Havens: Explicit Reliable Memory Regions for HPC Applications

Saurabh Hukerikar, Christian Engelmann
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA
Email: {hukerikarsr, engelmannc}@ornl.gov

***Abstract*—Supporting error resilience in future exascale-class supercomputing systems is a critical challenge. Due to transistor scaling trends and increasing memory density, scientific simulations are expected to experience more interruptions caused by transient errors in the system memory. Existing hardware-based detection and recovery techniques will be inadequate to manage the presence of high memory fault rates.**

In this paper we propose a partial memory protection scheme based on region-based memory management. We define the concept of regions called *havens* that provide fault protection for program objects. We provide reliability for the regions through a software-based parity protection mechanism. Our approach enables critical program objects to be placed in these havens. The fault coverage provided by our approach is application agnostic, unlike algorithm-based fault tolerance techniques.

I. Introduction

The management of faults and errors is a critical concern for the next generation of high-performance computing systems. With transistor scaling, the computing devices fabricated using miniaturized device feature sizes face serious reliability threats. The presence of frequent faults in the hardware, and the resulting errors, will have catastrophic effects on the scientific applications. The applications may not be able to run to completion, or may complete with wrong results due to undetected errors. Therefore, effective fault resilience techniques will be critical for the productive and efficient use of future exascale-class HPC systems [1].

Soft errors in memory are expected to increase dramatically due to increased memory densities in modern HPC systems. Traditionally, reliable memory systems have been equipped with hardware-based ECC, which is capable of correcting single bit errors and detecting double bit errors (SECDED) [2]. However scaling trends, which shrink transistor sizes and reduce supply voltage, result in reduction of charge stored per device, and a correspondingly higher exposure to soft error events. Furthermore, exascale systems are expected to embrace more complex and denser memory hierarchies, and utilize more diverse memory technologies in pursuit of higher memory capacity, and better energy efficiency and performance [3]. Therefore, existing hardware-based detection and correction techniques will be insufficient to cope with all memory errors. Silent data corruptions (SDC) are particularly problematic because they cause incorrect data to be delivered to applications without an error notification being raised.

From the perspective of an HPC application, there is an expectation of reliable program execution regardless of how unreliable a system is. However, for various scientific applications, not every computation needs to be performed reliably as long as specific critical computations and their input and result data are guaranteed to be highly reliable. Different variables in a program may exhibit different vulnerabilities to transient faults. Many classes of scientific algorithms are able to leverage such an expectation of selective robustness to continue to make progress towards an acceptable outcome in the presence of some system unreliability [4]. In this paper, we explore the use of selective reliability through a region-based approach to memory allocation. This approach enables the creation of software-enabled robust memory regions called *havens*. We guarantee highly-reliable behavior for these regions through comprehensive error protection based on a lightweight software-based parity protection scheme. Havens permit grouping program objects with similar robustness requirements. Based on this model for memory management, the HPC applications may develop resiliency strategies that trade-off application performance with reliability under program control, by declaring certain program data objects and computations to be highly reliable. Using accelerated fault-injection experiments, we measure the fault coverage of our approach, which confirms that havens provide high levels of fault protection for critical program objects with limited overhead to the overall application performance.

The rest of the paper is organized as follows: Section II motivates the concept and describes its important features. Section III describes abstract haven interfaces, while Section IV describes our prototype implementation of the interfaces and the parity-based protection algorithm. Section V presents the experimental evaluation and results.

This work was sponsored by the U.S. Department of Energy's Office of Advanced Scientific Computing Research. This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy.

Fig. 1. Havens-based Memory Management

II. Resilience-driven Memory Management

In the region-based memory management scheme for typed, high-level programming languages, the runtime memory is partitioned into *regions* and the assignment of program objects to regions is statically decided by the compiler. The memory allocation and deallocation is statically specified in the program at compile time and the program is annotated to include operations for managing regions [5]. The approach was proposed as an alternative to garbage collection in the context of declarative programming languages. The scheme was improved by relaxing the requirement for region lifetimes to be lexical [6]. However, the efficacy of using static memory management decisions in place of runtime garbage collection has been widely debated.

