Source To Source Translation in Hybrid High Performance Computing (HPC) Systems

Suman
CSE, H.K.B.K College of Engineering
Bangalore, Karnataka

***Abstract*--Hybrid computing is emerging as a requirement for power efficient system design. In hybrid architectures, more speed up is obtained by overlapping the computations of available computing elements. Therefore there is a need to make these computing elements work together by balancing the workload. In CPU + GPU architecture, GPU handles data parallel work by using large number of cores. While CPU handles serial work. In this paper we are going to study the source to source translation on OPENMP+CUDA architecture and verify the hypothesis that the performance is increased when the time taken by CPU is approximately equal to time taken by GPU. The goal is to reduce communication latency between CPU-GPU and other compute devices. Methodology will be tested against Rodinia 3.0 benchmarks.**

I. INTRODUCTION

High Performance Computing is a process of aggregating computing power in a way that delivers much higher performance than one could get out of a typical desktop computer or workstation in order to solve large problems in science, engineering or business [1]. Due to the increase in demand for power and speed, HPC will likely interest business of all sizes, particularly for transaction processing and data warehouses [3]. The requirement of hybrid architectures is increasing due to the growing trend of using the multiple computing resources as the sole computing resource. In such cases, the performance is strongly affected by the dependence that exists between the parallel code and architecture. The process of allocating the tasks to the processors is often a problem that requires considerable consuming process and it also requires the programmer to have in depth knowledge of the target architecture and also the programming knowledge. Therefore there is a tremendous need for the efficient automatic parallelization tools. In hybrid architectures, the CPU is idle most of the time. In the computations performed by the CPU and GPU, CPU performs the serial portion of the task and the GPU performs the parallel portion of the task. The CPU time and the communication time is more. This had a direct impact on the total execution time. It is possible to divide the work between CPU and GPU by using source to source translation. OpenMP work for CPU and CUDA work for GPU. This paper is structured as follows: in section 2 introduces some issues that motivated this research and the main goals to achieve. Section 3 shows some of the existing automatic parallelization tools and there drawbacks. Section 4 discusses about NAS and Rodinia Benchmarks and observations on NAS Benchmarks obtained through manual approach. Related work is shown in section 5 .Section 6 closes with some conclusions and future research to be done.

II. BACKGROUND AND MOTIVATION

Researchers and programmers have been attracted by parallel programming for many years. Parallel programming is a difficult task that may lead to many run time errors. So some tools or methods facilitating the process are required. Parallel programming can be achieved by a programmer manually or it can be achieved automatically by a compiler. Both these approaches has some advantages and some disadvantages. A high level of optimization can be achieved with the help of automatic parallelization compiler .Careful analysis and fine tuning can take this to another level to produce more optimized code. Programming on hybrid systems is obviously dependent on the type of architecture used and the performance obtained is strongly conditioned by the set of machines performing the computations. This means, in most of the cases, the techniques used on homogeneous architectures must be modified to be applied to the systems that have hybrid architectures [4]. In hybrid architectures, since the task is shared is between the available computing elements, there is a need to make these computing elements work together by balancing the workload so that the performance can be increased. The load can be distributed in many ways. One such way is to divide the data into several chunks and obtain performance measures. Based on the measures obtained assign number of chunks to each type of core [18] .The other way to distribute the load is to run the benchmarks in different compute resources using different balancing configurations and use suitable techniques to predict chunk distribution for newer application.

Traditionally source to source translation was done only for OPENMP or CUDA programming model, not for both at the same time and for the same application. But it can be done for the combination of both OPENMP as well as CUDA at the same time by which it can distribute the work and give more optimization.

The motivation of this research is, the load distribution behavior varies for different workload. There is no standard model to predict the performance. More speedup is

obtained by overlapping the computations of CPU and GPU, was this the best speedup?. The load distribution behavior varies for different workload, what is the impact of this on performance? Our objective is to study the behavior of load distribution for varying workload on OPENMP+CUDA architecture.To verify the hypothesis that the total execution time is minimized when time taken by CPU is approximately equal to the time taken by GPU.

