Received: 00 Month 0000 | Revised: 00 Month 0000 | Accepted: 00 Month 0000

DOI: xxx/xxxx

EXAMPI2017 SPECIAL ISSUE

Lazy Fault Detection for Redundant MPI

Corey Ford | Chris Lupo*

Computer Science and Software Engineering Department, California Polytechnic State University, California, USA

Correspondence
*Chris Lupo, Computer Science and Software Engineering Department, California Polytechnic State University, San Luis Obispo California, 93407, USA. Email: clupo@calpoly.edu

Abstract

As the scale of supercomputers grows, it is becoming increasingly important for software to efficiently withstand hardware and software faults. Process replication is one resilience technique, but typical implementations require the replicas to stay closely synchronized with each other. The work presented in this paper extends prior work to provide asynchronous methods of silent data corruption detection that avoid waiting for a full set of messages from sending replicas. We present algorithms to lazily detect faults in replicated MPI applications, allowing for more flexibility in replica scheduling. Evaluation on the NAS Parallel Benchmarks shows that this approach allows applications to complete substantially faster as compared to using a synchronized model, and often as fast as in non-replicated execution.

KEYWORDS:
MPI, Fault Detection, Fault Tolerance, Distributed Systems

1 | INTRODUCTION

High-performance computing (HPC) capabilities are growing quickly, with top supercomputers expected to reach exascale capacity within a decade. The Message Passing Interface (MPI) standard has become a de facto standard for large-scale distributed computing applications due to its portability and performance. The MPI standard is not inherently fault-tolerant, though there has been considerable research in this area.

As scale increases, so does the number of system components that can fail and disrupt application execution. Since the frequency of faults for a single component remains relatively constant, the probability of having no failures in the entire cluster during a given time period approaches zero as the cluster grows. At current scale, the mean time between failures (MTBF) of a system may be on the order of hours or less. Typical scientific HPC applications may have execution times measured in days and minimal tolerance for error. Understanding and efficient mitigation of failure are therefore becoming crucial. Proper and efficient application execution in spite of various faults, including *silent data corruption* (SDC) to memory, is the focus of the subfield of *resilience*, which encompasses a wide range of possible solutions, including process replication.

Process replication performs redundant computations on multiple nodes. The state of a process and its replicas are kept synchronized so that a replica can take over in case of a failure on the primary node. Additionally, replication can provide detection and potentially correction of transient errors on any node, such as silent data corruption. Synchronization between replicas to detect SDC has, in prior work, introduced inefficiencies and required replicas to execute at the same speed.

1.1 | Contributions

In this paper, we demonstrate a performant, practical, and portable extension to prior work on MPI runtime libraries that provide process replication. This extension allows for lazy fault detection in an HPC context. Lazy fault detection, where corrupted data is allowed to be identified as such after it has been used, improves the efficiency and flexibility of process replication while retaining the ability to detect SDC.

We develop algorithms and a practical implementation of lazy fault detection that can be transparently integrated into existing replicated MPI programs, and show that runtime performance is generally better than an equivalent synchronized approach across varied benchmark workloads.

Our extensions are open-source, and can be downloaded from the following location:

https://github.com/Cal-Poly-HPC/RedMPI-Extensions.git

1.2 | Outline

The remainder of this paper is organized as follows. Section 2 introduces the MPI system along with the problem of resilience and its various solutions. Section 3 develops the design of lazy fault detection within a software library, and Section 4 discusses our implementation. The performance of our implementation is evaluated in Section 5. Section 6 reviews related work in resilience. Finally, Section 7 concludes and discusses directions for future work.

2 | BACKGROUND

2.1 | MPI

The Message Passing Interface (MPI) is a standardized interface providing communication and other capabilities to parallel programs. The MPI standard defines APIs in C and Fortran; implementations typically also provide C++ interfaces, and bindings to many other languages are available. The features of MPI discussed in this paper are available in version 1 of the MPI standard (1), and do not include additional features introduced in later versions (2, 3).

There are several core MPI operations providing point-to-point communication. The work presented in this paper deals primarily with the following operations.

- `MPI_Isend` starts a nonblocking send to another process. The rank of the receiving process is specified, along with the memory location, size, and type of a buffer to send. This initializes an `MPI_Request` object.
- `MPI_Irecv` starts a nonblocking receive from another process. The rank of the sending process is specified, along with the memory location, size, and type of a buffer to be filled with the received message. This initializes an `MPI_Request` object.
- `MPI_Wait` takes a previously initialized `MPI_Request` object and blocks until the operation is complete. For send requests, a return from `MPI_Wait` guarantees that the buffer is no longer needed and can be reused or deallocated by the application. For receive requests, a return from `MPI_Wait` indicates that the buffer has been filled with a received message.
- `MPI_Test` takes a previously initialized `MPI_Request` object and sets a flag to indicate whether the operation is complete. A return from `MPI_Test` with $*$flag `==` 1 gives the same guarantees as a return from `MPI_Wait`.

The semantics of several other common MPI operations that initiate or determine the status of communications can be defined in terms of those already mentioned.

2.2 | Resilience

Resilience in HPC is a complex research area (4, 5). A wide variety of different approaches have been developed.

We first introduce some necessary terminology for precisely defining the problem of resilience, following a well-established taxonomy (6).

- A *fault* is an erroneous hardware behavior. Faults include transistor malfunctions and memory bits that change state or remain stuck. *Silent data corruption* (SDC) is a specific class of fault.
- An *error* is an erroneous system state caused by a fault. Errors include incorrect values stored in registers or memory.
- A *failure* is an erroneous system behavior resulting from an error. Failures include incorrect results (*Byzantine* failures) and node crashes (*fail-stop* failures).

Given these definitions, resilience is approximately synonymous with *fault-tolerance*, with the goal of preventing failures or mitigating their impact. While faults cannot be prevented in general, there are techniques to prevent some faults from silently causing errors. Software checksums might catch errors before they cause failures. When these resilience techniques are unsuccessful, failures will remain and must be addressed for the benefit of the application.

Substantial work has been done to analyze the faults that occur in real HPC systems and their components. Analysis of logs in production HPC environments give statistics on the frequency of various types of errors (7, 8, 9, 10, 11, 12). Characterization of individual element failures can also be used in resource allocation to minimize faults (13).

2.3 | Checkpointing

Checkpointing (commonly referred to as *checkpoint/restart*) is a long-standing resilience technique, where application state is periodically saved to persistent storage. When a fault is detected, the state is restored from the most recent checkpoint and execution continues.

Checkpointing can be traced back several decades (14). More recently, Elnozahy provides a good survey (15). One widely used implementation of checkpointing is Berkeley Lab Checkpoint / Restart (BLCR) (16), which uses a Linux kernel module to save process state to a filesystem and integrates with MPI implementations to coordinate checkpointing across nodes.

The efficiency of checkpointing is a significant concern. If coordinated checkpointing is performed at the level of the entire system, both the frequency of failures and the resources required to store a checkpoint increase linearly with the number of nodes. With sufficiently many nodes, most time will be spend on checkpointing, restarting, and recomputing lost work. Efficiency can be improved somewhat through incremental checkpointing approaches and compression of checkpoint storage (17). The optimal checkpointing interval has been derived in terms of the cost and frequency of checkpointing (18, 19).

2.4 | Replication

Replication (or redundancy) is another common fault-tolerance approach. Classically, *state machine replication* considers the program as a deterministic state machine and executes replicated copies of it (20, 21).

More generally, communicating processes can be replicated by introducing special handling of their communications. This results in r real processes being mapped to one "virtual" process, with a shared identity but with distinct replica ranks that are hidden from the application. Processes sharing the same replica rank form a replica set that mirrors the equivalent non-replicated system. In doing so, replication can handle failures more seamlessly than checkpointing, at the cost of generally requiring double or triple the computing resources of a single job.

