PERFORMANCE EFFICIENT HETEROGENEOUS MULTI CORE SCHEDULING STRATEGY BASED ON GENETIC ALGORITHM

A. S. Radhamani¹ and E. Baburaj²

¹Department of Computer Science and Engineering, Manonmanium Sundaranar University, Tirunelveli, Tamil Nadu, India
²Department of Computer Science and Engineering, Sun College of Engineering and Technology, Nagercoil, Tamil Nadu, India

E-mail: asradhamani@gmail.com

ABSTRACT

Multi-core processors offer a significant performance increase over single core processors. Therefore, they have the potential to enable computation-intensive real-time applications with stringent timing constraints that cannot be met on traditional single-core processors. However, with the number of cores on a single chip continuing to increase, it has been a great challenge to effectively manage the energy efficiency of multicore based systems. Power and temperature management are also two concerns that can increase exponentially with the addition of multiple cores. Design innovations in multicore processor architectures bring new optimization opportunities and challenges in the computing era. System performance will be further enhanced by addressing these challenges. In particular, the process (task) scheduler is one of the critical challenge is garnering great interest. High performance in a heterogeneous multicore system is essential which is achieved by effective scheduling, which remains a challenging problem. Further multi core technology opens research opportunities for energy reduction through efficient scheduling. There may be different hardware and software solutions for the above issue; hardware solutions are based on adjusting dynamic voltage per core, alternatively software approach includes, scheduling task among cores, in heterogeneous environment. Task scheduling in multicore architecture is an extremely difficult problem, because it requires a large combinatorial search space and also precedence constraints between the processes; for the effective utilization of multi core processor system, efficient assignment and scheduling of jobs is more important. Many of the existing algorithms are not focused on task scheduling and core utilization in heterogeneous multi core systems. This paper formulates task scheduling as an optimization problem and the results are compared with the earlier faster scheduler in use. Findings show that, in addition to its optimum solution for large scale problem, the Genetic Algorithm (GA) proposed here fits the heterogeneous multi core parallel scheduling problem of minimizing the completion time as well as in effective core utilization.

Keywords: heterogeneous multi core, task scheduling, genetic algorithm.

1. INTRODUCTION

The board trend in processor development has moved from dual-, tri-, quad-, hexa-, octo-core chips to ones with tens or even hundreds of cores. Besides, promising performance and efficiency gains of multicore processors in processing multimedia, recognition and networking applications can be achieved with the characteristics of multithreading, memory-on-chip, and special-purpose heterogeneous cores. There is also a challenge of improving energy-efficiency by focusing on performance-per-watt with advanced fine-grain or ultra fine-grain power management and dynamic voltage and frequency scaling (i.e., notebook computers and portable media players). However, many issues remain unsolved. In order to utilize a multicore processor at the maximum extent the applications run on the system must be scheduled.

Scheduling, in general, is concerned with allocation of resources to certain tasks to optimize few performance criterions, like the completion time, waiting time or cost production. In scheduling, task scheduling is measured as popular unsolved problem for heterogeneous multicore processors. The importance of scheduling has increased in recent years due to the extravagant development of new process and technologies. Scheduling, in multiprocessor architecture, can be defined as assigning the tasks of precedence constrained task graph onto a set of processors and determine the sequence of execution of the tasks at each processor. This multi core scheduling problem is known to be Non-deterministic Polynomial (NP) complete except in the case of large scale problems.

Because of, the classical algorithms are not dynamic, they cannot achieve the optimal scheduling for all situations, and therefore these algorithms cannot adapt themselves with all situations. This difficulty can be widely solved by Genetic algorithm. In the field artificial intelligence, Genetic Algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of Evolutionary Algorithms (EA), which generate solutions using procedure stimulated by natural evolution, such as inheritance, mutation, selection, and crossover to optimization problems. In most cases the methods are quite effective but not efficient enough, and some important aspects such as the time of transferring data between processes are ignored. Even in some researches, which almost all aspects are considered, only part of the problem is solved by genetic algorithm.