III. EXISTING SYSTEMS

Automatic parallelization is the process of converting sequential code into multi-threaded(parallel) code in order to utilize multiple processors simultaneously in shared memory multiprocessor machine. The goal of automatic parallelization is to relieve programmers from tedious and error-prone manual parallelization process. Though the quality of automatic parallelization has improved in the past several decades, fully automatic parallelization of sequential programs by compilers remains a grand challenge due to its need for complex program analysis and unknown factors(such as input data range) during compilation. The compiler usually conducts two passes of analysis before actual parallelization. The first pass of the compiler performs a data dependence analysis of the loop to determine whether each iteration of the loop can be executed independently. The second pass attempts to justify the parallelization effort by comparing the theoretical execution time of the code after parallelization to the code's sequential execution time. There are number of automatic parallelization tools available which supports different programming languages like FORTRAN and C such as SUIF compiler, Polaris compiler, Rose compiler, Bones compiler, Cetus compiler etc. In this paper we are going to discuss source to source compilers called Cetus, Bones and GCC.

A. Cetus

Cetus is the source to source compiler infrastructure for transformation of programs. Cetus has the ability to represent the given program in symbolic terms. It was created out of the need for the compiler research environment to support the development of interprocedural analysis and some of the parallelization techniques for C, C++ and java programs [22] .The internal representation (IR) in Cetus is visible to the pass writer through an interface called IR-API. In Cetus, it is easy to write source to source transformations and optimization modification. Portability of the infrastructure to a wide variety of platforms will make Cetus useful to larger community.
Cetus drawbacks: Cetus performs poorly, the deficit in number of parallel loops ranges from 10 to 40 percent. In LU's (NAS benchmark) case, Cetus generates fewer parallel loops because it exploits more outer-level parallelism.Cetus detects important parallel loops or their inner-loop parallelism in CG, IS, SP, and art (NAS benchmarks), but fails to parallelize such loops in EP, equake, and ammp.
CETUS tools are able to generate parallel codes, still more efforts are required to make those codes optimum in terms of performance. These tools should try to skip the loops that have smaller execution time. The performance of parallel code will increase when all the threads are mapped to physical cores. For task level parallelization, the task should have optimal size and less dependencies.

B. GCC compiler

GCC stands for "GNU Compiler Collection". GCC is an integrated distribution of the compilers for several major programming languages. These languages includes C, C++, java, Fortran etc. when GCC is invoked, it normally does preprocessing, compilation, assembly and linking. In GCC the process can be stopped at an intermediate state. The output consists of object files output by the assembler. The operands to GCC program are options and file names. GCC is one of the most robust compiler. It generates highly optimized code for variety of architectures. Its open source distribution and continuous updates make it more attractive. GCC was not designed for source to source transformations. Most of its passes operate on lower-level RTL representation.
Disadvantages: GCC does not provide friendly API for the pass writers. GCC's IR uses an ad-hoc type system, which is not reflected in its implementation language. This makes it difficult to the debuggers to provide meaningful information to the user. It requires extensive modifications to support interprocedural analysis across multiple files.

C.Bones compiler

Bones is the source to source compiler. It is written in Ruby programming language. It is based on Algorithmic Skeleton Technique. Algorithmic Skeleton Technique is a technique which revolves around a set of parameterisable skeleton implementation. Each skeleton implementation can be seen as template code for specific algorithm class on target architecture. Programmers are able to generate efficient target code by first identifying a number of lines of the code of certain class, followed by invoking the corresponding skeleton implementation. If no skeleton implementation is available for specific class-architecture combination, it can be manually added. Future algorithms of the same class can then be benefit from reuse of skeleton code.
Drawbacks: performance can still be improved. Code readability can be improved. Programmer's effort is required.

IV. BENCHMARKS

We worked on some of the NAS 2.3 Parallel Benchmarks and some Rodinia 3.0 Benchmarks.

A. NAS Parallel Benchmarks

NAS Parallel Benchmarks are the set of benchmarks used to evaluate the performance of supercomputers. They are

developed and maintained by NASA Advanced Supercomputing (NAS) division. Some of the NAS benchmarks along with there description is listed below.