Considering only fail-stop failures, it may suffice to simply run replica processes independently. For Byzantine failures, regular synchronization is necessary. To detect SDCs, assuming that processes execute deterministically, either the messages sent by replica processes can be compared, or their local state can be compared periodically.

RedMPI (22) is a library that implements replication with SDC detection via message comparison for MPI applications. When an application linked with RedMPI makes calls to `MPI_Isend` or `MPI_Irecv`, messages are sent to or received from several replicas of the target virtual process. When the application calls `MPI_Wait` on a receive request, the library waits for all messages to be received and compares them for consistency, choosing the majority value to return to the application if possible.

Unfortunately, this comparison introduces synchronization between replica processes, requiring them to execute at the same speed and in lockstep. Lazy (asynchronous) comparison of messages provides similar fault-detection benefits while reducing synchronization and allowing for more flexibility in replica process execution.

3 | DESIGN

Our primary goal in implementing lazy fault detection for MPI applications is to avoid the need for a message receiver to wait for multiple sending nodes, as in RedMPI. This enables the receiver to proceed as soon as one message is received, avoiding further delay in the common case that this message is not corrupted. We assume the same model of message corruption proposed for RedMPI: SDC affects MPI send buffers, or other values in memory from where corruption eventually propagates to send buffers.

This model allows combining SDC detection with shadow computing (23), where only the primary replica processes proceed at full speed and others can be operated at lower power. Slower replicas would provide eventual confirmation of results; additionally, if two replica processes complete in agreement, any further replicas can be terminated entirely. Lazy fault detection may even (subject to network limitations) allow an application to complete faster than without replication, since only the fastest replica response is required at each step.

As requirements, this design must preserve the expected semantics of MPI calls and error correction:

1. If `MPI_Recv` or `MPI_Wait` returns, or `MPI_Test` indicates completion, the application's buffer must contain a valid received message.
2. If, subsequently, the value that was placed in the application's buffer is determined to be incorrect, corrective action must be taken.

3. The application cannot exit until all messages have been verified.

With asynchronous fault detection, any call to `MPI_Recv`, or to the `MPI_Wait`/`MPI_Test` family of functions with a nonblocking receive request, is satisfied by any single message received from a sending replica. Further replica messages, which may support or contradict the original, are processed opportunistically at later times. If the application is confirmed to have proceeded with a correct message, no action needs to be taken (and any remaining messages can be ignored), whereas if the initial message is found to be incorrect, correction can be initiated. To satisfy the final requirement, when the application calls `MPI_Finalize`, the library waits until all messages have been verified.

The messages considered in satisfying the first two requirements must correspond to (and, absent corruption, match) the messages that would have been received in synchronized or non-replicated execution. The MPI standard includes a non-overtaking requirement that helps to guarantee this.

> (3.5) Messages are *non-overtaking*: If a sender sends two messages in succession to the same destination, and both match the same receive, then this operation cannot receive the second message if the first one is still pending. If a receiver posts two receives in succession, and both match the same message, then the second receive operation cannot be satisfied by this message, if the first one is still pending.
>
> …
>
> (3.7.4) Nonblocking communication operations are ordered according to the execution order of the calls that initiate the communication. The non-overtaking requirement of Section 3.5 is extended to nonblocking communication, with this definition of order being used. (3)

Lazy fault detection does not change the order in which communications are initiated at either the sender or receiver. Therefore, by the non-overtaking requirement, it does not affect which messages will satisfy which receive calls. In turn, since replicated sending as in RedMPI sends all replica messages in immediate succession, the ordering of these corresponds directly to the ordering of messages in non-replicated execution.

3.1 | Correction

If message corruption can be detected before the message is returned to the application, as in RedMPI, correction is simple: replace the message contents with the correct value. A lazily detected fault, however, is not readily corrected, since the process may have already used the incorrect value and performed further computation and communication. Instead, some form of rollback is required so that the process can proceed as if the correct value was initially received.

One simple approach would be to integrate the SDC detection protocol with a standard checkpoint/restart system, and when a corrupted message is detected, restart the entire application from a point before this message was received. However, a full application restart costs substantial computation time, and this has identical scaling behavior to the checkpointing system alone.

It would also suffice to roll back just the receiving process to the point just before receiving the corrupted message, and replay this and the following messages received up until the point of detection. Given conservative handling of requests, any messages received later than the corrupted one will also be validated later and thus still available in memory for correction. Outgoing messages sent by the process during replay can be discarded; if the original versions of these were tainted by the initial corruption, their receivers will be responsible for performing their own correction if necessary.

For more general memory corruption, it is impossible to determine with certainty when the corruption occurred by monitoring messages; a from-scratch restart or rollbacks to successively earlier points would be necessary.

Fault correction is not part of the implementation or design goal for the lazy fault detection extensions presented in this paper. This subsection serves only to illustrate what may be involved in incorporating correction for this lazy detection scheme.

4 | IMPLEMENTATION

We extended the RedMPI library to provide asynchronous methods of SDC detection that avoid waiting for a full set of messages from sending replicas.

As in RedMPI, this functionality is achieved using the MPI profiling layer, enabling compatibility with any existing MPI implementation or application. In addition to exporting `MPI_*` symbols for functions, MPI implementations also export a parallel set of `PMPI_*` symbols that reference the same functions. Our library, like RedMPI, exports new `MPI_*` symbols while using the `PMPI_*` symbols internally.

RedMPI tracks outstanding send and receive requests in an internally managed array. Each request entry references the real underlying MPI requests and, in the case of receive requests, receive buffers for each of these.

4.1 | Async Method

Our first extension is an *Async* SDC detection method added to RedMPI. The *Async* method functions similarly to the existing *AllToAll* method: complete messages are sent from each sending replica to each receiving replica, and receiving replicas perform error detection and correction independently of each other.

To support asynchronous operation, we extended the library to also track, when using the *Async* method,

- whether a request entry has returned an initial result to the application (and is therefore *incomplete*)
- which request's buffer was used to return an initial result
- a pointer to the application's buffer, which is independent of any of the underlying receive buffers
- hashes of each received buffer, for more efficient comparisons across multiple rounds of verification

```
Persistent state: requests, buffers, appBuf, firstIndex
  for i = 0, r − 1 do
      completed[i] ← PMPI_Test(requests[i])
  end for
  if firstIndex is unset then
      firstIndex ← smallest i such that completed[i]
      appBuf ← buffers[firstIndex]
  end if
  if all completed then
      if a majority of buffers match buffers[firstIndex] then
          free request information
      else if a majority of buffers match then
          initiate correction using majority value
      else
          abort execution
      end if
  else
      mark request as incomplete
  end if
```

FIGURE 1 Async integrity verification algorithm

The core algorithmic addition is an integrity-verification routine for the *Async* method, which may be invoked repeatedly on a single request. When provided with an application receive request, the routine operates as outlined in Figure 1 .

Asynchrony is introduced in the modified implementation of `MPI_Wait`. The new version performs a `PMPI_Waitany` instead of a `PMPI_Waitall` on the underlying requests, and invokes the integrity-verification routine (thus returning the first received buffer immediately to the application).

To fully decouple replicas, send requests must also be made asynchronous; otherwise, a fast process may be blocked waiting for a single slow receiving replica to post a matching receive call. Since a successful return from `MPI_Wait` for a send request indicates that the application is allowed to reuse the send buffer, this necessitates making an internal copy of each send buffer for the underlying requests to use. When using the *Async* method, `MPI_Isend` makes this copy, and `MPI_Wait` returns immediately for a send request. The integrity-verification routine cleans up such request entries if all underlying requests have completed.

To handle later-arriving replica messages, every execution of `MPI_Wait` first invokes the integrity-verification routine on each *incomplete* request. The execution of `MPI_Finalize` similarly invokes integrity verification, after performing a `PMPI_Waitall` to catch all outstanding messages, on each *incomplete* request. Figure 2 gives an overview of the steps that may be taken in processing an application receive request `req` in terms of operations on its underlying requests `reqs`.