We propose an adaptation of the region-based memory management concept in which our primary aim is to provide error resilient memory regions for HPC applications. Based on the insight that HPC applications dominantly use imperative programming paradigms, rather than declarative programming, we do not limit the concept of havens to any specific language style. Our proposed approach retains the core concept of well-defined regions for program objects. However, rather than assigning objects to regions solely based on their lexical scopes, we relax several constraints of the original concept and define features that make havens more suitable to HPC applications and programming paradigms and provide reliability. The salient features of our resilience-driven, region-based memory management scheme are:

- Havens provide robust containers in which program objects may be allocated. The memory region is protected by a predefined error detection and/or correction scheme that is agnostic to the algorithm features or structure of the data objects placed in havens.
- The havens enable separation between the memory allocation and the implementation of the robustness scheme. In theory, a region can be made robust using various protection mechanisms, e.g., replication, parity, hashing, etc. As illustrated in Figure 1 various regions in the program memory may be offered coverage using separate protection schemes. Since each of these methods carry a different level of overhead, for an application this enables the creation of logical grouping of objects that require similar robustness characteristics.
- Havens enable applications to exert fine-grained control on the resilience properties of individual program objects. Since different havens may have varying guarantees of reliability, based on the strength of the protection mechanism and its performance overhead, object placement in havens may be driven by the trade-off between criticality of the object to program correctness and the associated overhead.
- The objects in a haven are all freed at once by deleting the entire pool of memory. Therefore, havens enable the association of lifetime to the robust memory regions.

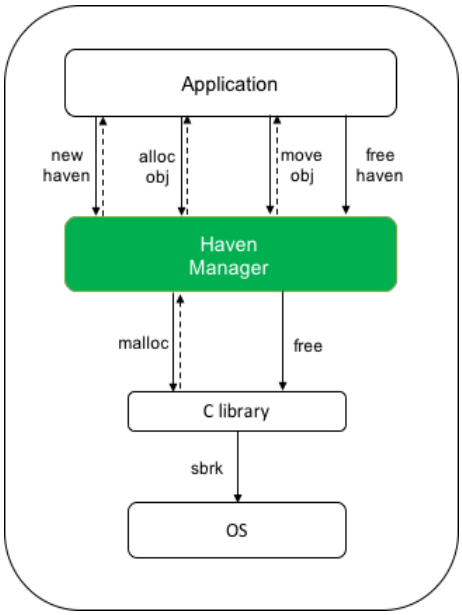

Fig. 2. Havens Interfaces

The key advantage of havens-based memory management is that it makes it straightforward for HPC applications to reason about the reliability requirements of programs and specific program fragments.

III. Using Havens in HPC Applications

In this section we introduce the abstract interface between the runtime memory manager and the application program. The overall goal of the havens approach is to enable applications to allocate program objects into regions for which there is a well-defined robustness scheme. Each program object that is resident to a region is provided with error protection. In imperative languages such as C/C++, the memory allocation interface consists of two operations: an *alloc* and *free* which create new heap objects. In our scheme, the application program interacts with a haven manager, which implements the haven abstraction and its error protection mechanism and handles the subdivision of the heap memory.

A. Creation

The havens are part of the heap where memory is allocated. When the region is first created, no memory is allocated. The creation is requested by an application using an abstract interface to the haven manager. When the application requests the creation of a new region, the haven manager returns a handle to the memory region, which is to be used in subsequent operations. The abstract interface needs to provides a `__haven_create__` primitive for haven creation. Also, the robustness technique may be explicitly specified for the memory region. Upon creation, the haven is an abstract entity for which an application expresses an expectation of reliability.

B. Allocation

An application requests allocation of memory within a region using the handle to the haven. This entails allocation of a block consisting of a specified amount of memory. The abstract interface must provide an interface `__haven_alloc__`, to which the application specifies a handle to the haven of which the new memory block will be a part of. The haven manager maintains state for the robustness feature of the memory region and the allocation requests initialization of such state.

C. Deallocation

When the haven is destroyed, all memory blocks allocated in the region are deallocated and the memory is available for reuse. Since this is the only way to deallocate memory in the region model, no further memory allocation requests are valid. Therefore, all the program objects located in a particular region are all simultaneously deallocated. The robustness characteristic of the haven and the state maintained by the haven manager is also destroyed. The abstract interface must provide a `__haven_destroy__` primitive, which uses the haven handle.