The paper presents a scheduling problem in multi-core systems in a new intelligent method so as to improve the present scheduling for minimizing the total completion time of all process and maximizing their utilization.
2. RELATED WORK

Several research works has been carried out in the past decades, in the heuristic algorithms for job scheduling and generally, since scheduling problems is NP-hard i.e., as the problem size increases the time required to complete the problem also increases exponentially, therefore it is extremely significant and necessary to develop algorithms to fine solution to these problems. Some heuristic methods like branch and bound and prime and search [1], have been proposed earlier to solve this kind of problem. Also, the major set of heuristics for job scheduling onto multiprocessor architectures is based on list scheduling [2-8] [9]. However the time complexity increases exponentially for these conventional methods and becomes excessive for large problems. Then, the approximation schemes are often utilized to find an optimal solution. It has reported in [2, 6] that the critical path list scheduling heuristic is within 5% of the optimal solution 90% of the time when the communication cost is ignored, whereas in rare cases any list scheduling is within 50% of the optimum solution. General task scheduling for cloud applications with multicore processors are implemented in [10].

It is worth to mention that due to sever complexity of scheduling in multi core process; the proposed classical algorithms do not have high performance. Therefore the intelligent methods have been used to find the solution of scheduling problem. In these methods the genetic algorithm is widely used repeatedly [11-14]. The problem based on scheduling of independent jobs on non-identical parallel machines in order to minimize makespan is described in [15]. In [16], new Improved PSO (ImPSO) algorithm is used for solving job scheduling in multiprocessor architecture with the objective of minimizing the job finishing time and waiting time. Multiprocessor task scheduling problem can be stated as finding a schedule for a general task graph to execute on a multiprocessor system so that the schedule length can be minimized. Hybrid Flow Shop Scheduling with Multiprocessor Task (HFSM) problem is known to be NP-hard. In this study [17] an effective parallel greedy algorithm to solve HFSMT problem is presented.

3. PROBLEM DEFINITION

The job scheduling problem of multi processor architecture is scheduling problem to partition the tasks between different cores by accomplishing minimum completion time and minimum waiting time simultaneously. If M different cores M = {ci, i = 1, 2,…,m} and T different tasks T = {tj, j = 1, 2,…,n} are considered in a heterogeneous environment, every core works on different speeds and processing capabilities. It is assumed that the cores C1 is much faster than C2, C3 and so on. If the processing speed is V (i, j) then the execution time has calculated on the basis of size of the tasks by processing speed on different core. Completion Time (or) Processing Time = Task size/Speed of the core.

Maximum Completion Time

\[ C_{\text{max}} = \max_i \left( \sum_{j=1}^{n} \frac{T(i, j)}{V(i, j)} \right) \]  

To evolve good solutions and to implement natural selection, we need a measure for distinguishing good solutions from bad solutions. This can be mathematical model or a computer simulation and the measure is said to be objective function.

Therefore the devised objective function is described below.

\[ \text{Min} \left( C_{\text{max}} \right) = \max_i \left( \sum_{j=1}^{n} \frac{T(i, j)}{V(i, j)} \right) \]  

To maintain the concept of load balancing, effective core utilization is desired. The average utilization is calculated based on the individual performance of the core. The individual core utilization is given by

Utilization = Completion time/ Max.span

An initial population is the set consisting of 15 chromosomes. The chromosomes number, i.e., population size, is one of the important parameters of GA. The efficiency of the algorithm is highly dependent on their “quality” of members of the initial population (chromosomes) which are the parents of the next generations Genetic Algorithm was involved with the number of populations to be 100 and 800 generations. The crossover rate was 0.6 and the mutation rate was 0.01. Randomly the populations were generated and for various trails of the number of cores and task sizes, the completed fitness values of execution time is calculated.

4. CACHE FAIR THREAD SCHEDULING WITH WAIT FREE DATA STRUCTURE (CFTS WF)

At present there are different types of scheduling algorithms are available viz shortest job first, round robin, etc, each with its own merits and demerits. Scheduling algorithms to be of priority based since tasks of some category need an immediate attention while others can be listened later. All threads accessing the wait free data structure can complete its process within a restricted number of steps not considering the behavior of other threads. Therefore when batching CFTS scheduling with wait free data structure where all the threads are given a priority based on the either by the programmer (via system API) or the operating system. There are several scheduling queues are available each with its own priority. Initially the scheduler tries to schedule threads waiting with top priority queue then subsequently queue with low priority. Thus no queue left unattended. In CFTS, if a thread is ready to execute but the only available cores are not in its processor affinity set, then the thread is forced to wait. To satisfy this requirement, Cache Fair Thread Scheduling (CFTS) is integrated with Wait Free data structure to
access the internal scheduler and to achieve optimal efficiency in terms of energy, processing power, and line rate and network traffic stability. Hence CFTS-WF schedules such that maximum throughput, minimum response time, minimum waiting time and utmost CPU consumption is obtained [10].