Compared to *AllToAll*, *Async* requires a single extra message buffer to be allocated for each request, plus r additional small hash buffers and several more bytes of information. More importantly, these allocations will tend to have greater temporal overlap, since the application may create

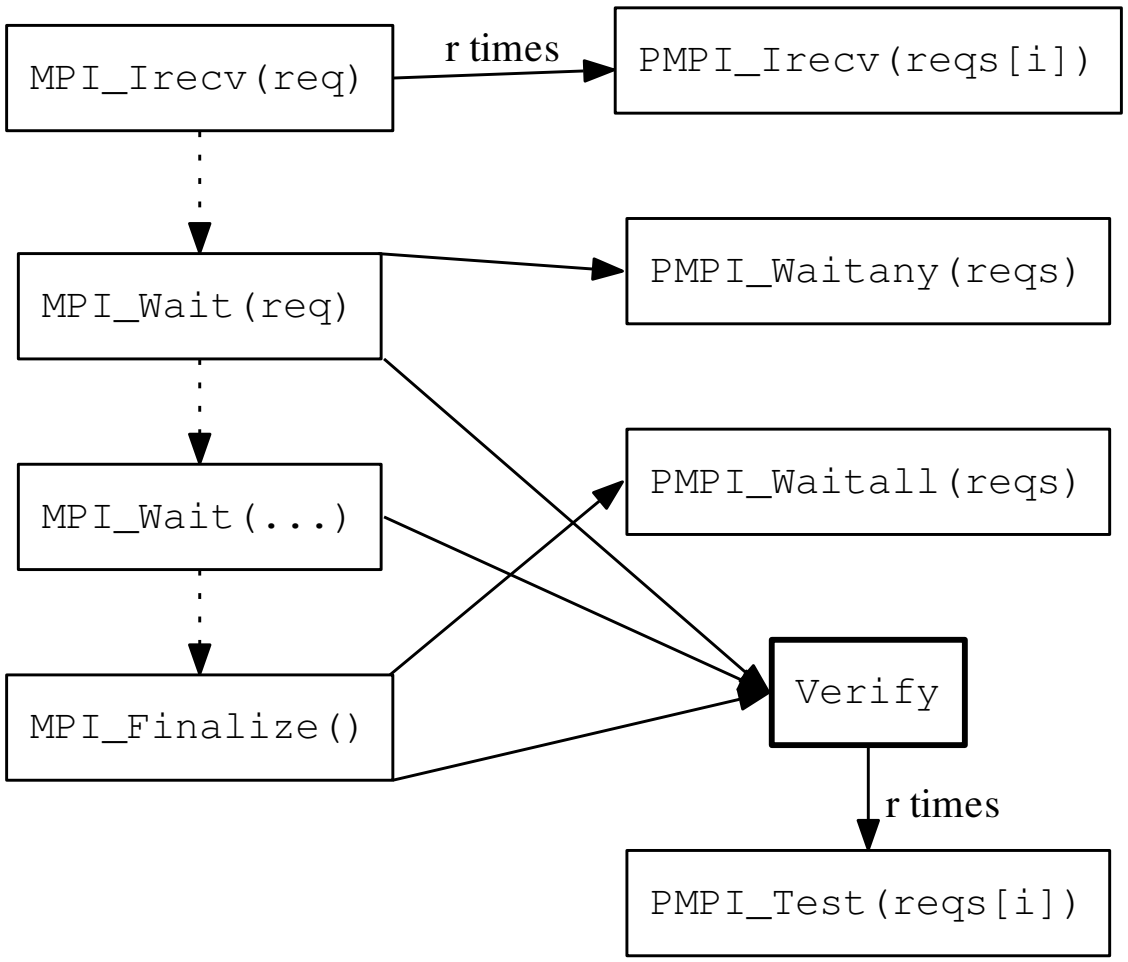

FIGURE 2 Stages of handling a receive request using Async detection

new requests while others are still in an *incomplete* state. The execution time of the integrity-verification algorithm is minimal, but it may in general be invoked an indeterminate number of times on a single request before successfully completing.

4.2 | AsyncHash Method

Our second extension is an *AsyncHash* SDC detection method based on the *MsgPlusHash* method in RedMPI. In *MsgPlusHash*, complete messages are only sent to the replica with the same replica rank k as the sender, and a hash of the message contents is sent to the replica with rank $(k + 1)$ mod r. Receivers check the hash against the message received and, if necessary, coordinate with adjacent replicas to determine the correct message contents. In *AsyncHash*, the hash message is allowed to be sent and received without blocking the application.

Additional per-request state for the *AsyncHash* method consists of

- an *incomplete* flag, as for *Async*
- a locally-computed hash of the full message received

```
Persistent state: hashReq, hashBuf, appBuf, localHash
  if localHash is unset then
      localHash ← H(appBuf)
  end if
  completed ← PMPI_Test(hashReq)
  if completed then
      if hashBuf = localHash then
          free request information
      else
          initiate correction
      end if
  else
      mark request as incomplete
  end if
```

FIGURE 3 AsyncHash integrity verification algorithm

The integrity-verification routine for *AsyncHash* is shown in Figure 3 . Once the hash message has been received, the routine checks for consistency before cleaning up the request.

With *AsyncHash*, `MPI_Wait` invokes `PMPI_Wait` only to wait for the full message to be sent or received. This is safe since the hash buffers for both sender and receiver are not exposed directly to the application. Receipt of the hash message is checked in later calls to `MPI_Wait` and `MPI_Finalize`, as with *Async*.

For SDC correction, *AsyncHash* would require coordination between two replica processes. If corruption occurs at a single sending process, two receiving processes will discover mismatched hashes, but the higher-ranked replica will have a correct message and the lower-ranked replica a matching hash from a distinct sender. These can be used to determine the correct message contents and perform correction at the lower-ranked replica, which will have proceeded in its execution with a corrupted message.

Unlike the *Async* method, *AsyncHash* does not allow the application to complete faster than any single set of processes, since each process is dependent upon the rest of its replica set (those processes sharing the same replica rank) for data. However, it does admit a shadow computing configuration in which some replica sets operate at higher power than others. In this configuration, the faster replica sets will buffer hash messages to be exchanged with the slower.

The additional storage requirements for *AsyncHash* are small and constant, and the integrity-verification algorithm is very simple, but the same considerations regarding allocation lifespan apply as for *Async*.

5 | EVALUATION

We evaluated the performance of our modified SDC detection methods relative to the original RedMPI methods and to non-replicated Open MPI 1.6. This measured the performance impact only in the absence of faults, since SDC correction was not implemented. Two different environments were used for evaluation:

1. 32 nodes, each with a 2.3GHz 14-core Intel Xeon processor and 32GB of RAM, running Linux 2.6 and connected via a gigabit Ethernet switch. We launched up to 12 processes per node for effective utilization without interference from system background tasks.

2. 27 nodes[1], each with a 4-core NVIDIA Tegra K1 processor and 2GB of RAM, running Linux 3.10 and connected via a gigabit Ethernet switch. This environment is the Astro cluster designed at Cal Poly (24). We launched up to 4 processes per node.

The target applications were selected from the NAS Parallel Benchmark suite (NPB) (25). The MPI implementations of NPB benchmarks use a simple subset of MPI features:

- point-to-point nonblocking communications, (does not include `MPI_Test` or its variants)

[1]Although 46 nodes are available in this environment, at most 27 were used in these executions due to benchmark sizing constraints.

