Leaping Shadows: Fault Tolerance with Forward Progress

Xiaolong Cui, Taieb Znati, Rami Melhem
Computer Science Department
University of Pittsburgh
Pittsburgh, USA
Email: {mclarencui, znati, melhem}@cs.pitt.edu

***Abstract*—**

***Keywords*-Leaping; Lazy Shadowing; extreme-scale computing; forward progress; reliability;**

I. Introduction

Computing power of supercomputers has become a vital factor in determing a country's competiveness in research, technology, and even national defence. As the demand for massive computing power continues to increase, the HPC community is striving to develop larger computing systems in pursuit of the power of scale. Nowadays, a race among various countries is underway to build the world's first exascale supercomputer to accelerate scientific discoveries, big data analytics, etc. It is expected that the first exascale supercomputer will enter service by 2020. However, making the transition to extreme-scale poses numerous unavoidable scientific and technological challenges.

As today's HPC and Cloud Computing systems grow to meet tomorrow's compute power demand, two of the biggest challenges will be system resilience and power consumption, both being direct consequences of the dramatic increase in the number of system components [1], [2]. Regardless of the reliability of individual component, the system level failure rate will continue to increase as the number of components increases, possibly by several orders of magnitude. It is projected that the Mean Time Between Failures (MTBF) of future extreme-scale systems will be at the order of hours or even minutes, meaning that many failures will occur every day [3]. Without an efficient fault tolerance mechanism, faults will be so frequent that the applications running on the systems will be continuously interrupted, requiring the execution to be restarted from a previous point for every failure.

Also thanks to the continuous growth in system components, there has been a steady rise in power consumption in large-scale distributed systems. In 2005, the peak power consumption of a single supercomputer reached 3.2 Megawatts. This number was doubled only after 5 years, and reached 17.8 Megawatts with a machine of 3,120,000 cores in 2013. Recognizing this rapid upward trend, the U.S. Department of Energy has set 20 megawatts as the power limit for future exascale systems, turning power from an optimzation goal to a leading system design constraint.

Today, two classic approaches to fault tolerance are dominant. The first approach is rollback recovery, which rolls back and restarts the execution every time there is a failure. This approach is often equipped with checkpointing to periodically save the execution state to a stable storage so that execution can be restarted from a recent checkpoint in the case of a failure [4], [5], [6]. On the other hand, the second approach, referred to as process replication, exploits hardware redundancy and executes multiple instances of the same task in parallel to overcome failure [7], [8], [9].

However, neither of the above two approaches applies well to future extreme-scale systems. Checkpointing/restart lacks of forward progress, meaning that its efficiency drops as failure rate increases. Given the anticipated increase in system level failure rates and the time to checkpoint large-scale compute-intensive and data-intensive applications, the time required to periodically checkpoint an application and restart its execution will approach the system's MTBF [10]. On the other hand, process replication has a low system efficiency (no more than 50%) by default because of its need for dedicated resources for replica processes. Therefore, should an exascale system be built in the next few years with any of the two fault tolerance approaches, a large portion of its capacity would be wasted due to fault tolerance. Furthermore, neither of them addresses the power cap issue.

To address the above shortcomings, we proposed Lazy Shadowing as an adaptive and power-aware approach to achieve high-levels of resilience in extreme-scale, failure-prone computing environments. Lazy Shadowing is a novel computational model that goes beyond adapting or optimizing well known and proven techniques, and explores radically different methodologies to fault tolerance [11]. The basic tenet of Lazy Shadowing is to associate with each main process a suite of shadows whose size depends on the "criticality" of the application and its performance requirements. Each shadow process is an exact replica of the original main process, and a consistency protocol is used to assure that the main and the shadow are consistent. When possible, the shadow executes at a reduced rate to save power. Lazy Shadowing achieves QoS along with power awareness by dynamically responding to failures. Experimental results demonstrate that Lazy Shadowing achieves higher performance and significant energy savings compared

to existing approaches in most cases.

Recently, however, several limitations of Lazy Shadowing have been identified. Firstly, Lazy Shadowing can only tolerate one failure per shadowed set. As a result, failures, as they occur, reduce the vulnerability of the system. Secondly, shadows are substitutes for mains, increasing the implementation complexity to deal with failures. Thirdly, the assumsion of crash failures limits the efficiency of Lazy Shadowing.