- MG (MultiGrid) : Approximate the solution to a three-dimensional discrete Poisson equation using the V-cycle multigrid method

- CG (Congugate Gradient) : Estimate the smallest eigenvalue of a large sparse symmetric positive-definite matrix using the inverse iteration with the conjugate gradient method as a subroutine for solving systems of linear equations

- FT(Fast Fourier Transform) : Solve a three-dimensional partial differential equation (PDE) using the fast Fourier transform (FFT)

- IS(Integer Sort) : Sort small integers using the bucket sort

- EP (Embarrisingly Parallel) : Generate independent Gaussian random variates using the Marsaglia polar method

- LU(Lower Upper Symmetric Gausseide) : Solve a synthetic system of nonlinear PDEs using three different algorithms involving block tridiagonal, scalar pentadiagonal and symmetric successive over-relaxation (SSOR) solver kernels, respectively

- BT (Block Tridiagonal) :Tests different parallel I/O techniques

- SP (Scalar Pentadiagonal) : Multiple independent systems of non-diagonally dominant, scalar pentadiagonal equations are solved

The table below shows the amount of time taken by each benchmark with manual parallelization approach.

	BT	**LU**	**CG**	**EP**	**FT**	**MG**	**SP**
1 THREAD	284.17	284.44	3.93	101.54	20.26	6.94	288.33
4 THREADS	146.45	85.93	1.61	12.10	4.37	2.25	99.91
16 THREADS	125.19	50.92	0.87	4.44	4.26	1.73	83.96

TABLE 1: NAS observations

The graph shows the observation of different bench marks BT, LU, CG, EP, FT, MG and SP for 3 set 1, 4 and 16 threads. The performance is observed in the time taken by each benchmark in seconds. One can clearly see the optimization in performance for each bench marks as the number of threads increases.

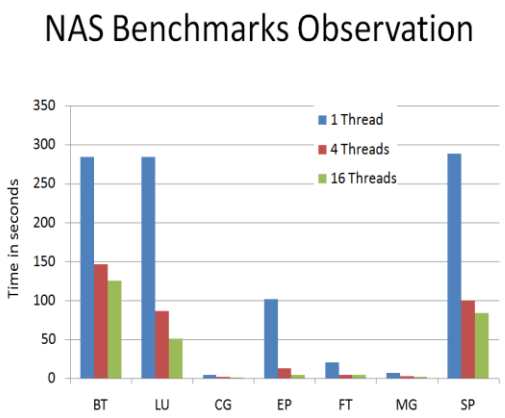

Figure 1 NAS Benchmarks

B. Rodinia Benchmark

The suite consists of four applications and five kernels. They have been parallelized with OpenMP for multicore CPUs and with the CUDA API for GPUs. We use various optimization techniques in the applications and take advantage of various on-chip compute resources. Some of the applications alond with their domains is tabulated below.

Application/Kernel	***Domain***	***Dwarf***
K-means	Data Mining	Dense Linear Algebra
Needleman-Wunsch	Bioinformatics	Dynamic Programming
HotSpot*	Physics Simulation	Structured Grid
BackPropagation*	Pattern Recognition	Unstructured Grid
SRAD	Image Processing	Structured Grid
Leukocyte Tracking	Medical Imaging	Structured Grid
Breadth-First Search*	Graph Algorithms	Graph Traversal
Stream Cluster*	Data Mining	Dense Linear Algebra
Similarity Scores*	Web Mining	MapReduce

TABLE 2: showing applications/kernel of Rodinia benchmarks with domain and dwarf

Rodinia Benchmark [16] uses OPEN MP + CUDA architectures. We are trying to mix both the architectures such that more optimizations are obtained and suitable DR (Distribution Ratio) is identified. The generalized algorithm used is given below

Algorithm

[1] Read input sequential code
[2] Analyze the input and identify CPU and GPU variables
[3] Identify parallelizable portions
[4] Calculate block grid and dimension
[5] Allocate memory to GPU and copy the code on

GPU
[6] Place suitable parallel code on the kernel and GPU code on main function
[7] Make a call to kernel at appropriate place in main
[8] Return the result to CPU
[9] Free memory on GPU and CPU

Basic diagram to show the source to source translation by taking any application of Rodinia benchmarks 3.0. These are the following steps which are used generally.

