A Framework for Resilient and Energy-efficient Computing in GPU-accelerated Systems

Zheng Miao, Jon C. Calhoun (Advisor) and Rong Ge (Advisor)
Clemson University
{*zmiao, jonccal, rge*}*@clemson.edu*

***Abstract*—High-performance computing systems must simultaneously address both resilience and power. In heterogeneous systems, the trade-offs between resilience and energy-efficiency are more complex for applications using both CPUs and GPUs. A deep understanding of the interplay among energy efficiency, resilience, and performance is required for heterogeneous systems to address them simultaneously.**

In this work, we present a new framework for resilient and energy-efficient computing in GPU-accelerated systems. This framework supports partial or full redundancy and checkpointing for resilience, and provides users with flexible hardware resource selection, adjustable precision and power management to improve performance and energy-efficiency. We further perform cuda-aware MPI to reduce resilience overhead, mainly in message communication between GPUs. Using CG as an example, we show that our framework provides about 40% time and 45% energy savings, comparing to simple extension of RedMPI, a redundancy based resilience framework for homogeneous CPU systems.

1. Introduction

Resilience and power are two major concerns in high performance computing (HPC) systems. The overall failure rate and power consumption of systems increase with component counts. Future exascale systems are expected to have a similar power budget of 20 MW, and a mean time between failure (MTBF) within an hour [1]. In faulty systems, resilience techniques allow long running applications to successfully finish executions. However, resilience incurs power and time overhead, and aggravates power problem.

Heterogeneous systems composed of CPUs and GPUs have become the mainframe building blocks. However, the power and resilience challenges are more complex for GPU applications. As GPUs have a larger throughput and power consumption than CPUs, load balancing and power management between these two architectures are complex. Higher power and temperature on GPUs cause higher error rates. Resilience incurs power and time overhead and exacerbates the power challenge for heterogeneous systems with GPUs.

Previous work has investigated resilience for heterogeneous systems of GPUs, but is limited to traditional checkpointing [6]. Current studies do not have a comprehensive solution to simultaneously address both fault-tolerance and energy-efficiency for GPU applications. Shadow replication [5] shows energy-efficiency in fault tolerance mechanisms, while few studies in redundancy are developed and optimized for heterogeneous systems with GPUs.

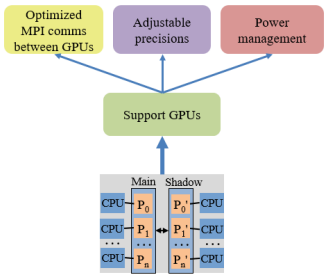

Figure 1. Design of our framework.

Current heterogeneous systems provide several opportunities and challenges for achieving energy-efficient redundancy. There are plenty of hardware resources (CPUs and GPUs) for redundancy in heterogeneous systems. Techniques such as adjustable precision and power management can improve performance and reduce power consumption. Although various hardware resources (low-end/high-end CPUs/GPUs) are available in heterogeneous systems, they are difficulty to coordinate with each other for redundancy due to the difference between their performance. Improving reliability usually requires higher resilience overhead in time and power, which hurts the energy-efficiency.

In this work, we present a framework for resilient and energy-efficient computing in heterogeneous systems. We make the following contributions:

- We present the first of its kind framework to provide redundancy in GPU-accelerated systems.
- This framework allows main and shadow processes to take over each other's role alternatively upon faults.
- This framework configures various hardware resources (low/high-power CPUs or low/high-power GPUs), flexible precision [2] and power management depending on user's requirement of applications' performance and power consumption.
- We perform optimization of resilience overhead [4] to reduce communication time between GPUs.

2. Framework Design

As Figure 1 shows, our framework is designed and implemented based on previous CPU replication [5]. We extend the framework to support main and shadow processes

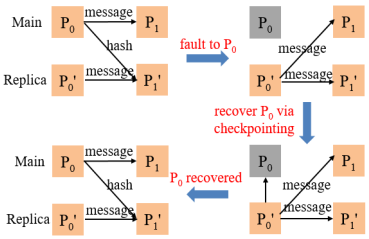

Figure 2. MPI implementation for redundancy and fault recovery. The faulty main MPI process is replaced by the corresponding replica process until the faulty process is recovered from checkpointing.