5. GENETIC ALGORITHMS FOR SCHEDULING

Genetic algorithms are a kind of random search algorithms coming under evolutionary strategies which uses the natural selection and gene mechanism in nature for references. The key concept of genetic algorithm is based on natural genetic rules and it used random search space. GA was formulated by J. Holland with a key advantage of adopting population search and exchanging the information of individuals in population [11, 12 and 14]. The algorithm used to solve scheduling problem is as follows: The flowchart depicting the approach of genetic algorithm is as shown in Figure-1.

Step-1: Initialize the population to start the genetic algorithm Process. For initializing population, it is necessary to input number of processors, task size and core speed.

Step-2: With the generated populations the fitness function is calculated. For the defined problem, the fitness function can be obtained using the equation \( f_i = \alpha x e^{-\beta x E_{\text{max}}} \) where ‘\( \alpha \) and \( \beta \)’ should be set to select an appropriate positive real number for ensuring the fitness of all good individuals to be positive in the solution space. And \( E_{\text{max}} \) is the objective function value.

Step-3: Perform selection process to select the best individual based on the fitness evaluated to participate in the next generation and eliminate the inferior. The job with the minimal execution time and waiting time is the best individual corresponding to a particular generation.

Step-4: Two crossover points are generated uniformly in the mated parents at random, and then the two

Step-5: In this step, mutation operation is performed to further create new offsprings, which is necessary for adding diversity to the solution set. Using turn over operation mutation is performed. Generally, mutation is adopted to avoid loss of information about the individuals the process of evolution. Mutation is performed by setting a random selected job to a random processor.

Step-6: Next stopping criteria must be tested. Stopping condition may be obtaining the best fitness value with minimum execution time and minimum waiting time for the given objective function. If the stopping criterion is satisfied then go to step-7 else go to step-2.

Step-7: From the completed generations optimum result is declared. Stop.

6. SCHEDULING HETEROGENEOUS MULTI CORES USING GENETIC ALGORITHM

High performance in a heterogeneous multicore system can be achieved by an efficient scheduling onto the process which minimizes the execution time and also average of response time. In this method genetic algorithm present the optimal scheduling directly from the final chromosome in final generation. A matrix \( X \) is composed of variables \( x \ (i, j) \) and it has following significance. The row \( i \) of the matrix \( X \) consists of \( n \) cores to process the different task based on the core speed. Rows are called “genes” \( (c_1, c_2, c_3, \ldots, c_n) \) and they represent cores to \( (ci) \) depend on size of the task and speed of the core.

A. Initial population

An initialization operator is applied first. An intermediate population of \( n \) “parent” individuals is created by the initialization operator. Then \( n \) independent removal of an individual from the old population is carried out [18] to produce “parents”. Each individual being removed must be in linear relationship with the fitness of that individual. The average individuals which lie in the above range should have more copies in the new population; while below average individuals should have few to no copies present. The selection operator chooses two members of the present generation to participate in later crossover and mutation operations. Initialization
process raises the issue of fitness function. A requirement of the selection methods is that the probability \( f_i \) of an organism must be the best to be selected.

B. Fitness evaluation

The performance of an individual with respect to the current optimum so that different individuals can be compared to get a numerical value is called the fitness value. There exists a range of solutions for fitness from worst to best. The degree of success in the application of evolutionary algorithms depend significantly on the definition of a fitness that changes neither too rapidly nor too slow wit the design parameters of the optimization problem. Performance evaluation of chromosomes is merely based on fitness function. After the generation of new population, fitness value of each chromosome is calculated \( f_i \). The higher the fitness value, the better the performance of the chromosome (i.e., parent). Hence, parents with higher fitness value have more chances to survive. Because the objective function is to minimize the execution time, fitness values can be obtained using the following function:

\[
f_i = \alpha x e^{-\beta x E_{max}},
\]

where \( \alpha \) and \( \beta \) are positive real number and \( E_{max} \) is the objective function value (execution time) of the chromosome \( k \).