TABLE 1 Execution times (seconds) for NPB (Environment 1)

Benchmark (Class)	Procs	Open MPI	AllToAll		Async		Improvement		MsgPlusHash		AsyncHash		Improvement	
		1x	2x	3x	2x	3x	2x	3x	2x	3x	2x	3x	2x	3x
BT (B)	36	61.52	215.25	399.67	82.90	117.56	61.5%	70.6%	67.48	67.85	60.50	58.52	10.3%	13.8%
	64	73.87	213.82	359.82	82.49	100.39	61.4%	72.1%	94.36	92.53	72.14	79.85	23.5%	13.7%
	121	80.09	187.13	317.48	83.76	99.33	55.2%	68.7%	97.28	118.70	90.29	104.34	7.2%	12.1%
CG (C)	32	86.42	578.17	968.71	205.40	318.80	64.5%	67.1%	198.54	220.15	146.73	130.29	26.1%	40.8%
	64	172.92	614.23	1078.34	213.96	344.13	65.2%	68.1%	247.40	294.24	174.54	169.72	29.5%	42.3%
	128	295.35	646.69	1157.28	193.90	233.12	70.0%	79.9%	374.68	414.55	285.29	282.79	23.9%	31.8%
EP (D)	32	175.99	177.17	178.62	178.52	179.82	-0.8%	-0.7%	178.48	178.82	180.12	176.64	-0.9%	1.2%
	64	91.01	91.17	90.89	91.32	90.10	-0.2%	0.9%	91.56	90.84	91.28	91.36	0.3%	-0.6%
	128	46.54	48.20	49.15	48.79	48.84	-1.2%	0.6%	48.35	51.95	48.56	55.09	-0.4%	-6.0%
IS (D)	32	263.10	560.24	898.02	541.45	858.65	3.4%	4.4%	297.12	299.30	293.88	289.61	1.1%	3.2%
	64	185.63	366.21	595.17	314.80	491.52	14.0%	17.4%	189.01	268.07	171.37	177.29	9.3%	33.9%
	128	150.14	276.94	445.46	194.10	296.71	29.9%	33.4%	286.60	259.64	140.73	147.41	50.9%	43.2%
LU (B)	32	33.03	91.49	207.56	45.51	65.78	50.3%	68.3%	40.52	49.17	51.90	52.29	-28.1%	-6.3%
	64	26.90	114.72	227.96	43.75	61.68	61.9%	72.9%	67.61	87.09	49.82	54.92	26.3%	36.9%
	128	56.33	139.93	267.97	53.19	54.42	62.0%	79.7%	127.32	152.41	81.20	90.10	36.2%	40.9%
SP (B)	36	117.11	374.56	634.59	143.82	200.20	61.6%	68.5%	120.28	115.75	100.76	99.30	16.2%	14.2%
	64	148.76	404.16	688.51	153.46	182.74	62.0%	73.5%	176.59	174.46	138.23	146.89	21.7%	15.8%
	121	144.59	306.32	495.65	147.35	160.74	51.9%	67.6%	169.78	214.90	160.69	188.91	5.4%	12.1%

- blocking communications, which RedMPI implements in terms of nonblocking communications
- various collective operations, which RedMPI implements in terms of point-to-point communications

The NPB programs exhibit strong scaling: the amount of work for a given problem size does not depend on the number of processes. Therefore, varying the number of processes changes the communication/computation ratio. In addition, the different benchmarks exhibit a range of communication patterns. The FT benchmark (in Environment 1) and the MG benchmark (in both environments) were excluded due to difficulties encountered when running these under RedMPI.

5.1 | Runtime Measurements

We executed each benchmark in each environment, first using only Open MPI, then using RedMPI under double and triple replication and with each of the four SDC detection methods (*AllToAll* and *MsgPlusHash* from RedMPI along with our *Async* and *AsyncHash*). We selected three different process counts for each benchmark, as appropriate for that benchmark's requirements. Benchmark classes (problem sizes) were selected to give approximately a 1–2 minute running time under Open MPI (notably with the IS benchmark) to fit within system memory constraints.

Table 1 and Table 2 show measurements of runtime for each execution in the two environments. Time was measured across the entire execution of the `mpirun` command rather than using the benchmark code's own timing, which uses `MPI_Wtime` and thus (due to RedMPI's handling of this operation) only represents the elapsed time for the first replica set. Figures 4 and 5 chart runtimes for double and triple replication respectively, normalized against the runtime under Open MPI, and restricted to Environment 1 and the process count of 64 common to all benchmarks. Figures 6 and 7 show similar data for Environment 2 and a process count of 16.

We see that, in our environments, there is a significant penalty for most benchmarks due to the doubled or tripled message traffic with *AllToAll*, and a lesser one for *MsgPlusHash*. In general, our *Async* and *AsyncHash* SDC detection methods alleviate these respective overheads substantially.

TABLE 2 Execution times (seconds) for NPB (Environment 2)

Benchmark (Class)	Procs	Open MPI	AllToAll		Async		Improvement		MsgPlusHash		AsyncHash		Improvement	
		1x	2x	3x	2x	3x	2x	3x	2x	3x	2x	3x	2x	3x
BT (B)	16	101.53	200.01	321.81	132.35	163.10	33.8%	49.3%	118.72	120.64	113.11	113.53	4.7%	5.9%
	36	97.61	222.97	377.83	88.75	117.35	60.2%	68.9%	114.59	129.21	99.96	101.78	12.8%	21.2%
CG (C)	16	163.95	429.28	779.35	299.17	406.45	30.3%	47.8%	199.04	219.87	191.56	191.86	3.8%	12.7%
	32	186.25	478.98	885.79	180.65	253.85	62.3%	71.3%	214.59	255.85	154.78	153.66	27.9%	39.9%
EP (C)	16	41.22	41.45	41.24	41.06	41.58	0.9%	-0.8%	41.06	41.20	41.07	41.19	-0.0%	0.0%
	32	21.07	21.46	21.73	21.24	41.51	1.0%	-91.0%	21.31	21.38	21.27	21.43	0.2%	-0.2%
FT (B)	16	53.47	100.08	155.93	97.66	145.12	2.4%	6.9%	56.01	55.78	55.93	56.19	0.1%	-0.7%
	32	58.90	63.51	96.13	57.73	83.59	9.1%	13.0%	35.46	35.52	35.84	35.62	-1.1%	-0.3%
IS (C)	16	24.50	53.80	82.62	52.34	78.46	2.7%	5.0%	28.81	28.79	29.11	28.59	-1.0%	0.7%
	32	29.24	36.03	57.47	32.22	47.58	10.6%	17.2%	19.36	19.58	18.94	19.82	2.2%	-1.2%
LU (B)	16	99.51	153.07	225.65	211.55	240.66	-38.2%	-6.7%	133.15	139.17	216.30	219.76	-62.4%	-57.9%
	32	54.88	140.33	214.05	153.09	182.45	-9.1%	14.8%	94.03	98.24	187.82	192.84	-99.7%	-96.3%
SP (A)	16	84.22	203.52	323.69	75.24	101.08	63.0%	68.8%	116.84	124.69	86.09	87.60	26.3%	29.7%
	36	97.48	223.61	363.09	58.05	75.27	74.0%	79.3%	120.10	131.23	101.64	105.27	15.4%	19.8%

For most benchmarks, the improvement is consistent across a range of communication/computation ratios, with IS especially sensitive to this ratio and LU exhibiting a few anomalies. The EP benchmark, which performs minimal communication, verifies the reliability of our measurements and shows that there is negligible fixed overhead introduced by SDC detection. For BT, CG, IS, and SP, the runtime using *AsyncHash* is comparable to or slightly better than the baseline Open MPI runtime; the apparent improvements are likely due to different process placement under replication.

Occasional negative effects are also observed, notably for triple-replicated IS; we conjecture that the memory-intensive nature of this benchmark may cause it to perform worse in the presence of the additional per-request allocations needed to support replication.

5.2 | Memory Overhead Measurements

Our *Async* and *AsyncHash* SDC detection methods make a time-memory tradeoff by continuing execution before fully verifying and freeing an application request. To measure the impact on memory usage, we instrumented the allocation of internal request structures in RedMPI across all processes in a replicated job. With this instrumentation, each benchmark was executed again in Environment 1 with double replication and a process count of 64 under each of the SDC detection methods.