In this paper, we present our latest work on addressing the above limitations. Firstly, we combine dynamic process creation with shadow leaping to efficiently keep the system from becoming vulnerable. Then, we deviate from the concept of shadow as a replica, where a shadow is promoted to a new main when the original main fails, and use shadow as an "assistant" to a main whereby a shadow helps a main to overcome failures until the main completes execution. Last but not least, for different types of failures, we study different schemes to optimize Lazy Shadowing. The main contributions of this paper are as follows:

- An enhanced scheme of Lazy Shaodwing that incorporates rejuvenation techniques for consistent reliability and improved performance.
- A full-feature implementation for Message Passing Interface
- A through evaluation of the overhead and performance of the implementation with various benchmarks and real applications.

The rest of the paper is organized as follows. We begin with a survey on related work in Section II. Section III introduces system design and fault model, followed by discussion on implementation details in Section IV. Section V presents empirical evaluation results. Section VI concludes this work and points out future directions.

II. RELATED WORK

III. SYSTEM DESIGN

The computational model of Lazy Shadowing has been continuously optimized to improve its scalability and efficiency, to eliminate vulnerability, and to reduce implementation complexity and overhead. This section presents the system design of the comprehensive Lazy Shadowing model.

A. Fault model

As demonstrated in [11], Lazy Shadowing is able to tolerate both hardware and software failures under the fail-stop fault model. In this work, we continue with the assumption of fail-stop model. In order to further improve resilience and efficiency, however, we differentiate between temporary failures and permanent failures. Temporary failures include memory bit flips, kernel panic, etc., and can be recoverd by rebooting the machine, while permanent failures, such as failure in power supply and network switch, needs the device to be replaced in order to recover. Later in this paper we will show that Lazy Shadowing maximizes the resource utilization for resilience by adopting different schemes for different types of failures.

B. Shadowing

Shadowing is the essential concept in Lazy Shadowing. With shadowing, each original process (referred to as main) is associated with a replica process (referred to as shadow) that executes at a potentailly lower rate to save power. Then fault tolerance comes from the property that if one process fails, its associated process can continue to complete the task. For example, if a main process fails, its shadow is promoted to a new main and continue to carry out the assigned task. In this way, shadow processes are substitutes for mains in the case of failure. In this work, we deviate from the original concept of shadow as a replica and use a shadow as "assitant" to a main to complete computation in the presence of failures. Specifically, if a main fails, a new main process will be rejuvenated from its associated shadow by our leaping technique (discussed below), while the shadow remains a shadow.

C. Leaping

Leaping is initially proposed as a technique to boost performance in [11]. As the shadows execute slower than mains, failure recovery will introduce delay to the execution. Leaping opportunistically takes advantage of the recovery time and copies state from healthy mains to their associated shadows. As a result, shadows achieve forward progress with minimal overhead, and recovery time for future potential failures is minimized.

Recently, we have identified an empirical problem to which leaping is a solution. When Lazy Shadowing is used to execute an MPI application, application messages are generated at the rate of the mains, but consumed by the shadows at a lower rate because the shadows are slower. As a result, messages accumulate on the shadow side and could possibly result in a buffer overflow. Leaping is naturally a solution to this problem as it can move the shadows forward and synchronize their execution states with those of the mains. After the synchronization, accumulated messages at the shadows become obsolete and thus can be safely discarded. To differentiate the two cases where leaping is used, leaping during failure recovery is referred to as failure induced leaping while leaping to avoid buffer overflow is referred to as forced leaping.

D. Rejuvenation

It has been discussed in [11] that with shadow collocation each shadowed set can only tolerate one failure. After the first failure, all main processes in the shadowed set would lose their shadows and become vulnerable. Although quantitative study show that a second failure in a shadowed set is unlikely to occur even with over one million processes, in

practice the system will become more and more vulnerable as failures occur. In many cases, it would be too costly to take such risk, especially for long-running, large-scale, and mission-critical applications. Therefore, it is preferable to maintain the same level of resilience across failures.