C. Reproduction

Based on roulette-wheel selection, is a genetic operation for selecting potentially useful solutions for recombination. In fitness proportionate selection, as in all selection methods, the fitness function assigns fitness to possible solutions or chromosomes. This fitness level is used to associate probability of selection with each individual chromosome. Selection probability is given below if \( f_i \) is the fitness of individual \( i \) in the population,

\[
P_i = \frac{f_i}{\sum_{j=1}^{N} f_j}
\]

where \( N \) is the number of individuals in the population.

This could be imagined similar to a Roulette wheel in a casino. Based on their fitness value a proportion of the wheel is assigned to each of the possible selections. Normalization could be achieved by dividing the fitness of a selection by the total fitness of all the selections.

Table-1 shows the selection probability for 12 individuals, linear ranking and selective pressure of 2 together with the fitness value. Individual 1 is the fit individual and occupies the largest interval, whereas individual 11 as the second least fit individual has the smallest interval on the line. Individual 12, the least fit interval, has a fitness value of 0 and get no change for reproduction for selecting the mating population the appropriate number of uniformly distributed random numbers (uniform distributed between 0.0 and 1.0) is independently generated.

<table>
<thead>
<tr>
<th>Number of individual</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitness value</td>
<td>2.2</td>
<td>2.0</td>
<td>1.8</td>
<td>1.6</td>
<td>1.4</td>
<td>1.2</td>
<td>1.0</td>
<td>0.8</td>
<td>0.6</td>
<td>0.4</td>
<td>0.2</td>
<td>0.0</td>
</tr>
<tr>
<td>Selection probability</td>
<td>0.2</td>
<td>0.19</td>
<td>0.17</td>
<td>0.15</td>
<td>0.13</td>
<td>0.11</td>
<td>0.09</td>
<td>0.07</td>
<td>0.06</td>
<td>0.03</td>
<td>0.02</td>
<td>0.0</td>
</tr>
</tbody>
</table>

D. Recombination (Crossover) operators

A new generation is formed by (a) selecting, according to fitness values, some of the parents and offspring and (b) rejecting others so as to keep the population size constant. Filter chromosomes have higher probabilities of being selected. Crossover is the major genetic operator which works on two chromosomes at a time and generates offspring by combining both chromosomes features. A simple way to achieve crossover would be to choose a random cut-point and generate the offspring by combining the segment of one parent to the left of cut-point with the segment of the other parent to the right of the cut-point. This method works well with the bit string representation. To a great extent the performance of genetic algorithm depends, based on the performance of the crossover operator used. In our method, crossover operator deals with genes and not chromosomes like most of the applications in literature. Each gene “\( i \)” consists of \( n \) cores to execute the various task on core \( i \), \( \alpha (i, j) \neq 0, j = 1, 2…n \); each chromosome contains \( m \) genes and each one consists in one feasible schedule. Crossover operator described here combines two genes of same chromosome, in proper order to obtain a new chromosome giving a better solution. Therefore crossover operation is carried out by moving one core to another for different task size and core speed to get a better solution. Figure-2 depicts the process of arithmetic cross over for different cores.

![Figure-2. Arithmetic cross over.](image-url)

The first gene to combine, \( g_i \), is the one that indicates the core with maximum execution time:
The core with the minimum processing time of gene \( g_i \) is given by the equation below:

\[
P(i', j') = \min_{i=1}^{m} P(i', j')
\]

8. RESULTS AND DISCUSSIONS

We have used simulation to validate the genetic algorithm. The simulation was performed on simulation environment for parallel program execution that was developed using MATLAB programming language. Genetic Algorithm was invoked with the number of populations to be 100 and 800 generations. The crossover rate was 0.5 and the mutation rate was 0.01 randomly the populations were generated and for various trials of the number of cores and task sizes, the completed fitness values of execution time is calculated. Simulation results of genetic algorithm shows that when the task sizes are larger and close to neighborhood sizes the core utilization is higher than that of the smaller tasks. Figure-1 shows the result of scheduling with GA and CFTS-WF algorithm with respect to task size and Figure-2 shows the result of scheduling with genetic algorithm and fast scheduler (Cache Fair Thread Scheduler with Wait Free data structure) CFTS-WF based on number of processing cores. From Figure-2, it can be observed that as the number of core increases, the completion time has got reduced for GA than CFTS-WF. Figure-3 shows the variations in the task size with respect to number of cores for GA, and CFTS-WF scheduler. The execution time and waiting time is observed based on the number of tasks allocated to each cores. Also it shows the variation in execution time for the assigned size of tasks and speed of the core.