The resulting data, after applying a moving average, sampling this every second, and normalizing to a per-process count, is plotted for each benchmark in the sub-plots of Figure 8 . Data for *AllToAll* was excluded since it is nearly identical to that for *MsgPlusHash* but extends for a longer time period in most cases.

Most of the benchmarks appear to frequently block on requests, so under *MsgPlusHash* (or *AllToAll*) there are at most a handful of request entries allocated at a time. As a general trend, *Async* maintains approximately an order of magnitude more requests, while *AsyncHash* shows more dramatic variation over time but has up to 20 times again as many requests allocated. The exceptions are EP, which makes no MPI requests during the majority of its execution (and therefore also does not provide our integrity-verification routine an opportunity to clean up leftover initial requests), and IS, which performs all-to-all communications that appear as 128 individual point-to-point requests and that are not effectively streamlined by our asynchronous methods.

The largest overall memory overhead observed in this data (excluding IS, which has large memory usage regardless of SDC detection) is for the CG benchmark using the *Async* SDC detection method, where an average of 16 requests per process were concurrently allocated at one point in time. A typical message buffer in this benchmark is 150KB, giving a per-process memory overhead of 3.6MB for redundant buffers (one for send requests, two for receive requests) in this situation. With the *AsyncHash* method, request buffers are not copied, so a single allocated request

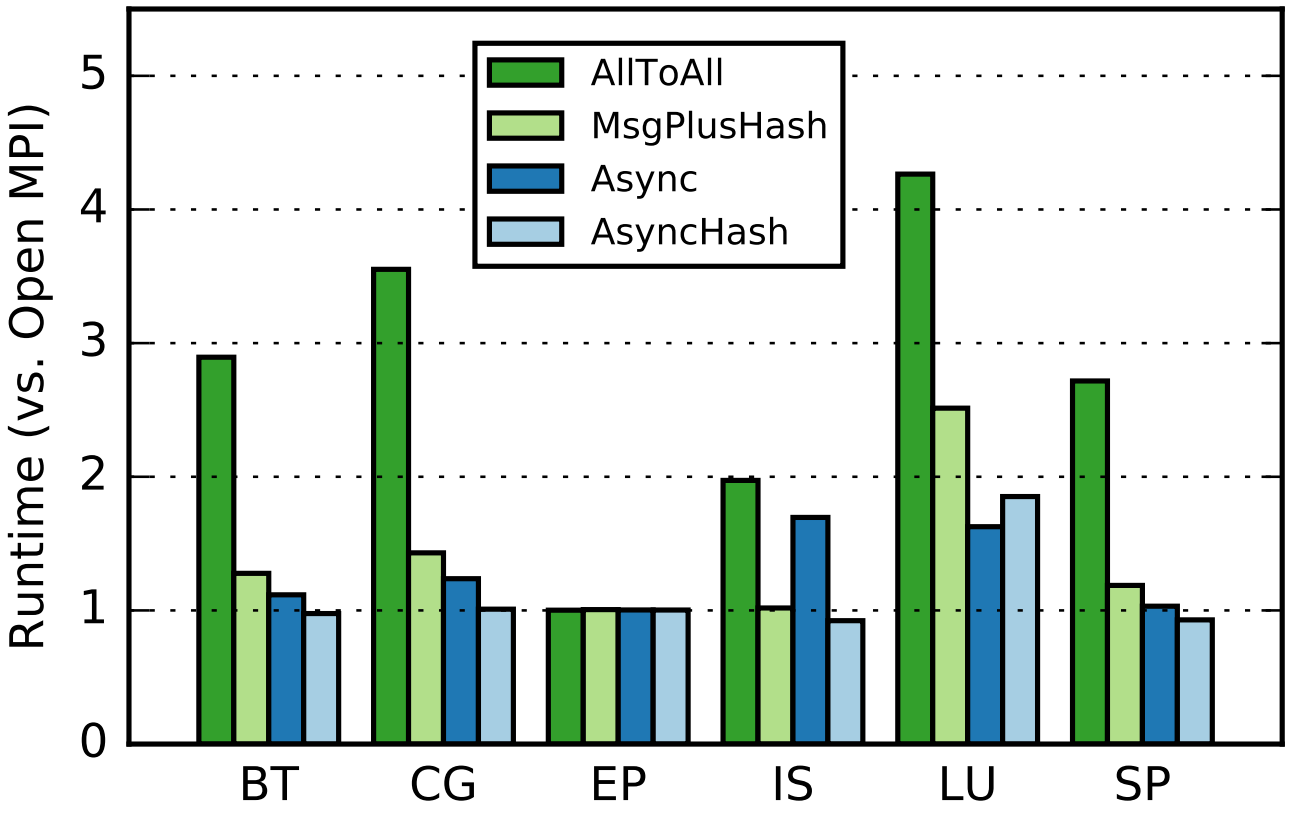

FIGURE 4 Runtime at double replication (64 processes, Environment 1)

represents less than 200 bytes of memory. The largest number of allocated requests observed in the *AsyncHash* data is 330 per process for the SP benchmark, resulting in approximately 66KB memory overhead. We believe these overheads are inconsequential for current systems and do not present a hindrance to the use of lazy fault detection.

6 | RELATED WORK

6.1 | Replicated MPI Implementations

Several libraries implementing transparent replication for MPI applications have previously been developed. One early implementation is rMPI, which and replicates MPI processes on multiple nodes to protect against fail-stop failures (26, 27, 17). MR-MPI is a more sophisticated implementation of redundancy featuring MPI collectives, wildcard receives, and partial replication (28). PaRep-MPI (29) implements adaptive replication based on failure prediction. All three of these libraries use the MPI profiling layer to transparently interpose between existing MPI implementations and applications.

RedMPI extends the approach of MR-MPI to handle silent data corruption (22). Its core functionality consists of protocols for MPI point-to-point communication that provide SDC detection. The first method, *AllToAll*, directs (by instrumenting `MPI_Isend`) sent messages to all replicas of the receiving process. The second, *MsgPlusHash*, saves bandwidth by sending a full message to only one receiving replica and a hash to another. The same combination of messages is seen when receiving; `MPI_Irecv` is instrumented to start multiple receive requests, while `MPI_Wait` waits for all these requests to complete and then checks for consistency. Either of these methods provides SDC detection under double replication and SDC correction (by voting) under triple replication. RedMPI supports many MPI features including collectives and wildcards.

VolpexMPI (30) and FMI (31) are complete runtime implementations that provide replication, replacing a normal MPI implementation. These approaches consider only fail-stop failures, so while they do not require close synchronization between replicas, SDC will not be detected.

