

Tytuł rozprawy doktorskiej: Optymalizacja wykonania hybrydowych aplikacji równoległych w heterogenicznych systemach obliczeniowych wysokiej wydajności z uwzględnieniem czasu wykonania i poboru mocy

(Streszczenie)

Wiele istotnych problemów obliczeniowych wymaga wykorzystania systemów obliczeniowych wysokiej wydajności, w których skład wchodzi kilkupoziomowa struktura łącząca coraz większe liczby urządzeń o różnych charakterystykach. Wykorzystanie pełnej mocy takich systemów wymaga programowania aplikacji równoległych, które są hybrydowe w dwóch znaczeniach: potrafią jednocześnie wykorzystać równoległość na wielu poziomach oraz łączą ze sobą interfejsy programistyczne charakterystyczne dla różnych typów urządzeń obliczeniowych.

Głównym celem przetwarzania równoległego jest zwiększenie wydajności przetwarzania, a więc zmniejszenie czasu wykonania aplikacji. Międzynarodowa społeczność skupiona wokół systemów obliczeniowych wysokiej wydajności postawiła sobie za cel zbudowanie do roku 2020 superkomputerów "skali Exa", to znaczy mających możliwość wykonania 10^{18} operacji zmiennoprzecinkowych na sekundę. Jedną z głównych przeszkód stojących na drodze do tego celu jest pobór mocy systemów obliczeniowych przekraczający możliwości dostawy energii. Nowe modele programistyczne i algorytmy uwzględniające to kryterium są jednym z kluczowych pól, na których istotne postępy są konieczne, aby osiągnąć postawiony cel.

Celem rozprawy jest wyodrębnienie ogólnego modelu wykonania hybrydowych aplikacji równoległych w heterogenicznych systemach obliczeniowych wysokiej wydajności będącego syntezą istniejących podejść szczegółowych oraz opracowanie metodologii optymalizacji takiego wykonania aplikacji z uwzględnieniem minimalizacji przeciwstawnych kryteriów czasu wykonania aplikacji i poboru mocy wykorzystywanego sprzętu obliczeniowego. Oba znaczenia hybrydowości aplikacji równoległych wiążą się z mnogością parametrów wykonania o nietrywialnych współzależnościach i wpływie na rozpatrywane kryteria optymalizacyjne. Istotny wpływ na owe kryteria ma także sposób mapowania procesów aplikacji na urządzenia obliczeniowe.

Rozprawa składa się ze wstępu, dwóch rozdziałów literaturowych, trzech rozdziałów empirycznych i podsumowania. We wstępie umotywowano podjęcie badań, sformułowano problem badawczy, przedstawiono zakres, główne oryginalne osiągnięcia, tezy oraz przegląd rozdziałów rozprawy. W rozdziale drugim zawarto przegląd istniejących rozwiązań w zakresie wykonania, modelowania i symulacji hybrydowych aplikacji równoległych oraz opisano przykłady takich aplikacji z różnych dziedzin wraz ze znaczeniem ich hybrydowości. W rozdziale trzecim dokonano krytycznej analizy istniejących podejść do optymalizacji aplikacji równoległych z uwzględnieniem czasu wykonania i poboru mocy, ze szczególnym uwzględnieniem metod optymalizacji wielokryterialnej, zarządzania zasobami obliczeniowymi oraz automatycznego strojenia parametrów wykonania aplikacji.

W rozdziale czwartym opisano pięć praktycznych aplikacji równoległych z różnych dziedzin zastosowań oraz pięć różnorodnych systemów obliczeniowych wykorzystywanych w eksperymentach zawartych w rozprawie. W rozdziale piątym zaproponowano metodologię optymalizacji wykonania hybrydowych aplikacji równoległych w heterogenicznych systemach obliczeniowych wysokiej wydajności, składającą się z określonych kroków wykonania oraz metody symulacji do szybkiej ewaluacji punktów w przeszukiwanej przestrzeni rozwiązań. W rozdziale szóstym przedstawiono wyniki eksperymentów polegających na zastosowaniu poszczególnych proponowanych kroków do wybranych rzeczywistych aplikacji oraz zastosowaniu metodologii optymalizacji w całości do aplikacji treningu głębokiej sieci neuronowej do automatycznego rozpoznawania mowy.

Jak wykazano, wykonanie specyficznych w kontekście proponowanego modelu kroków wstępnej optymalizacji procesów, mapowania procesów, strojenia parametrów i właściwego uruchomienia, pozwala na optymalizację czasu wykonania hybrydowych aplikacji równoległych w heterogenicznych systemach obliczeniowych wysokiej wydajności, a proponowana metoda modelowania i symulacji umożliwiła szybkie i dokładne wyznaczenie zbioru optymalnych rozwiązań w problemie wielokryterialnej optymalizacji ich wykonania z uwzględnieniem czasu wykonania i poboru mocy.

Title of Ph.D. Dissertation: Optimization of hybrid parallel application execution in heterogeneous high performance computing systems considering execution time and power consumption

(Abstract)

Many important computational problems require utilization of high performance computing (HPC) systems that consist of multi-level structures combining higher and higher numbers of devices with various characteristics. Utilizing full power of such systems requires programming parallel applications that are hybrid in two meanings: they can utilize parallelism on multiple levels at the same time and combine together programming interfaces specific for various types of computing devices.

The main goal of parallel processing is increasing the processing performance, and therefore decreasing the application execution time. The international HPC community is targeting development of "Exascale" supercomputers (able to sustain 10^{18} floating point operations per second) by the year 2020. One of the main obstacles to achieving this goal is power consumption of the computing systems that exceeds the energy supply limits. New programming models and algorithms that consider this criterion are one of the key areas where significant progress is necessary in order to achieve the goal.

The goal of the dissertation is to extract a general model of hybrid parallel application execution in heterogeneous HPC systems that is a synthesis of existing specific existing approaches and developing an optimization methodology for such execution aiming for minimization of the contradicting objectives of application execution time and power consumption of the utilized computing hardware. Both meanings of the application hybridity result in multiplicity of execution parameters characterized by nontrivial interdependences and influence on the considered optimization criteria. Mapping of the application processes on computing devices has also a significant impact on these criteria.

The dissertation consists of an Introduction, two theoretical Chapters, three empirical Chapters and a Summary. The Introduction includes motivations for the study, research problem formulation, scope, main contributions, claims and overview of the thesis. Chapter 2 contains a review of existing approaches in the area of executing, modeling and simulation of hybrid parallel applications along with descriptions of examples of such applications and the meaning of their hybridity. Chapter 3 contains a critical analysis of existing approaches to parallel application optimization considering execution time and power consumption with a particular emphasis on multi-objective optimization methods, computing resource management and auto-tuning of application execution parameters.

Chapter 4 describes five real-life parallel applications from various practical fields and five diverse computing systems that were the subject of the experiments included in the dissertation. In Chapter 5, the author proposes an optimization methodology for hybrid parallel application execution in heterogeneous HPC systems that consists of specific execution steps and a simulation method for fast evaluation of points in the solution search space. Chapter 6 presents results of experiments considering applying the consecutive execution steps to chosen real-life applications and using the proposed optimization methodology as a whole to one application of deep neural network training for automatic speech recognition.

As shown in the dissertation, the execution steps specific in the context of the proposed model, including preliminary process optimization, process mapping, parameter tuning and actual execution allow to optimize execution time of hybrid parallel applications in heterogeneous high performance computing systems, while the proposed modeling and simulation method allows for fast and accurate identification of the set of optimal solutions to the problem of multi-objective execution time and power consumption optimization.