<table>
<thead>
<tr>
<th>Core</th>
<th>Scheduled tasks</th>
<th>Execution time</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.1</td>
<td>Task2</td>
<td>15.5</td>
</tr>
<tr>
<td>C.2</td>
<td>Task3 Task5</td>
<td>14.8</td>
</tr>
<tr>
<td>C.3</td>
<td>Task7 Task6 Task4</td>
<td>16.0</td>
</tr>
<tr>
<td>C.4</td>
<td>Task13 Task8</td>
<td>15.0</td>
</tr>
<tr>
<td>C.5</td>
<td>Task12 Task10 Task20 Task1</td>
<td>15.4</td>
</tr>
<tr>
<td>C.6</td>
<td>Task11 Task17</td>
<td>15.8</td>
</tr>
<tr>
<td>C.7</td>
<td>Task15 Task19</td>
<td>16.2</td>
</tr>
<tr>
<td>C.8</td>
<td>Task14 Task16 Task18 Task9</td>
<td>16.0</td>
</tr>
</tbody>
</table>

8. RESULTS AND DISCUSSIONS

We have used simulation to validate the genetic algorithm. The simulation was performed on simulation environment for parallel program execution that was developed using MATLAB programming language. Genetic Algorithm was invoked with the number of populations to be 100 and 800 generations. The crossover rate was 0.5 and the mutation rate was 0.01 randomly the populations were generated and for various trials of the number of cores and task sizes, the completed fitness values of execution time is calculated. Simulation results of genetic algorithm shows that when the task sizes are larger and close to neighborhood sizes the core utilization is higher than that of the smaller tasks. Figure-1 shows the result of scheduling with GA and CFTS-WF algorithm with respect to task size and Figure-2 shows the result of scheduling with genetic algorithm and fast scheduler (Cache Fair Thread Scheduler with Wait Free data structure) CFTS-WF based on number of processing cores. From Figure-2, it can be observed that as the number of core increases, the completion time has got reduced for GA than CFTS-WF. Figure-3 shows the variations in the task size with respect to number of cores for GA, and CFTS-WF scheduler. The execution time and waiting time is observed based on the number of tasks allocated to each cores. Also it shows the variation in execution time for the assigned size of tasks and speed of the core.

<table>
<thead>
<tr>
<th>Core</th>
<th>Scheduled tasks</th>
<th>Execution time</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.1</td>
<td>Task12</td>
<td>15.0</td>
</tr>
<tr>
<td>C.2</td>
<td>Task13 Task15</td>
<td>14.9</td>
</tr>
<tr>
<td>C.3</td>
<td>Task17 Task16 Task14</td>
<td>16.2</td>
</tr>
<tr>
<td>C.4</td>
<td>Task3 Task18</td>
<td>16.1</td>
</tr>
<tr>
<td>C.5</td>
<td>Task20 Task1 Task2 Task11</td>
<td>15.8</td>
</tr>
<tr>
<td>C.6</td>
<td>Task10 Task7</td>
<td>16.0</td>
</tr>
<tr>
<td>C.7</td>
<td>Task5 Task9</td>
<td>16.0</td>
</tr>
<tr>
<td>C.8</td>
<td>Task4 Task6 Task8 Task9</td>
<td>15.8</td>
</tr>
</tbody>
</table>

Table-2. Scheduling Task with GA
9. CONCLUSIONS

Most of the scheduling problems are NP-hard. Mathematical optimization techniques can give an optimal solution for a reasonably sized problem, however, in the case of a large scale problem, their application is limited. Scheduling schemes without GA are suitable only for small scale problems and no scheduling schemes without GA guarantees optimum result in various problems. Research efforts have therefore concentrated on heuristic approaches. Among these approaches, GA outperforms others in view of its characteristic such as high adaptability, near optimization and easy realization. As many practical job shop and open shop scheduling problems can be simplified as heterogeneous multi core scheduling problems under certain conditions, the same has received a great deal of attention in the academic and engineering circle. There are many applications of GA to solve heterogeneous multi core scheduling problem; but, even though it is a common problem in the industry, only a small number of them deal with heterogeneous multi core systems. We therefore decided to concentrate our research effort on scheduling heterogeneous multi core with different parameters using GA so that better results may be produced.

REFERENCES