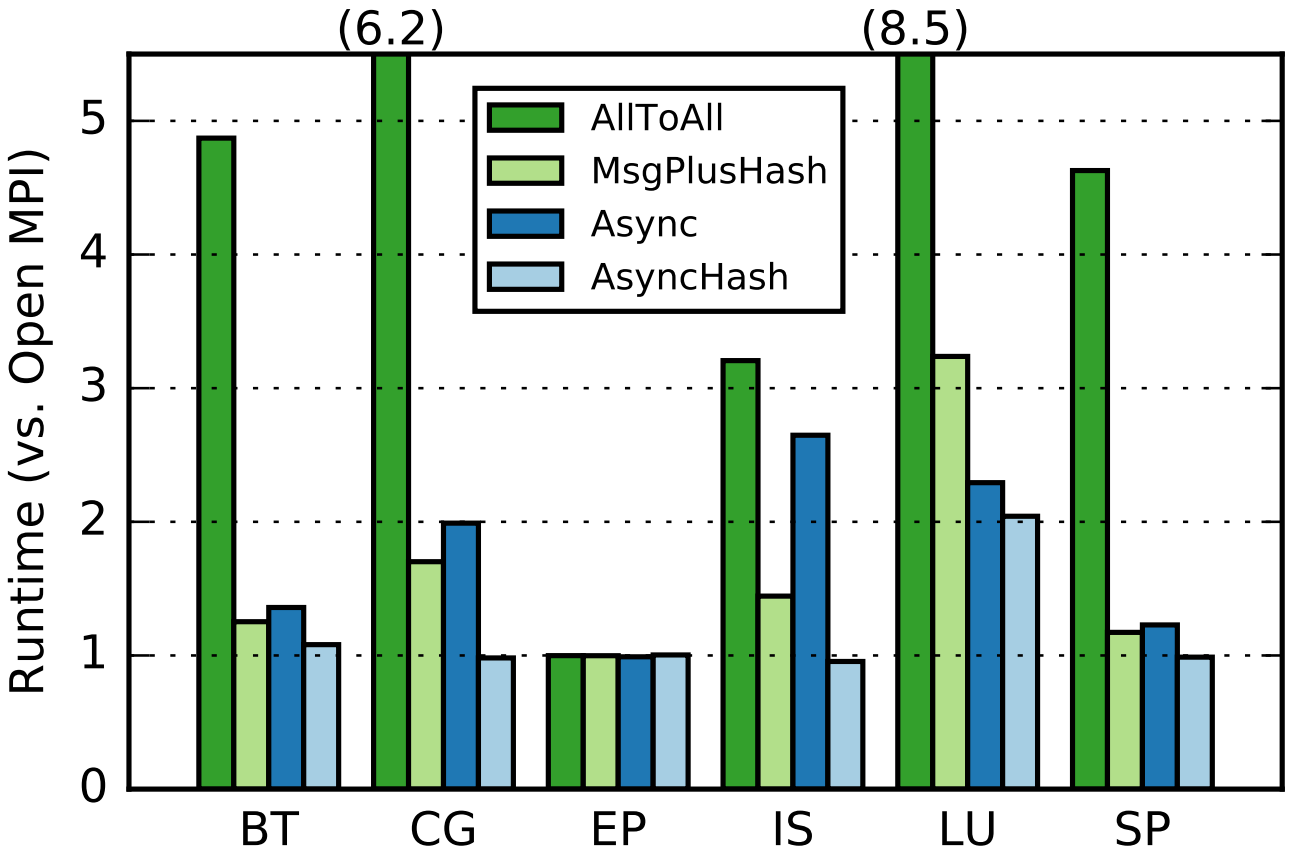

FIGURE 5 Runtime at triple replication (64 processes, Environment 1)

Purdy et al. (32) propose replicating MPI processes to improve runtime in the face of highly variable latency in a public cloud environment. This work does not include an implementation and does not consider SDC detection.

6.2 | Process Management

Another body of work within resilience concentrates on management of the replica processes themselves. Shadow computing executes non-primary replica processes at slower speeds (23, 33, 34). When the primary processes are successful, the replicas can be terminated early. When one of the primary processes fails, a replica can take over and accelerate to full speed, having only a fraction of the work to make up compared to a complete restart. Allowing for SDC detection within shadow computing is one goal of our work.

Node cloning dynamically creates replicas of running processes (35, 36). This is achieved by copying memory pages to a new node after quiescing MPI communication. Cloning is useful to consolidate divergent node states, or to regain replication after nodes are lost to failures. This complements fault detection by providing a way to recover from even fail-stop failures.

Sliding substitution attempts to give replacement nodes similar communication positions to the originals (37).

6.3 | Redundant Multithreading

Redundant multithreading is similar to distributed replication, but in a shared-memory multithreading context. In redundant multithreading, network failures are not a concern, but transient hardware faults are still important (38). Work by Hukerikar et al. develops redundant multithreading as a fault-detection strategy (39). Later extensions to this work incorporate lazy fault detection (40, 41). This minimizes synchronization between replica threads by periodically saving state to memory buffers for later comparison. We use this design as inspiration for our MPI implementation of lazy fault detection.

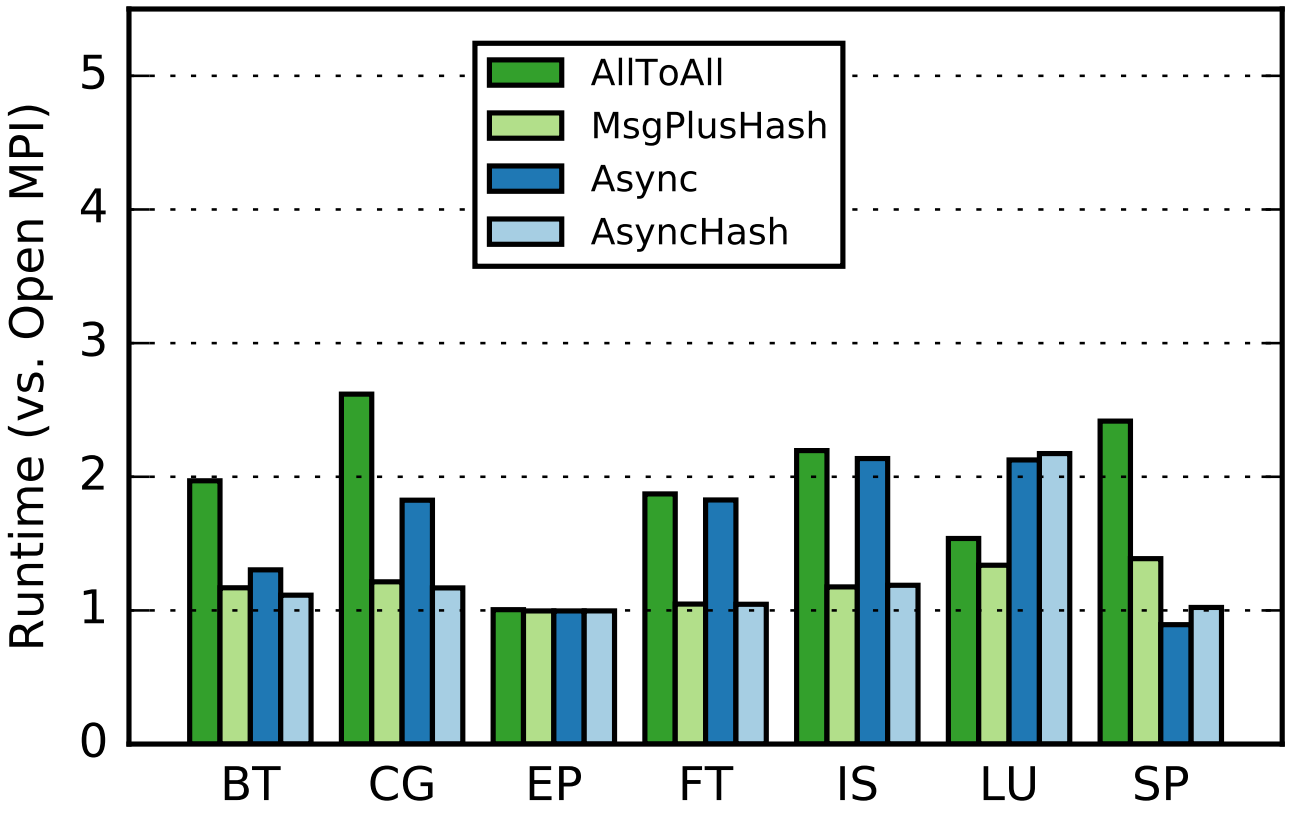

FIGURE 6 Runtime at double replication (16 processes, Environment 2)

7 | CONCLUSIONS & FUTURE WORK

We have developed an approach to lazy fault detection for MPI programs that improves on previous replication models in efficiency and scheduling flexibility. The evaluation results suggest that our *Async* and *AsyncHash* methods of SDC detection for RedMPI provide a noticeable performance benefit, matching the speed of non-replicated execution for some benchmarks, with reasonable memory overhead. We believe this represents a valuable advance in addressing resilience.

The most significant area for future work is implementation of SDC correction (discussed in Section 3.1) for lazy fault detection. With correction added, the efficiency of the library can be evaluated again in the presence of injected faults that corrupt MPI messages.