D. Relaxation

Among the key features of the havens-based memory management is the association of reliability lifetime with the memory contained in the haven. Many HPC applications can naturally divide their work into nearly independent pieces or phases for which the reliability requirement may be more or less stringent. The error protection scheme may be turned on and off based on the needs of the application program. The abstract interface must provide primitives `__haven_relax__` and `__haven_robust__` in order to relax and impose the protection scheme on the memory region.

IV. Implementation

The description of the haven interfaces in the previous section is intentionally abstract and language-independent, and a variety of implementations for these interfaces are possible. To simplify porting existing HPC applications to use havens, our initial prototype implementation provides a minimal set of haven library interface calls, which will require limited changes to application source codes.

A. Basic Haven Operations

The heap is divided into fixed-size pages, and each new haven creation is aligned on a page-size boundary. The haven manager maintains a linked list of these pages. The library provides concrete functions for each of the primitives that enable basic haven operations. The library provides a `haven-type` that is used to create handle objects, and which contains the bounds of page addresses that make up the memory region. An identifier is bound to a haven handle and the handle to the haven is passed as argument to the allocation and destruction functions. The functions `__hmalloc()` and `__hnew()` implement the abstraction `__haven_create()` for the allocation of objects into the associated region. The implementation of the haven system imposes no changes on the representation of pointers and permits the access to individual objects in the havens using pointer operations. However, we only support per-region allocation and deallocation and therefore per-object deallocation is an illegal operation. The `__hdestroy()` operation concatenates all the haven's page list to a global list of free pages. The present implementation does not support techniques that prevent dangling-pointer dereferences and require the programmer to ensure safety when haven deallocation function is called.

B. Parity-based Protection

The goal of the error protection mechanism is to provide comprehensive fault coverage for the memory of an entire haven. In order to implement a generic scheme that is agnostic to the actual program data structure, yet carries low performance overhead, we employ a parity-based detection/correction mechanism. The region is the unit of memory over which the library applies the parity-based detection/correction.

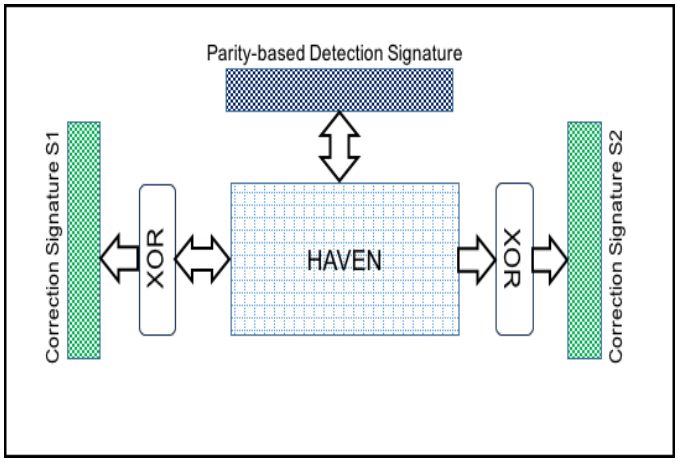

Fig. 3. Parity-based Protection for Haven Memory

The Figure 3 illustrates the implementation of the parity-based detection/correction scheme applied to the haven. The implementation is based on the bit-level representation of the memory in the haven and is independent of the higher-level data type. The haven manager maintains a pair of correction

signatures for each memory region, which are of word length, and an additional word length detection signature per 64 words in the memory block in the region. The detection signature contains one parity bit per word in the memory region. As memory is allocated for the region and initialized, the correction signature S1 retains the XOR of all words that are written to the memory region. We apply an XOR operation on every word that is updated in the memory region and the correction signature S2. Silent data corruptions or multi-bit errors are detected based on the state of the detection signature containing a parity bit for each word contained in the memory region.

When the parity signature for a memory location in the haven detects a parity violation, the location of the corrupted memory word may be identified. The value at the corrupted memory location may be recovered using the XOR signature words. The XOR of the two signatures S1 and S2 equals the XOR of all the uncorrupted locations in the haven. The corrupted value in the memory region is recovered by performing an XOR operation on the remaining words in the haven with the XOR of the two signatures S1 and S2. The recovered value overwrites the corrupted value, and the detection signature is recomputed. Using this correction scheme, multibit corruptions may be recovered from (unlike hardware ECC, which only offers single-bit error correction and double-bit error detection). Each of these of detection/recovery operations are transparent to the application. This parity-based protection is an adaptation of an erasure code and it maintains very limited state for the detection and correction capabilities and therefore carries very little space overhead in comparison to other software-based schemes such as software-based ECC and checksums schemes. The detection is a constant time operation while the recovery is a O(n) operation based on the size of the haven.

