Failure Detection and Propagation in HPC systems

George Bosilca*, Aurelien Bouteiller*, Amina Guermouche*,
Thomas Herault*, Yves Robert*†, Pierre Sens‡, and Jack Dongarra* § ¶
*ICL, University of Tennessee, USA †LIP, ENS Lyon, France ‡Sorbonne Univ., UPMC, CNRS, Inria, LIP6, France
§Oak Ridge National Lab., USA ¶Manchester University, UK

***Abstract*—Building an infrastructure for Exascale applications requires, in addition to many other key components, a stable and efficient failure detector. This paper describes the design and evaluation of a robust failure detector, able to maintain and distribute the correct list of alive resources within proven and scalable bounds. The detection and distribution of the fault information follow different overlay topologies that together guarantee minimal disturbance to the applications. A virtual observation ring minimizes the overhead by allowing each node to be observed by another single node, providing an unobtrusive behavior. The propagation stage is using a non-uniform variant of a reliable broadcast over a circulant graph overlay network, and guarantees a logarithmic fault propagation. Extensive simulations, together with experiments on the Titan ORNL supercomputer, show that the algorithm performs extremely well, and exhibits all the desired properties of an Exascale-ready algorithm.**

***Index Terms*—MPI, Failure Detection, Fault-Tolerance**

I. Introduction

Failure detection is a prerequisite to failure mitigation and a key component to any infrastructure requiring resilience. This paper is devoted to the design and evaluation of a reliable algorithm that will maintain, and distribute the updated list of alive resources with a guaranteed maximum delay. We consider a typical High-Performance Computing (HPC) platform in steady-state operation mode. Because in such environments the transmission time can be considered as bounded (although that bound is unknown), it becomes possible to provide a *perfect failure detector* according to the classical definition of [7]. A failure detector is a distributed service able to return the state of any node, alive or dead (subject to a crash)[1]. A failure detector is *perfect* if any node crash is eventually suspected by all surviving nodes, and if no surviving node ever suspects a node that is still alive. Critical fault-tolerant algorithms for HPC, and implementations of communication middleware for unreliable systems rely on the strong properties of perfect failure detectors (see e.g. [9], [14], [5], [6], [19]). Their cost, in terms of computation and communication overhead, as well as their properties in terms of latency to detect and notify failures and of reliability, have thus a significant impact on the overall performance of a fault-tolerant HPC solution.

While we focus primarily on of one the most widely used programming paradigms, the Message Passing Interface (MPI), the techniques and algorithms proposed have a larger scope, and are applicable in any resilient distributed programming environment. We consider the platform as being initially composed of N nodes, but with a high probability, some of these resources will become unavailable throughout the execution. When exposed to the crash of one node, traditional applications would abort. However, the applications that we consider, are augmented with fault tolerant extensions that allow them to continue across failures (e.g. [4]), either using a generic or an application-specific fault tolerant model. The design of this model is outside the scope of this paper, but without loss of generality, we can safely assume that any fault tolerant recovery model requires a robust fault detection mechanism. Our goal is to design such a *robust* protocol that can detect all failures and enable the *efficient repair* of the execution platform.

Platform parameters	
N	Initial number of nodes
τ	Upper bound on the time to transfer a message
Protocol parameters	
η	Period for heartbeats
δ	Time-out for suspecting a failure

TABLE I: List of Notations.

By *repairing the platform*, we mean that all surviving nodes will eventually be notified of all failures, and will therefore be able to compute the list of surviving nodes. The state of the platform where all failed nodes are known to all processes is called a *stable configuration* (note that nodes may not be aware that they are in a stable configuration).

By *robust*, we mean that regardless of the length of the execution, if a set of up to f failures disrupt the platform and precipitate it into an unstable configuration, the protocol will bring the platform back into a stable configuration within $T(f)$ time units (we will define $T(f)$ later in the paper). Note that the goal is not to tolerate up to f failures overall. On the contrary, the protocol will tolerate an arbitrary number of failures throughout an unbounded-length execution, provided that no more than f successive overlapping failures strike within the $T(f)$ time-window. Hence, f induces a constraint on the frequency of failures, and not on the total number of failures.

By *efficiently*, we aim at a low-overhead protocol that limits the number of messages exchanged to detect the faults and repair the platform. While we assume a fully-connected platform (any node may communicate with any other), we use a realistic one-port communication model [3], where a node can send and/or receive at most one message at any time-step. Independent communications, involving distinct sender/receiver pairs, can take place in parallel: however, two messages sent by the same node will be serialized.