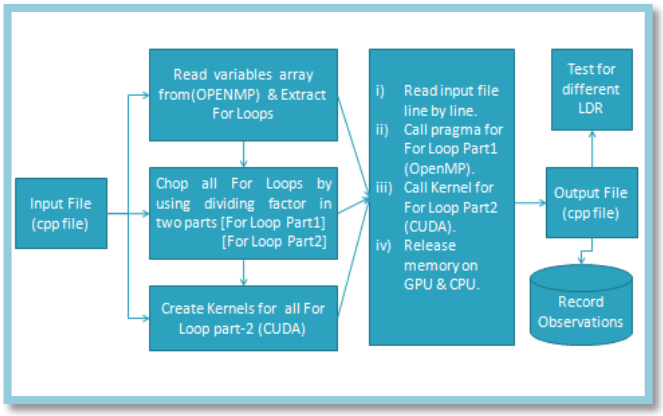

FIG-Source to Source Translation Steps

V. RELATED WORK

Some of the transformations were checked in GCC compiler. Loop optimization is the process of increasing execution speed and reducing the overheads associated of loops. It plays an important role in improving the cache performance and making effective use of parallel processing capabilities. Loop optimizations offer a good opportunity to improve the performance of data parallel applications. These optimizations are usually targeted at improving the granularity, load balance and data locality, while minimizing synchronization and other parallel overheads. Common loop transformations are fission, fusion, interchange, scheduling, software pipelining, unrolling, etc. unrolling duplicates the body of the loop multiple times, in order to decrease the number of times the loop condition is tested and the number of jumps, which may degrade the performance by impairing the instruction pipeline. Loop fission can improve data locality of reference, both on cache being accessed in the loop and the code in the loop's body. Loop fusion attempts to reduce the overhead. Loop interchange improve data locality of reference, depending on the array's layout. Loop transformations found in GCC compiler is tabulated below.

Loop Transformatios	Unrolling	Loop Fission	Loop Fusion	Loop Interchange
GCC compiler	Yes	No	No	Yes

TABLE 3: Loop Transformations in GCC

This work will be related for a particular application of Rodinia Benchmark such as loop fission and loop fusion will be applied on the particular loops of the application.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have presented how to distribute the work between CPU-GPU. We also showed benchmarking criteria for evaluation of automatic parallelization tools and compilers. Though the automatic parallelization tools mentioned in this paper are able to generate parallel codes, more efforts are required to make those codes optimum in terms of performance. These tools should try to skip the loops that have smaller execution time. Although there has been a lot of research over the past few decades on automating the task, manual parallelization continues to outperform automatic parallelization tools. This trend can be expected to continue as the performance gap in terms of efficiency and computing resource utilization between automatic and manual parallelization. So hybrid architectures are introduced into the market. The methodology is tested against Rodinia Benchmarks 3.0 for the particular application where we are trying to merge OPEN MP+CUDA architecture so that the load can be distributed to CPU and GPU to reduce the execution time such that performance will be increased.

