utility of our workload model and evaluate it through a number of case studies involving comparisons to detailed simulation models.

Zusammenfassung

Um ein erfolgreiches Computersystem zu entwerfen, müssen die Entwickler die Rechenanforderungen für das System kennen. Daher ist es notwendig diese Anforderungen mittels eines formalen Auslastungsmodells zu charakterisieren. Dieses Modell repräsentiert eine Abstraktion der konkreten Rechenauslastung und dient beim Entwurf als Eingabe für verschiedene kritische Entwurfsaufgaben. Die Qualität des Auslastungsmodells wirkt sich hierbei direkt auf die Qualität der hierauf basierenden Entwurfsentscheidungen aus.

Oftmals ist es schwierig, ein geeignetes Modell für die Auslastung von Computersystemen in verschiedenen Einsatzgebieten zu finden. Ein solches Gebiet, mit welchem sich auch diese Arbeit befasst, ist der Systementwurf von eingebetteten Computern, deren Hauptfunktion die Echtzeit-Verarbeitung von Media-Datenströmen beinhaltet (z.B. Datenströme von Audio- und Video-Daten). In letzter Zeit ist die Nachfrage nach solchen Computern stark gewachsen, da sie zunehmend in den meisten elektronischen Produkten verwendet werden. Besonders im Unterhaltungselektroniksbereich finden sich viele Beispiele wie digitale Fernseher, Audio- und Video-Recorder, digitale Videokameras, Digitalempfänger, Multimedia-Mobiltelefone und andere elektronische Geräte, die Multimedia-Anwendungen unterstützen. Um die hohen Ansprüche an die Leistung eines solchen Systems zu erfüllen, gleichzeitig aber die Budgets bezüglich Kosten, Grösse und Energieverbrauch nicht zu sprengen, werden diese eingebetteten Computer als komplexe, heterogene Multiprozessorsysteme entworfen. Die ständig wachsende Komplexität dieser Systeme und der darauf ausgeführten Multimedia-Anwendungen führen zu einem sehr komplexen Verhalten der Rechenauslastung, das die Modellierung erschwert.

In dieser Arbeit zeigen wir, dass die Variabilität verschiedener Kenngrössen von der Multimedia-Rechenauslastung die Haupteigenschaft ist, die ein geeignetes Auslastungsmodell umfassen sollte. Wir zeigen weiterhin, dass herkömmliche Auslastungsmodelle diese dynamischen Eigenschaften der Rechenauslastung nicht genau modellieren und demzufolge zu pessimistische Abschätzungen der Systemleistung liefern, besonders dann, wenn die Extremwerte der Leistung von Interesse sind. Als Lösung schlagen wir ein neuartiges Auslastungsmodell vor, welches die dynamischen Eigenschaften der Rechenauslastung gut charakterisieren kann. Wir zeigen die Vorteile des vorgeschlagenen Modells gegenüber herkömmlichen Auslastungsmodellen, und entwickeln einige Systementwurfs-

methoden, welche auf diesem Modell beruhen. Diese Methoden umfassen die Leistunganalyse auf Systemebene, die automatische Identifizierung der charakteristischen Rechenauslastung für die System-Simulation, den Entwurf und die Optimierung der Strategien zum Management der Systemressourcen, und ein Verfahren für die Anpassung der Prozessortaktfrequenz zur Laufzeit für eine energieeffiziente Verarbeitung von Media-Datenströmen auf heterogenen eingebetteten Multiprozessorsystemen mit Speicherplatzeinschränkungen. Wir zeigen den Nutzen unseres Auslastungsmodells und evaluieren es durch eine Reihe von Fallstudien, unterstützt durch detaillierte Simulationen.