[1]We use the words *failure*, *crash*, or *death* indifferently.

SC16; Salt Lake City, Utah, USA; November 2016
978-1-4673-8815-3/16/$31.00

These goals seem contradictory but they only call for a carefully designed trade-off: as shown in [10], [17], [20], system noise created by the messages and computations of the fault detection mechanism, can impose significant overheads in HPC applications, hence the efficiency of the approach must be carefully assessed. The overhead should be kept minimal in the absence of failures, while failure detection and propagation should execute quickly, which usually implies a robust broadcast operation that introduces many messages. The major contributions of this work are as follows:

• A proven algorithm for failure detection, based on a robust protocol that tolerates an arbitrary number of failures, provided that no more than f consecutive failures strike within a time window of duration $T(f)$;

• The protocol has minimal overhead in failure-free operation, with a unique observer per node;

• The protocol achieves failure detection and propagation in logarithmic time for up to $f_{\max} = \lfloor \log n \rfloor - 1$ where n is the number of alive nodes. More precisely, the bound $T(f_{\max})$ is deterministic, and logarithmic in n, even in the worst case;

• All performance guarantees are expressed within a realistic one-port communication model;

• Extensive simulations and experiments with ULFM [4] show very good performance of the algorithm.

The rest of the paper is organized as follows. We start with an informal description of the algorithm in Section II. We detail the model, the proof of correctness and the time-performance analysis in Section III. Then we assess the efficiency of the algorithm in a practical setting, first by reporting on a comprehensive set of simulations in Section IV, and then by discussing experimental results on the ORNL Titan supercomputer in Section V. Section VI provides an overview of related work. Finally, we outline conclusions and directions for future work in Section VII.

II. Algorithm

This section provides an informal description of the algorithm. We refer to Section III for a detailed presentation of the model, a proof of correctness and a time-performance analysis. We maintain two main invariants in the algorithm:

1) Each alive node maintains its own list of known dead resources;
2) Alive nodes are arranged along a ring and each node observes its predecessor in the ring. In other words, the successor/observer receives heartbeats from its predecessor/emitter (see below).

When a node crashes, its observer broadcasts the information and reconnects the ring: from now on, the observer will observe the last known predecessor (accounting for locally known failures) of its former predecessor. The rationale for using a ring for detection is to reduce the overhead in the failure free case: with only one observer, a minimal number of heartbeat messages have to be sent. We use the protocol suggested in [8] for fault detection. Consider a node q observing a node p. The observed node p is also called the emitter, because it emits periodic heartbeat messages $m_1, m_2, \ldots$ at time $\sigma_1, \sigma_2, \ldots$ to its observer q, every η time units. Now let

Algorithm 1 Sketch of the failure detector for node i.

```
1:  task Initialization
2:      emitter_i ← (i − 1) mod N
3:      observer_i ← (i + 1) mod N
4:      HB-Timeout ← η
5:      Susp-Timeout ← δ
6:      D_i ← ∅
7:  end task
8:
9:  task T1: When HB-Timeout expires
10:     HB-Timeout ← η
11:     Send heartbeat(i) to observer_i
12: end task
13:
14: task T2: upon reception of heartbeat(emitter_i)
15:     Susp-Timeout ← δ
16: end task
17:
18: task T3: When Susp-Timeout expires
19:     Susp-Timeout ← 2δ
20:     D_i ← D_i ∪ emitter_i
21:     dead ← emitter_i
22:     emitter_i ← FindEmitter(D_i)
23:     Send NewObserver(i) to emitter_i
24:     Send BcastMsg(dead, i, D_i) to Neighbors(i, D_i)
25: end task
26:
27: task T4: upon reception of NewObserver(j)
28:     observer_i ← j
29:     HB-Timeout ← 0
30: end task
31:
32: task T5: upon reception of BcastMsg(dead, s, D)
33:     D_i ← D_i ∪ {dead}
34:     Send BcastMsg(dead, s, D) to Neighbors(s, D)
35: end task
36:
37: function FindEmitter(D_i)
38:     k ← emitter_i
39:     while k ∈ D_i do
40:         k ← (k − 1) mod N
41:     return k
42: end function
```

$\sigma'_i = \sigma_i + \delta$. At any time $t \in [\sigma'_i, \sigma'_{i+1})$, q trusts p if it has received heartbeat m_i or higher. Here, δ is the time-out after which q suspects the failure of p. Assume there are initially N alive nodes numbered from 0 to $N-1$, and node $i+1 \bmod N$ observes node i according to the previous protocol, for all $0 \leq i \leq N-1$. Tasks T_1 and T_2 in Algorithm 1 execute this basic observation node, with the time-out delay being reset upon reception of a heartbeat. Note that [8] shows that this protocol, where the emitter spontaneously sends heartbeats to its observer, exhibits better performance than the variant where observers reply to heartbeat requests.