V. Experimental Evaluation

A. Methodology

The fault model for these experiments is transient errors, particularly silent data corruptions. Therefore, we have designed an accelerated fault injection system in which we inject multiple random faults into the address space of a program under execution. We perturb multiple bits of the selected injection site and no error notification is raised. The fault location and distribution of fault intervals is randomized, and applications experience at least 5 faults per application run.

For the evaluation of the haven-based memory management system, we modify a conjugate gradient code to include the havens library API calls. We use a pre-conditioned iterative CG algorithm, and we validate the outcome of the solver with a solution produced using a direct solver. In the CG algorithm, which solves a system of linear equations A.x = b, the application allocates the matrix A, the vector b and the solution vector x. Additionally, the conjugate vectors p and the residual vector r are referenced during each iteration of the algorithm. We perform two sets of experiments, in both of which the various objects in the CG code are allocated using havens. In the first experiment, we evaluate the CG application under accelerated fault injections. For each experiment run, we allocate only one structure using the haven, while the remaining structures are allocated using the standard memory allocation interfaces.

From an error recovery perspective, the application state may be classified into three types of state: the static state, which represents variables initialized at the start and remain unmodified during the computation; the dynamic state, which represents the solution data structures that are continuously updated; and, the computational environment, which represents the program code, environment variable, pointer variables, etc. In the second set of experiments, this classification forms the basis of selecting which data structures in the CG code are allocated into havens. We evaluate the following combinations of CG objects' placements: (i) allocation of only the static state, i.e., the matrix A and vector B, the preconditioner M into havens, while the dynamic state, i.e., all the solution vectors, are allocated using standard memory allocation functions; (ii) allocation of only matrix A and vector B into havens; (iii) only the dynamic state is provided fault coverage using havens.

B. Results

Providing fault coverage to all the CG's variables results in a success rate in the excess of 90% for the accelerated fault injection experiments, as shown in Figure 4(a). The operand matrix A occupies a dominant part of the solver's memory, occupying over 50% of the active address space, whereas the solution vector x, the conjugate vectors p and the residual vector r and the preconditioner matrix M account for the remaining space. The preconditioner matrix M demonstrates lower sensitivity to the silent errors, as do the vectors x, p, r. This is because silent errors in the operand matrix A or vector b fundamentally changes the linear system being solved. Even if the CG solver converges to a solution, it may be significantly different from a correct solution. The allocation of these structures into havens provides a substantially higher resilient behavior in terms of completion rates of the CG algorithm. The algorithm is naturally resilient to silent corruptions in the solution vectors, and the results suggest that there is little incentive to allocate these vectors into havens. The preconditioning matrix is also tolerant to errors, but may sometimes require a substantially higher number of iterations to converge to a solution. The performance overhead of using havens to the convergence time of the CG code (Figure 5(b)) is higher for higher coverage, but correlating the coverage results demonstrates that higher coverage for low sensitivity program state does not translate to higher application resilience.

The results of the evaluation of the strategic allocation of program objects using havens (Figure 4(b)), and their correlation with the performance overhead analysis (Figure 5(b)), shows that the allocation of only the matrix A and vector B yields the best return in terms of completion success rate of the application. The remaining combinations of allocation using havens provides diminishing returns. The results indicate that the CG solver suits the model of selective reliability, in