In addition, further MPI features could be made to avoid replica synchronization, though often with a loss of complete consistency between replicas. RedMPI implements features such as wildcard receives, `MPI_Test`, `MPI_Iprobe`, and `MPI_Wtime` by sharing one replica's result with others. A given application might or might not diverge based on the results of these time-sensitive calls. Adapting these calls to use local results would risk replica divergence for some applications that use them, but improve performance for others.

8 | ACKNOWLEDGMENTS

The authors would like to thank Sandia National Laboratories for supporting this work. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration (NNSA) under contract DE-AC04-94AL85000. This work was funded by NNSA's Advanced Simulation and Computing (ASC) Program.

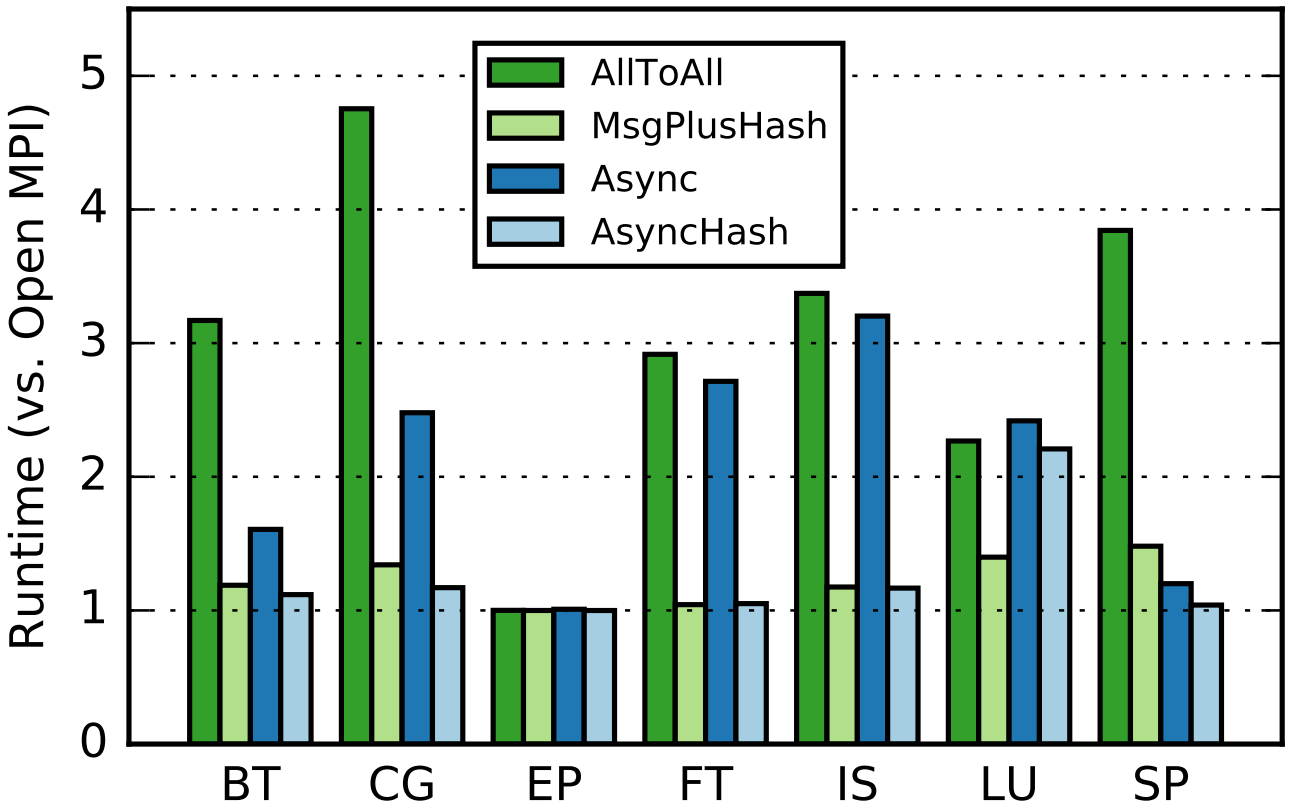

FIGURE 7 Runtime at triple replication (16 processes, Environment 2)

References

[1] *MPI: A Message-Passing Interface Standard: Version 1.3.* : Message Passing Interface Forum; 2008.

[2] *MPI: A Message-Passing Interface Standard: Version 2.2.* : Message Passing Interface Forum; 2009.

[3] *MPI: A Message-Passing Interface Standard: Version 3.1.* : Message Passing Interface Forum; 2015.

[4] Bouteiller Aurélien. Fault-Tolerant MPI. In: Herault Thomas, Robert Yves, eds. *Fault-Tolerance Techniques for High-Performance Computing*, Computer Communications and Networks. Springer International Publishing 2015 (pp. 145–228). DOI: 10.1007/978-3-319-20943-2_3.

[5] Dongarra Jack, Herault Thomas, Robert Yves. Fault Tolerance Techniques for High-Performance Computing. In: Herault Thomas, Robert Yves, eds. *Fault-Tolerance Techniques for High-Performance Computing*, Computer Communications and Networks. Springer International Publishing 2015 (pp. 3–85). DOI: 10.1007/978-3-319-20943-2_1.

[6] Avizienis A., Laprie J.-C., Randell B., Landwehr C.. Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing.* 2004;1(1):11–33.

[7] Cappello Franck, Geist Al, Gropp William, Kale Sanjay, Kramer Bill, Snir Marc. Toward Exascale Resilience: 2014 update. *Supercomputing frontiers and innovations.* 2014;1(1):5–28.

[8] Schneider Fred B.. What Good are Models and What Models are Good?. In: New York, NY, USA: ACM Press/Addison-Wesley Publishing Co. 1993 (pp. 17–26).

[9] Hacker Thomas J., Romero Fabian, Carothers Christopher D.. An analysis of clustered failures on large supercomputing systems. *Journal of Parallel and Distributed Computing.* 2009;69(7):652–665.

[10] Stearley Jon, Ballance Robert. A preliminary report on red storm RAS performance. In: ; 2006; Lugano, Switzerland.

[11] Schroeder Bianca, Gibson Garth A.. Understanding failures in petascale computers. *Journal of Physics: Conference Series.* 2007;78(1):012022.

[12] Oliner A., Stearley J.. What Supercomputers Say: A Study of Five System Logs. In: :575–584; 2007.