What happens when an observer (node i) suspects the crash of its predecessor in the ring? Task T_3 in Algorithm 1 implements two actions. First, it updates the local list $\mathcal{D}_i$ of

dead nodes with the identity of its emitter and then reconnects the ring (lines 19 to 23); and second, it initiates a reliable broadcast informing all nodes in its current list of alive nodes about the crash of its predecessor (line 24).

The first action, namely the reconnection of the ring, is taken care of by the procedure *FindEmitter*($\mathcal{D}_i$): node i searches its list of dead resources $\mathcal{D}_i$ and finds the first (believed) alive node, j, preceding it in the ring. It assigns j as its new emitter and sends a message NEWOBSERVER informing j that i has become its observer. Node i also sets a timeout to 2δ time units, a period after which it will suspect its new emitter, j, if it has not received any heartbeat. Task T_4 implements the corresponding action at the emitter side.

The second action for node i is the broadcast of the crash to all alive nodes (according to its current list). A message BCASTMSG($\texttt{dead}, i, \mathcal{D}_i$) containing the identity of the crashed node `dead`, the source of the broadcast i, and the locally known list of dead nodes $\mathcal{D}_i$ is broadcast to all alive nodes (according to the current knowledge of node i). We now detail how this procedure works. Let $\mathcal{A}$ be the complement of $\mathcal{D}_i$ in $\{0, 1, \ldots, N-1\}$, and let $n = |\mathcal{A}|$. The elements of $\mathcal{A}$ are labeled from 0 to $n-1$, where the source i of the broadcast is labeled 0. The broadcast is tagged with a unique identifier and involves only nodes of the labeled list $\mathcal{A}$ (this list is computable at each participant as $\mathcal{D}_i$ is part of the message). Because n is not necessarily a power of two, we have a complication[2]. Letting $k = \lfloor \log n \rfloor$ (all logarithms are in base 2), we have $2^k \leq n < 2^{k+1}$. We use twice the reliable hypercube broadcast algorithm (HBA) of [25]. The first HBA call is from the source (label 0) to the sub-cube of nodes j, where $0 \leq j \leq 2^k$, and the second HBA call is from the same source (label 0) to the sub-cube of nodes $n - j \bmod n$, where $0 \leq j \leq 2^k$. Each HBA call thus involves a complete hypercube of 2^k nodes, and their union covers all n nodes (with some overlap). The HBA algorithm delivers multiple copies of the broadcast message through disjoint paths to all the nodes in the system. Each node executes a recursive doubling algorithm and propagates the received information to up to k participants ahead of it, located at distance 2^k for $0 \leq j \leq 2^k$. For simplicity we refer to both HBA calls as a single broadcast in our algorithm.

Upon reception of a broadcast message including a source s and a list of dead nodes $\mathcal{D}$, any alive node i can reconnect the complement list $\mathcal{A}$ of nodes involved in the broadcast operation and their labels, and then compute the ordered set of neighbors *Neighbors*($s, \mathcal{D}$) to which it will then forward the message. We stress that the same list $\mathcal{D}$, or equivalently the same set of participating nodes, is used throughout the broadcast operation, even though some intermediate nodes might have a different knowledge of dead and alive nodes. This feature is essential to preserving fault-tolerance in the algorithm of [25]. Indeed, we know from [25] that each sub-hypercube broadcast is guaranteed to complete provided that there are no more than $k-1$ dead nodes within participating nodes (set $\mathcal{A}$) while the broadcast executes.

[2]Delay-bounded fault-tolerant broadcasts are not easily obtained for arbitrary values of n, see the discussion in Section VI-B.

III. MODEL & PERFORMANCE ANALYSIS

This section provides a detailed presentation of the model, and a proof of correctness of the algorithm, together with a worst-case time-performance analysis.