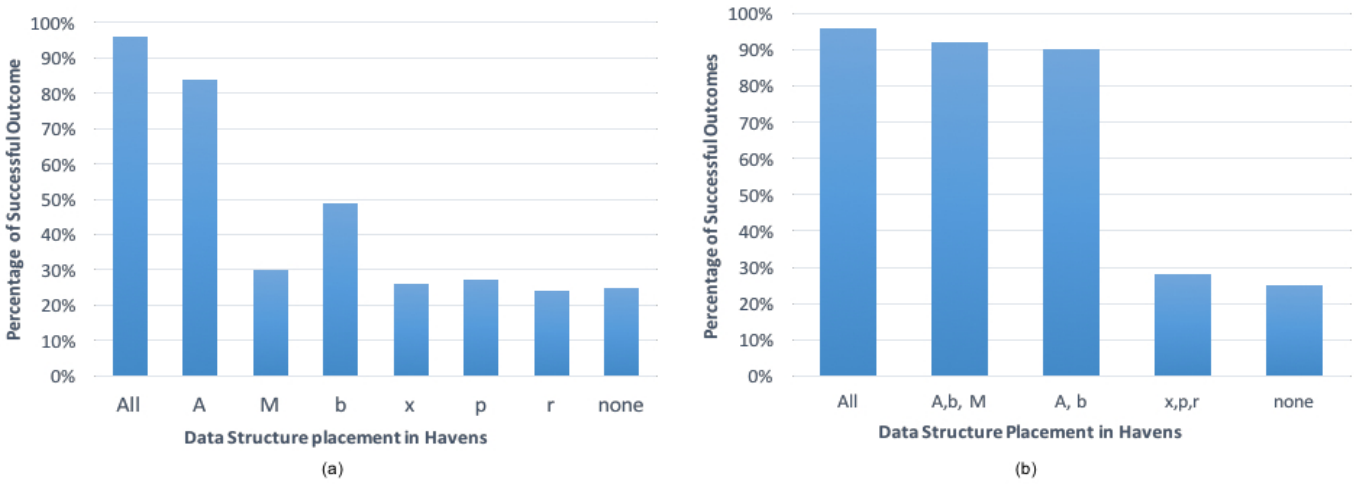

Fig. 4. Impact of havens' fault coverage on application outcomes

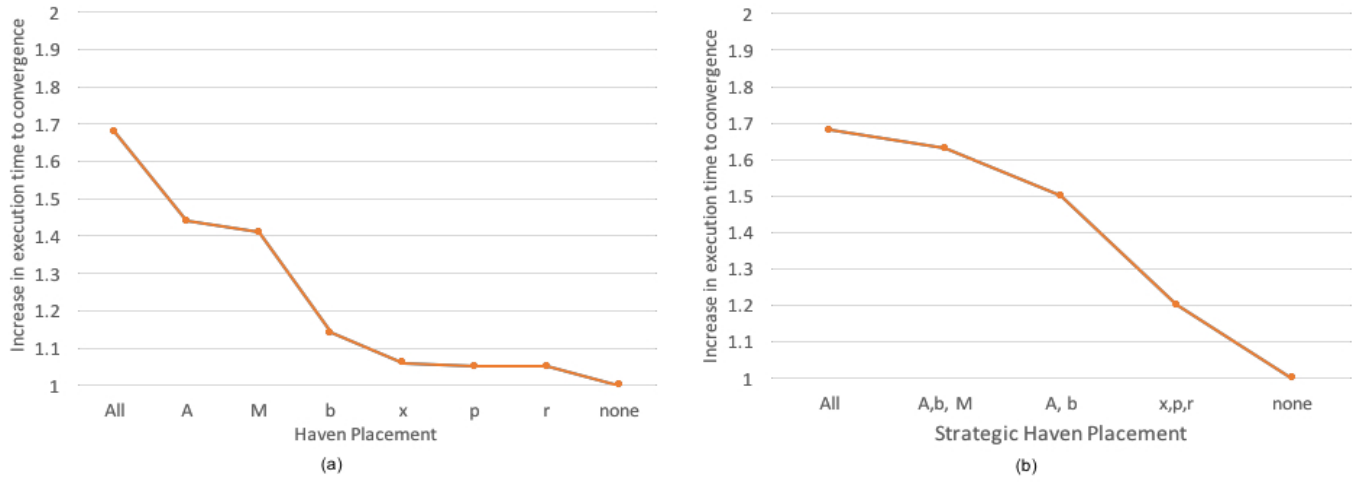

Fig. 5. Performance overheads of havens

which the various program objects show different sensitivities to errors and impact on the outcome of the application. The results indicate that placing only such critical objects in the parity-protected havens yields higher probability of successful outcome and we are able to avoid additional overhead to performance incurred in protecting lesser important program objects.