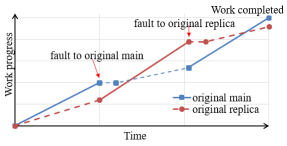

Figure 3. Recovery from multi-faults. Solid lines show current main.

to run on GPUs. The framework provides redundancy and fault recovery based on MPI message and hash [3] as Figure 2 shows. We utilize acceleration technologies like GPUDirect by the MPI library to provide high-bandwidth and low-latency communications with NVIDIA GPUs [4].

This framework has multiple features:

- provide resilience to multiple faults. Upon faults, main and replica replaces the role of each other as Figure 3.
- support double precision and mixed with single precision on CPUs and GPUs to provide different performance and power consumption.
- support DVFS on CPUs/GPUs for power management.
- Upon faults, main and replica dynamically adjust their execution rates by adjustable power management.

3. Evaluation

We implement a prototype of the framework and evaluate it on a NUMA cluster. Each node is built of dual 12-core Intel Haswell processers. There are a total of 4 Titan V GPUs. We use an iterative CG with a 27-point stencil matrix as a study case.

Figure 4-6 shows performance and energy-efficiency of our framework. By applying multiple techniques, our framework addresses performance and energy efficiency simultaneously for GPU applications in faulty environment. As problem size increases, mixed-precision on GPUs shows better performance. Mixed-precision and power management show a significant power saving. Among all the configurations, mixed-precision with higher GPU frequency shows the best performance while mixed-precision with lower GPU frequency shows the best energy saving.

4. Conclusion

We propose a novel framework for resilient and energy-efficient HPC applications for heterogeneous systems with GPUs. This work provides shadow redundancy based on

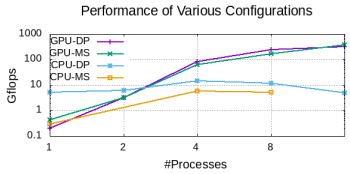

Figure 4. Performance of various resources and precisions. Performance of mixed precision exceeds that of DP (double precision) when the gain from mixed precision computing is larger than precision conversion overhead.

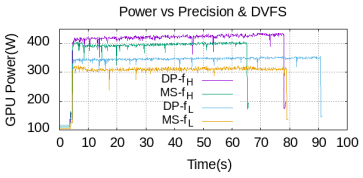

Figure 5. Power reduction with mixed-precision and lower GPU frequency. H—high frequency, L—low frequency. DP-H and mixed-L have similar execution time, but the latter has a 30% lower power consumption.

MPI implementation for GPU applications. Results show that our framework significantly reduces the resilience overhead while maintaining energy-efficiency.

For future work, we will optimize our resilience mechanism in recovering faulty processes to improve reliability. More advanced accelerators such as tensor cores or FPGAs will be incorporated into our framework. We will also incorporate mixed-precision with FP16 in those advanced accelerators for higher energy-efficiency.

References

[1] S. Ashby, P. Beckman, J. Chen, P. Colella, B. Collins, D. Crawford, J. Dongarra, D. Kothe, R. Lusk, P. Messina, et al. The opportunities and challenges of exascale computing. *ASCAC*, pages 1–77, 2010.

[2] A. Buttari, J. Dongarra, J. Kurzak, P. Luszczek, and S. Tomov. Using mixed precision for sparse matrix computations to enhance the performance while achieving 64-bit accuracy. *TOMS*, 34(4):17, 2008.

[3] D. Fiala, F. Mueller, C. Engelmann, R. Riesen, K. Ferreira, and R. Brightwell. Detection and correction of silent data corruption for large-scale high-performance computing. In *SC12*.

[4] J. Kraus. An introduction to cuda-aware mpi. *Weblog entry]. PARALLEL FORALL*, 2013.

[5] B. Mills, T. Znati, and R. Melhem. Shadow computing: An energy-aware fault tolerant computing model. In *2014 International Conference on Computing, Networking and Communications (ICNC)*, pages 73–77. IEEE, 2014.

[6] L. Shi, H. Chen, and T. Li. Hybrid cpu/gpu checkpoint for gpu-based heterogeneous systems. In *International Conference on Parallel Computing in Fluid Dynamics*, pages 470–481. Springer, 2013.

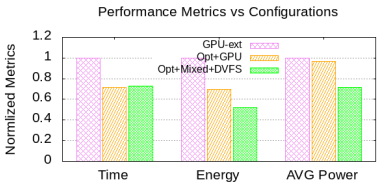

Figure 6. Normalized performance metrics for various configurations. Base is a simple extension of CPU replication to support GPUs in center layer of Figure 1. The following three bars shows performance with techniques in top layer of Figure 1.