A. Model

1) General Framework: Nodes can communicate by sending messages in communication channels, expected to be lossless and not ordered. Any node can send a message to any other node. Messages in the communication channel (p, q) take a random time $T_{p,q}$ to be delivered, which has an upper bound τ. We consider executions where nodes can crash permanently at any time. If a node p crashes, then all communication channels to p are emptied, p does not send any message nor execute any local assignment.

Note that τ is a property of the platform, that represents the maximal time that separates a process entering a send operation, and the destination process having the corresponding message ready to read in its memory. While the exact value for τ is generally unknown, it can be bounded in our case, using the techniques described in Section V-A. The algorithm uses $\delta > \tau$ as a bound to define the limit after which a node is suspected dead. Tuning the value of δ as close as possible to τ, without underestimating τ to guarantee that false positives are not detected, is an operation that must be fitted for each target platform. Thus, in the theoretical analysis, we use τ to evaluate the worst case of a communication that succeeds, while the algorithm must rely on δ to detect a failure.

2) Using the One-Port Model: While we assume a fully-connected platform (any node may communicate with any other), we use a realistic one-port communication model [3] where a node can send and/or receive at most one message at any time-step. Independent communications, involving distinct sender/receiver pairs, can take place in parallel: however, two messages involving the same node will be serialized. Using the one-port model while aiming at a low-overhead protocol is a key motivation to this work. It is not realistic to assume that each node would observe any other node, or even a large subset of nodes: while this would greatly facilitate the diffusion of knowledge about a new crash, and speed-up the transition back to a stable configuration, it would also incur a tremendous overhead in terms of heartbeat messages, and in the end dramatically impact the throughput of the platform.

Because all messages within our algorithm have a small-size, we model our communications using a constant time τ to send a message from one node to another. We could have used a traditional model such as LogP, or used a start-up overhead plus a time proportional to the message size, but since we use this only as an upper bound, this would complicate the analysis unnecessary. Under the one-port model, the HBA algorithm [25] with 2^k nodes executes in $2k\tau$, provided that no more than $k-1$ crashes strike during its execution. The time for one complete broadcast algorithm in Algorithm 1 would then be (upper bounded by) $4\tau \log n$ in the absence of any other messages, since we use two HBA calls in sequence. But our algorithm also requires heartbeats to be sent along the ring, as well as NEWOBSERVER messages when ring reconnection

is needed. Assuming that $\eta \geq 3\tau$ (where η is the heartbeat period), we can always insert broadcast and NEWOBSERVER messages in between two successive heartbeats, thereby guaranteeing that a broadcast in Algorithm 1 will always execute within $B(n) = 8\tau \log n$, assuming no new failure interrupts the broadcast operation.

3) Stable Configuration and Stabilization Time: Here we consider executions that, from the initial configuration, reached a steady state before a failure hit the system and made it leave that steady state. To prove the correctness of our algorithm, we show that in a given time, the system returns to a steady state, assuming that no more than a bounded number of failures strike during this time.

Connected Node A node p is *connected with its successor* in a configuration, if p is alive and $\texttt{emitter}_p$ is the closest predecessor of p that is alive (on the ring). It is *connected with its predecessor* if it is alive, and $\texttt{observer}_p$ is the closest successor of p that is alive in that configuration. It is *reconnected* if it is connected with both its successor and predecessor. If all processors are reconnected, we say the ring is reconnected.

Stable Configuration A configuration C is the global state of all processes plus the status of the network. A configuration is declared as *stable*, if any alive node p is reconnected in C and for any node q, $q \in \mathcal{D}_p \iff q$ is dead in C.

Stabilization Time $T(f)$, with f being the number of overlapping failures, is the duration of the longest sequence of non stable configurations during any execution, assuming at most f failures during the sequence.

B. Correctness and Performance Analysis

The main result is the following proof of correctness, that provides a deterministic upper bound on the *Stabilization Time* $T(f)$ of the algorithm with at most f overlapping faults:

Theorem 1. *With $n \leq N$ alive nodes, and for any $f \leq \lfloor \log n \rfloor - 1$, we have*

$$T(f) \leq f(f+1)\delta + f\tau + \frac{f(f+1)}{2} B(n) \qquad (1)$$

where $B(n) = 8\tau \log n$.

This upper bound is pessimistic for many reasons, which are discussed after the proof. But the key point is that the algorithm tolerates up to $\lfloor \log n \rfloor - 1$ overlapping failures in logarithmic time $O((\log n)^3)$.