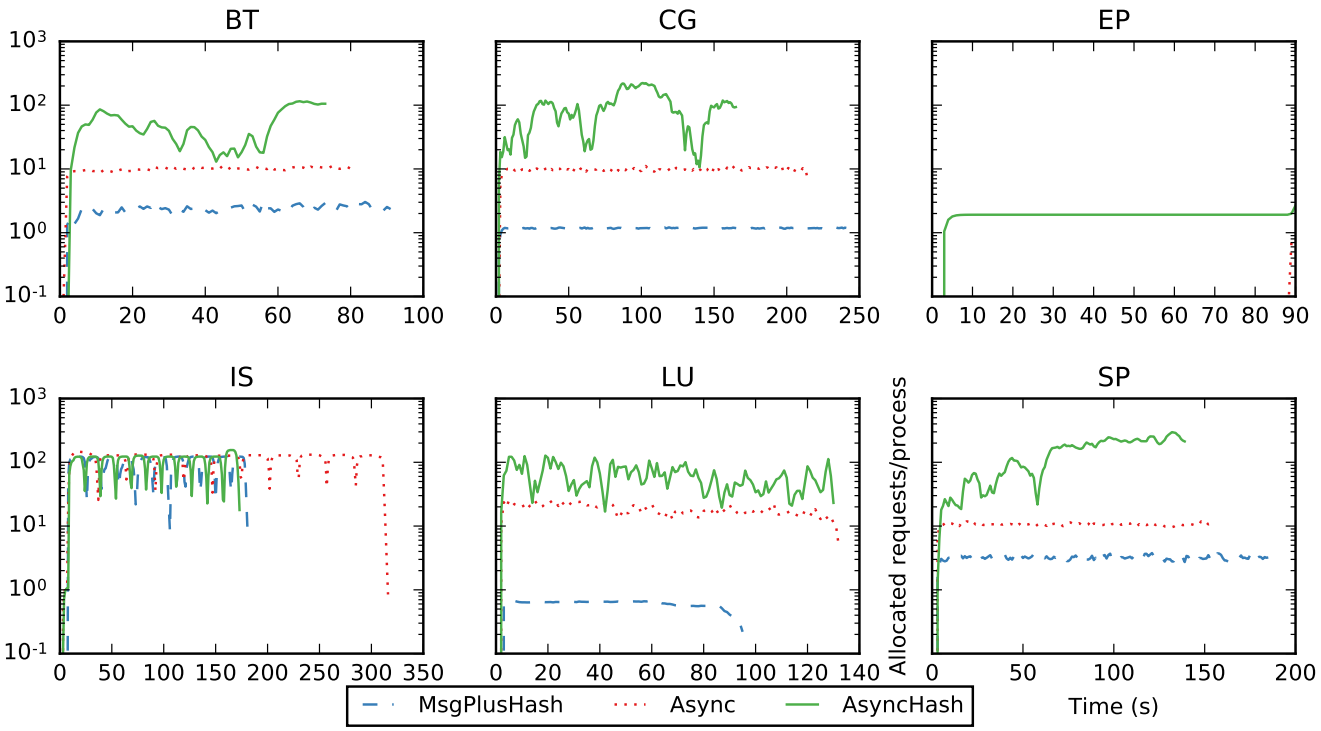

FIGURE 8 Allocated requests per process over time for double-replicated 64-process jobs

[13] Brandt J., Debusschere B., Gentile A., et al. Using Probabilistic Characterization to Reduce Runtime Faults in HPC Systems. In: :759–764; 2008.

[14] Koo R., Toueg S.. Checkpointing and Rollback-Recovery for Distributed Systems. *IEEE Transactions on Software Engineering*. 1987;SE-13(1):23–31.

[15] Elnozahy E. N. (Mootaz), Alvisi Lorenzo, Wang Yi-Min, Johnson David B.. A Survey of Rollback-recovery Protocols in Message-passing Systems. *ACM Comput. Surv.*. 2002;34(3):375–408.

[16] Hargrove Paul H., Duell Jason C.. Berkeley lab checkpoint/restart (BLCR) for Linux clusters. *Journal of Physics: Conference Series*. 2006;46(1):494.

[17] Ferreira Kurt, Riesen Rolf, Oldfield Ron, et al. Keeping checkpoint/restart viable for exascale systems. *Sandia National Laboratories, Tech. Rep. SAND2011-6815*. 2012;.

[18] Young John W.. A First Order Approximation to the Optimum Checkpoint Interval. *Commun. ACM*. 1974;17(9):530–531.

[19] Daly J. T.. A higher order estimate of the optimum checkpoint interval for restart dumps. *Future Generation Computer Systems*. 2006;22(3):303–312.

[20] Schneider Fred B.. Implementing Fault-tolerant Services Using the State Machine Approach: A Tutorial. *ACM Comput. Surv.*. 1990;22(4):299–319.

[21] Kumar Vinit, Agarwal Ajay. HT-Ring Paxos: Theory of High Throughput State-Machine Replication for Clustered Data Centers. *arXiv:1507.04086 [cs]*. 2015;. arXiv: 1507.04086.

[22] Fiala David, Mueller Frank, Engelmann Christian, Riesen Rolf, Ferreira Kurt, Brightwell Ron. Detection and Correction of Silent Data Corruption for Large-scale High-performance Computing. In: SC '12:78:1–78:12IEEE Computer Society Press; 2012; Los Alamitos, CA, USA.

[23] Mills B., Znati T., Melhem R.. Shadow Computing: An energy-aware fault tolerant computing model. In: :73–77; 2014.

[24] Sheen Sean. Astro - A Low-Cost, Low-Power Cluster for CPU-GPU Hybrid Computing. Master's thesisCalifornia Polytechnic State University, San Luis Obispo2016.

[25] Bailey D. H., Barszcz E., Barton J. T., et al. The NAS Parallel Benchmarks. *International Journal of High Performance Computing Applications*. 1991;5(3):63–73.

[26] Stearley Jon R., Riesen Rolf E., Laros III James H., et al. *Increasing fault resiliency in a message-passing environment.*. SAND2009-6753: Sandia National Laboratories; 2009.

[27] Ferreira Kurt, Stearley Jon, Laros James H., et al. Evaluating the Viability of Process Replication Reliability for Exascale Systems. In: SC '11:44:1–44:12ACM; 2011; New York, NY, USA.

[28] Engelmann Christian, Böhm Swen. Redundant Execution of HPC Applications with MR-MPI. In: ACTAPRESS; 2011.

[29] George C., Vadhiyar S.. Fault Tolerance on Large Scale Systems using Adaptive Process Replication. *IEEE Transactions on Computers*. 2015;64(8):2213–2225.

[30] LeBlanc Troy, Anand Rakhi, Gabriel Edgar, Subhlok Jaspal. VolpexMPI: An MPI Library for Execution of Parallel Applications on Volatile Nodes. In: Ropo Matti, Westerholm Jan, Dongarra Jack, eds. *Recent Advances in Parallel Virtual Machine and Message Passing Interface*, no. 5759 in Lecture Notes in Computer Science. Springer Berlin Heidelberg 2009 (pp. 124–133). DOI: 10.1007/978-3-642-03770-2_19.

[31] Sato K., Moody A., Mohror K., et al. FMI: Fault Tolerant Messaging Interface for Fast and Transparent Recovery. In: :1225–1234; 2014.

[32] Purdy Scott, Hunt Pete, Bindel David. Process Replication for HPC Applications on the Cloud. 2010;.

[33] Cui Xiaolong, Mills Bryan, Znati Taieb, Melhem Rami. Shadow Replication: An Energy-Aware, Fault-Tolerant Computational Model for Green Cloud Computing. *Energies*. 2014;7(8):5151–5176.

[34] Cui Xiaolong, Znati Taieb, Melhem Rami. Lazy Shadowing: An Adaptive, power-Aware, Resiliency Framework for Extreme Scale Computing. In: ; 2014.

[35] Rezaei A., Mueller F.. Sustained Resilience via Live Process Cloning. In: :1498–1507; 2013.

[36] Rezaei Arash, Mueller Frank. DINO: Divergent Node Cloning for Sustained Redundancy in HPC. In: ; 2015; Chicago, IL.

[37] Hori Atsushi, Yoshinaga Kazumi, Herault Thomas, Bouteiller AurÃl'lien, Bosilca George, Ishikawa Yutaka. Sliding Substitution of Failed Nodes. In: EuroMPI '15:14:1–14:10ACM; 2015; New York, NY, USA.

[38] Mukherjee S.S., Kontz M., Reinhardt S.K.. Detailed design and evaluation of redundant multi-threading alternatives. In: :99–110; 2002.

[39] Hukerikar Saurabh, Diniz Pedro C., Lucas Robert F.. A Case for Adaptive Redundancy for HPC Resilience. In: Mey Dieter an, Alexander Michael, Bientinesi Paolo, et al. , eds. *Euro-Par 2013: Parallel Processing Workshops*, no. 8374 in Lecture Notes in Computer Science. Springer Berlin Heidelberg 2013 (pp. 690–697). DOI: 10.1007/978-3-642-54420-0_67.

[40] Hukerikar S., Diniz P.C., Lucas R.F., Teranishi K.. Opportunistic application-level fault detection through adaptive redundant multithreading. In: :243–250; 2014.

[41] Hukerikar S., Teranishi K., Diniz P.C., Lucas R.F.. An evaluation of lazy fault detection based on Adaptive Redundant Multithreading. In: :1–6; 2014.