VI. Related Work

Reliable memory systems have been implemented using hardware-based encoding techniques such as ECC [2] [7]. Memory scrubbing techniques were proposed in order to increase memory reliability in noisy environments [8], which entail periodically reading all memory lines, correcting any bits through ECC, and writing corrected data back to the same location. The use of additional encoding bits for stronger error detection and correction codes adds overheads to the area, energy consumption, and increases the memory access latencies. Other approaches have focused on designing fault tolerant memory cells through cell-hardening techniques [9] [10]. The partial protection of cache memory has been accomplished by partitioning sensitive and non-sensitive data [11], and through dynamic assignment of protection priority to error-prone cache lines [12]. Partial protection was also explored through a reliability-aware data placement scheme that provided stronger fault protection for only the program code and data sections [13].

Software-based selective protection may be accomplished through compiler-directed techniques, in which all statements relevant to the computation of specific variables are extracted using static program analysis, and placed in protection domains [14]. Programming constructs called containment domains [15] offer transactional semantics for programmer-defined computational scope, where as Rolex [16] offers langauge-based extensions that support various resilience semantics on application data and computations. Global View Resilience (GVR) supports reliability of application data by

providing an interface for applications to maintain version-based snapshots of the data [17]. In support of fault tolerance of dynamic memory allocation, the `malloc_failable` may be used by the application to allocate memory on the heap; callback functions are used to handle error recovery of the memory block [18]. Much research has also been done on algorithm-based fault tolerance (ABFT) methods [19] [20]. These approaches apply encoding on the application data structures to detect and recover errors in them. However, such approaches require significant modifications to the application source code to incorporate the fault detection and correction capabilities into the algorithm. The notion of selective reliability was demonstrated as a viable solution for iterative methods for linear solvers, in which the GMRES algorithm is partitioned into reliable and unreliable phases [21].

VII. Conclusion

Future HPC systems may not be able to offer complete hardware-based protection mechanisms for memory errors due to constraints of cost and energy. Due to their frequency, which is only expected to increase in future systems, it is imperative to protect critical program objects from errors that are uncorrectable by hardware schemes, and from silent data corruptions. In this paper we presented a software-driven fault tolerance technique for program memory, which extends the concept of region-based memory management. HPC applications may identify the most critical memory objects and place them in robust regions, called havens. The havens are protected using a software-implemented, lightweight parity-based error detection and correction scheme, which provides transparent protection against memory errors without relying on any hardware-based detection/correction schemes. Our solution is agnostic to the algorithm design, and is applicable to a variety of HPC applications that require selective reliability. Our results show that providing such explicitly robust memory havens proves to be an effective way of providing low-cost fault tolerance for applications running on unreliable system fabrics.

References

[1] E.N.Elnozahy, R. Bianchini, T. El-Ghazawi, A. Fox, F. Godfrey, A. Hoisie, K. McKinley, R. Melhem, J. Plank, P. Ranganathan, and J. Simons, "System Resilience at Extreme Scale," DARPA, Tech. Rep., 2008.

[2] D. C. Bossen and M. Y. Hsiao, "A system solution to the memory soft error problem," *IBM Journal of Research and Development*, vol. 24, no. 3, pp. 390–397, May 1980.

[3] P. Kogge, K. Bergman, S. Borkar, D. Campbell, W. Carlson, W. Dallya, M. Denneau, P. Franzon, W. Harrod, K. Hill, J. Hiller, S. Karp, S. Keckler, D. Klein, R. Lucas, M. Richards, A. Scarpelli, S. Scott, A. Snavely, T. Sterling, R. S. Williams, and K. Yelick, "Exascale Computing Study: Technology Challenges in Achieving Exascale systems," DARPA, Tech. Rep., September 2008.

[4] N. DeBardeleben, J. Laros, J. Daly, S. Scott, C. Engelmann, and B. Harrod, "High-End Computing Resilience: Analysis of issues facing the hec community and path-forward for research and development," *Whitepaper*, December 2009.

[5] M. Tofte and J.-P. Talpin, "Implementation of the typed call-by-value λ-calculus using a stack of regions," in *Proceedings of the 21st ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, ser. POPL '94. New York, NY, USA: ACM, 1994, pp. 188–201. [Online]. Available: http://doi.acm.org/10.1145/174675.177855

[6] A. Aiken, M. Fähndrich, and R. Levien, "Better static memory management: Improving region-based analysis of higher-order languages," in *Proceedings of the ACM SIGPLAN 1995 Conference on Programming Language Design and Implementation*, ser. PLDI '95, 1995, pp. 174–185.

