Efficiently Energetic Acceleration EEA-Aware
For Scientific Applications of Large-Scale On Heterogeneous Architectures

John A. García. H.1,2, Esteban Hernandez B.1,3, Philippe O. A. Navaux2, C.J. Barrios Hernández2
1,2 jgarciachao@acm.org, 1,3 ejhernandezb@udistrital.edu.co, 2 navaux@inf.ufrgs.br, 1 cbarrios@uis.edu.co

Abstract

Heterogeneous parallel programming has two main problems on large computation systems: one is the increase of power consumption on supercomputers in proportion to the amount of computational resources used to obtain high performance, and the second problem is the underuse of resources by scientific applications with improper distribution of tasks. Select the optimal computational resources and make a good mapping task granularity is the fundamental challenge for building the next generation of Exascale Systems. This research proposes an integrated energy-aware scheme called Efficiently Energetic Acceleration (EEA) for large-scale scientific applications running on heterogeneous architectures. The EEA scheme uses statistical techniques to get GPU power levels to create a GPU power cost function and obtains the computational resource set that maximizes efficiency for a provided workload. The programmer or load balancing framework can use the computational resources obtained to schedule the map parallel task granularity in a static time.

Introduction

The evaluation of performance and power consumption is a main step in the design of applications for large computation systems, such as supercomputers and clusters with nodes that have many cores and multi-GPUs. Researchers must design several experiments for workload characterization to observing the architectural implications of different parameters combinations such as problem size, number of cores per GPU or accelerator, number MPI rank and others, the resulting clock frequency, memory usage, bandwidth and power consumption, which are factors that determine the performance and energy efficiency of their workload implementation.

A key observation from the study by DARP (Technology Challenges in Achieving Exascale Systems) in [1] is that it may be easier to solve the power problem associated with base computation than it will be to reduce the problem of transporting data from one site to another on the same chip, between closely coupled chips in a common package, or between different racks on opposite sides of a large machine room, or on storing data in the aggregate memory hierarchy. Therefore this research designed an integrated scheme called Energy Efficient Acceleration (EEA) as shown in the Figure 1. In which can be used by load balancing frameworks (e.g. StarPU, Ompi5s) for mapping tasks on the computational resources; then it uses a monitor called enerGpu and enerGphi for capturing metrics of performance and power consumption in order to characterize the application and computing architecture; finally the prediction system uses the scheme AEE to predict the combination of computational resources that maximize energy efficiency of applications running on heterogeneous architectures CPU-GPU type.

Step 1: Experimental Procedures and Results

At the first level the global parameters have chosen the workload and computational architecture according to the combination of resources that will be used. The experimental procedures were executed with a set of tests of HPL code variants using different workload and architectural parameters on Cluster nodes GUANE. The experimental procedures was chosen following the fractional factorial design principle proposed by Raj Jain [4].

Step 2: Architectural Characterization

The second level uses the enerGpu monitor in the post-processing for data visualization and statistical characterization of each architecture. In which the Figure 4 shows that the accuracy using the key factors is much more accurate than just using the number of GPUs. In addition, this monitor displays information via sequence data, statistics, histograms and tables showing results in terms of energy efficiency, for each experiment.

Step 3: EEA Prediction System

The third level uses the projected multivariable regression model results to see metrics such as time, performance, power consumption, power consumption and performance per watt, which are used to execute the HPL for each combination of computational resources and calculate the best combination of computational resources, as shown in Figure 5.

EEA-Aware Structure

This poster proposes an integrated energy-aware scheme called Efficiently Energetic Acceleration (EEA) for scientific applications of large-scale on heterogeneous architectures. The EEA structure has a workflow of three steps as shown in Figure 2. In the first step, the data is captured in runtime and executed the enerGpu or enerGphi monitor tool in parallel with the scientific application using different combinations of computational resources, applying the power capping technique for nodes with multi-GPUs. In the second step, the data visualization and statistics characterization used a separate level of enerGpu and enerGphi monitor tool for analysis the key factors by results of each experiment in terms of energy efficiency and estimated the power levels. A deep description of monitoring structure and utilization is present by Garcia John et al. in [2, 3]. Finally using the data collected by monitors cost functions are build and model prediction system running for obtaining the optimal computational resources in a static time to mapping parallel task granularity of scientific applications on heterogeneous architectures.

Figure 1. Organization processors and co-processors in different computational machines.

Figure 2. Efficiently Energetic Acceleration (EEA) Scheme

Figure 3. Efficiently Energetic Acceleration (EEA) Scheme

Figure 4. Efficiently Energetic Acceleration (EEA) Scheme

Figure 5. Efficiently Energetic Acceleration (EEA) Scheme

References