Not surprisingly, vulnerability could be avoided by creating a new process for every failed process (either main or shadow). In this way, every main is always guaranteed to have an associated shadow and shadow does not need to substitute a main. The problem, however, is that the newly created process will start from the beginning and may lag far behind the other processes in the system. Later on when the new processs needs to participate in a synchronization point or when it need to perform a failure recovery, significant delay will incur as a result of its lag. Fortunately, leaping can be used to deal with this issue. After a new process is created, we can use leaping to synchronize the new process' state with the existing one. Consider a pair of main process, M, and shadow process, S. If M fails, a new main process will be created to replace M, and then a leaping from S will advance the new process to the state of S. On the other hand, if S fails, a new shadow process will be created and immediately advanced to the state of M by leaping.

Depending on the type of failure, the new process will be placed at different locations. If it is a temporary failure, the node where failure occurs will be rebooted and then used to host the new process, whether it is a main or shadow. Although there is a delay from the rebooting, it is usually acceptable and can be accounted part of the recovery. For permanent failures, the node cannot be used and we have to migrate and collocate some processes. If the new process is a main, its existing shadow will be migrated to another node where shadow process(es) reside, and make room for the new main. Otherwise, if the new process is a shadow, it will be directely created on a shadow node.

IV. IMPLEMENTATION

V. EVALUATION

VI. CONCLUSION AND FUTURE WORK

ACKNOWLEDGMENT

This research is based in part upon work supported by the Department of Energy under contract DE-SC0014376.

REFERENCES

[1] S. Ashby, B. Pete, C. Jackie, and C. Phil, "The opportunities and challenges of exascale computing," 2010.

[2] M. Snir, R. W. Wisniewski, J. A. Abraham, S. V. Adve, S. Bagchi, P. Balaji, J. Belak, P. Bose, F. Cappello, B. Carlson *et al.*, "Addressing failures in exascale computing," *International Journal of High Performance Computing Applications*, p. 1094342014522573, 2014.

[3] K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, K. Yelick, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dally, M. Denneau, P. Franzon, W. Harrod, J. Hiller, S. Keckler, D. Klein, P. Kogge, R. S. Williams, and K. Yelick, "Exascale computing study: Technology challenges in achieving exascale systems peter kogge, editor & study lead," 2008.

[4] E. Elnozahy and et. al., "A survey of rollback-recovery protocols in message-passing systems," *ACM Computing Surveys*, vol. 34, no. 3, pp. 375–408, 2002.

[5] S. Kalaiselvi and V. Rajaraman, "A survey of checkpointing algorithms for parallel and distributed computers," *Sadhana*, vol. 25, no. 5, pp. 489–510, 2000. [Online]. Available: http://dx.doi.org/10.1007/BF02703630

[6] K. M. Chandy and L. Lamport, "Distributed snapshots: Determining global states of distributed systems," *ACM Trans. Comput. Syst.*, vol. 3, no. 1, pp. 63–75, Feb. 1985. [Online]. Available: http://doi.acm.org/10.1145/214451.214456

[7] J. F. Bartlett, "A nonstop kernel," in *Proceedings of the Eighth ACM Symposium on Operating Systems Principles*, ser. SOSP '81. New York, NY, USA: ACM, 1981, pp. 22–29. [Online]. Available: http://doi.acm.org/10.1145/800216.806587

[8] W.-T. Tsai, P. Zhong, J. Elston, X. Bai, and Y. Chen, "Service replication strategies with mapreduce in clouds," in *Autonomous Decentralized Systems,* 10^{th} *Int. Symp. on*, 2011, pp. 381–388.

[9] K. Ferreira, J. Stearley, J. H. Laros, III, R. Oldfield, K. Pedretti, R. Brightwell, R. Riesen, P. G. Bridges, and D. Arnold, "Evaluating the viability of process replication reliability for exascale systems," in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '11, 2011, pp. 44:1–44:12.

[10] F. Cappello, A. Geist, B. Gropp, L. Kale, B. Kramer, and M. Snir, "Toward exascale resilience," *Int. J. High Perform. Comput. Appl.*, vol. 23, no. 4, pp. 374–388, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1177/1094342009347767

[11] X. Cui, T. Znati, and R. Melhem, "Adaptive and power-aware resilience for extreme-scale computing," in *16th IEEE International Conference on Scalable Computing and Communications*, July 18-21 2016.