REFERENCES

[1] J. S. Vetter, R. Glassbrook, J. Dongarra, K. Schwan, B. Loftis, S. Mcnally, J. Meredith,J. Rogers,P. Roth, K. Spafford, And S. Yalamanchili, "Keeneland: Bringing Heterogeneous GPU Computing To The Computational Science Community," Computing In Science & Engineering, 2011
[2] C. Vömel, S. Tomov, and J. Dongarra, "Divide and Conquer On Hybrid GPU-Accelerated Multicore Systems," SIAM Journal on Scientific Computing, Vol. 34, No. 2, Pp. C70–C82, 2012
[3] S. Tomov, J. Dongarra, And M. Baboulin, "Towards Dense Linear Algebra For Hybrid GPU Accelerated Manycore Systems," Parallel Computing, Vol. 36, No. 5, Pp. 232–240, 2010.
[4] J. Agulleiro, F. Vazquez, E. Garzon, And J. Fernandez, "Hybrid Computing: CPU+ GPU Coprocessing And Its Application To Tomographic Reconstruction," Ultramicroscopy, Vol. 115, Pp. 109– 114, 2012.
[5] P. Guo, L. Wang, And P. Chen, "A Performance Modeling And Optimization Analysis Tool For Sparse Matrix-Vector Multiplication On GPUs" IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 5, May 2014
[6] A. K. Datta And R. Patel, " CPU Scheduling For Power/Energy Management On Multicore Processors Using Cache Miss And Context Switch Data" IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 5, May 2014
[7] X. Liu, M. Li, S. Li, S. Peng, X. Liao, And X. Lu, " IMGPU: GPU-Accelerated Influence Maximization In Large-Scale Social Networks", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 1, January 2014
[8] J. Zhong And B. He, "Medusa: Simplified Graph Processing On GPUs", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 6, June 2014
[9] H. Chen, J. Sun, L. He, K. Li, And H. Tan, "Bag: Managing GPU As Buffer Cache In Operating Systems", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 6, June 2014
[10] H. Huynh, A. Hagiescu, O. Liang, W. Wong, And R. S. M. Goh, " Mapping Streaming Applications Onto GPU Systems", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 9, September 2014

[11] M. Huang, And D. Andrews, "Modular Design Of Fully Pipelined Reduction Circuits On FPGAs", IEEE Transactions On Parallel And Distributed Systems, Vol. 24, No. 9, September 2013
[12] V. Karakasis, T. Gkountouvas, K. Kourtis, G. Goumas And N. Koziris, "An Extended Compression Format For The Optimization Of Sparse Matrix-Vector Multiplication", IEEE Transactions On Parallel And Distributed Systems, Vol. 24, No. 10, October 2013
[13] H. Zhibin, Z. Mingfa, And X. Limin, "Lvtppp: Live-Time Protected Pseudo partitioning Of Multicore Shared Caches", IEEE Transactions On Parallel and Distributed Systems, Vol. 24, No. 8, August 2013
[14] S. Che, M. Boyer, D. Tarjan, "Rodinia: A Benchmark Suit for Heterogeneous Computing", IISWC, Oct. 2009 © IEEE 2009
[15] A. N. Yzelman and D. Roose, "High-Level Strategies for Parallel Shared-Memory Sparse Matrix-Vector Multiplication", IEEE Transactions On Parallel And Distributed Systems, Vol. 25, No. 1, January 2014
[16] R. Buchty, V. Heuveline, W. Karl, J. Weib, "A Survey On Hardware-Aware And Heterogeneous Computing On Multicore Processors And Accelerators", Engineering Mathematics And Computing Lab (EMCL) 2009
[17] S. E. Droz, R. R. Torrado, "High Performance Computing Oil and Gas Industry", US Department Of Energy, Technical Report, August 2011
[18] H. Kelly And Hart, "Wind Energy Fundamental Science Issues Requiring HPC", US Department Of Energy, Technical Report, August 2011
[19] Sang-Ik Lee, Troy A. Johnson, And Rudolf Eigenmann, "Cetus – An Extensible Compiler Infrastructure For Source-To-Source Transformation", Purdue University, West Lafayette IN 47906, USA
[20] J. W. Sheaffer, M. Boyer, "A Characterization Of The Rodinia Benchmark Suit With Comparision To Contemporary CMP Workloads", Dept Of Computer Science, University Of Virginia, 2011.
[21] J Shen, A L Varbanescu, "A Detailed Performance Analysis Of The Openmp Rodinia Benchmark" Delft University Of Technology Parallel And Distributed Systems Report Series, ISSN 1387-2109, PDS-2011-011.
[22] P Harish And P. J. Narayanan, "Accelerating Large Graph Algorithms On The GPU Using CUDA", Center For Visual Information Technology, International Institute Of Information Technology, 2008.