[7] C. L. Chen and M. Y. Hsiao, "Error-correcting codes for semiconductor memory applications: A state-of-the-art review," *IBM Journal of Research and Development*, vol. 28, no. 2, pp. 124–134, March 1984.

[8] A. M. Saleh, J. J. Serrano, and J. H. Patel, "Reliability of scrubbing recovery-techniques for memory systems," *IEEE Transactions on Reliability*, vol. 39, no. 1, pp. 114–122, April 1990.

[9] F. Vargas and M. Nicolaidis, "Seu-tolerant sram design based on current monitoring," in *Twenty-Fourth International Symposium on Fault-Tolerant Computing (FTCS), Digest of Papers.*, June 1994, pp. 106–115.

[10] T. Calin, M. Nicolaidis, and R. Velazco, "Upset hardened memory design for submicron cmos technology," *IEEE Transactions on Nuclear Science*, vol. 43, no. 6, pp. 2874–2878, December 1996.

[11] K. Lee, A. Shrivastava, I. Issenin, N. Dutt, and N. Venkatasubramanian, "Mitigating soft error failures for multimedia applications by selective data protection," in *Proceedings of the 2006 International Conference on Compilers, Architecture and Synthesis for Embedded Systems*, ser. CASES '06, 2006, pp. 411–420.

[12] S. Kim and A. K. Somani, "Area efficient architectures for information integrity in cache memories," in *Proceedings of the 26th International Symposium on Computer Architecture*, 1999, pp. 246–255.

[13] M. Mehrara and T. Austin, "Exploiting selective placement for low-cost memory protection," *ACM Transactions on Architecture and Code Optimization*, vol. 5, no. 3, pp. 14:1–14:24, Dec. 2008.

[14] G. Chen, M. Kandemir, M. J. Irwin, and G. Memik, "Compiler-directed selective data protection against soft errors," in *Proceedings of the ASP-DAC 2005. Asia and South Pacific Design Automation Conference, 2005.*, vol. 2, January 2005, pp. 713–716 Vol. 2.

[15] J. Chung, I. Lee, M. Sullivan, J. H. Ryoo, D. W. Kim, D. H. Yoon, L. Kaplan, and M. Erez, "Containment domains: a scalable, efficient, and flexible resilience scheme for exascale systems," in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, 2012, pp. 58:1–58:11.

[16] S. Hukerikar and R. F. Lucas, "Rolex: Resilience-oriented language extensions for extreme-scale systems," *The Journal of Supercomputing*, pp. 1–33, 2016.

[17] A. Chien, P. Balaji, P. Beckman, N. Dun, A. Fang, H. Fujita, K. Iskra, Z. Rubenstein, Z. Zheng, R. Schreiber, J. Hammond, J. Dinan, I. Laguna, D. Richards, A. Dubey, B. van Straalen, M. Hoemmen, M. Heroux, K. Teranishi, and A. Siegel, "Versioned distributed arrays for resilience in scientific applications: Global view resilience," *Procedia Computer Science*, vol. 51, pp. 29 – 38, 2015.

[18] P. G. Bridges, M. Hoemmen, K. B. Ferreira, M. A. Heroux, P. Soltero, and R. Brightwell, "Cooperative application/os dram fault recovery," in *Proceedings of the 2011 International Conference on Parallel Processing - Volume 2*, ser. Euro-Par'11. Springer-Verlag, 2011, pp. 241–250.

[19] K.-H. Huang and J. Abraham, "Algorithm-based fault tolerance for matrix operations," *IEEE Transactions on Computers*, vol. C-33, no. 6, pp. 518 –528, june 1984.

[20] G. Bosilca, R. Delmas, J. Dongarra, and J. Langou, "Algorithmic Based Fault Tolerance Applied to High Performance Computing," *CoRR*, 2008.

[21] M. Hoemmen and M. A. Heroux, "Fault-tolerant iterative methods via selective reliability," in *Proceedings of the 2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC). IEEE Computer Society*, vol. 3, 2011, p. 9.