Proof. Starting from a non stable configuration, the next stable configuration will be reached when (i) all nodes are informed of the different failures via the broadcast, and (ii) processes of the ring are reconnected. Recall that every time a node has detected a failure, it initiates a broadcast that executes within $B = B(n) = 8\tau \log n$ time units, and which is guaranteed to reach all alive nodes as long as $f \leq \lfloor \log n \rfloor - 1$. Because we interleave reconnection messages within the broadcast, B encompasses both the broadcast and the reconnection. However, due to the one-port model, we cannot assume anything about the pipelining of several consecutive broadcast operations. In this proof, we make a first simplification by over-approximating $T(f)$ as the maximum time $R(f)$ to reconnect the ring after f overlapping failures, plus the time to execute all the broadcasts that were initiated, in sequence (assuming no overlap at all). We prove an upper bound on $R(f)$ by induction, letting $R(0) = 0$:

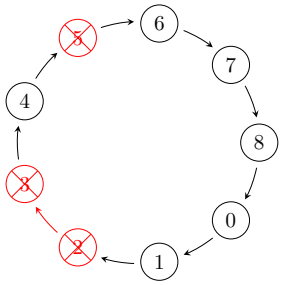

Fig. 1: Segments of dead nodes after $f = 3$ failures: $n = 9$, $k = 2$, $I_1 = \{2, 3\}$, $I_2 = \{5\}$, $d_1 = 2$ and $d_2 = 1$.

Lemma 1. *For $1 \leq f \leq \lfloor \log n \rfloor - 1$, we have*

$$R(f) \leq R(f-1) + 2f\delta + \tau \qquad (2)$$

Proof. We first prove Equation (2) when $f = 1$. Assume that node p, observed by node q, fails. After receiving the last heartbeat, q needs δ time units to detect the failure (line 15 of Algorithm 1). Thus, the worst possible scenario is when p fails right after sending a heartbeat, which will take τ time units to reach q. Thus q detects the failure after $\tau + \delta$ time units. Finally, q sends the reconnection message to the predecessor of p, which will take τ, hence $R(1) \leq 2\tau + \delta$. We keep the over-approximation $R(1) \leq \tau + 2\delta$ to simplify the formula in the general case.

Assume now that Equation (2) holds for all $f \leq \lfloor \log n \rfloor - 2$. Now consider an execution with $f+1$ overlapping failures, the first of them striking at time 0 (see Figure 2). The $(f+1)$-th failure strikes at time t. Necessarily $t \leq R(f)$, otherwise the ring would have been reconnected after f failures, and the last one would not be overlapping. There are f dead nodes just before time t among the original n alive nodes, which define $k \leq f$ segments I_i, $1 \leq i \leq k$. Here, segment I_i is an interval of $d_i \geq 1$ consecutive dead nodes (see Figure 1). Of course $\sum_{i=1}^{k} d_i = f$, and there remain $n - f$ alive nodes. There are multiple cases depending upon which node is struck by the $(f+1)$-th failure at time t:

(a) The new failure strikes a node that is neither a predecessor nor a successor of a segment (e.g., the failure strikes node 7 in Figure 1). In that case, a new segment of length 1 is created, and the ring is reconnected at time $t + R(1)$.

(b) The new failure strikes a node p that precedes a segment I_i. Let q be the successor of the last dead node in I_i. By definition, $q \neq p$. There are two sub-cases: (i) The predecessor p' of p is still alive (e.g., the failure strikes node 1 preceding segment I_1 in Figure 1, $q = 4$ and $p' = 0$ is alive). Then the size of segment I_i is increased by one. In the worst case, q is not aware of the death of any node in I_i at time t, and needs to probe all these nodes one after the other before reconnecting with p' (in the example, $q = 4$ needs to try to reconnect with

2 and 1 since it is not aware of their death). This costs at most $(d_i+1)(2\delta)+\tau \leq 2(f+1)\delta+\tau$, because $d_i+1 \leq f+1$, hence the ring is reconnected at time $t+2(f+1)\delta+\tau$. (ii) The predecessor p' of p is dead (e.g., the failure strikes node 4 preceding segment I_2 in Figure 1, $q=6$ and $p'=3$ is dead). Then p' is the last node of another segment I_j. In that case, segments I_i and I_j are merged into a new segment of size $d_i+d_j+1 \leq f+1$. Just as before, in the worst case, q is not aware of the death of any node in that new segment, and the reconnection costs at most $(d_i+d_j+1)(2\delta)+\tau \leq 2(f+1)\delta+\tau$ (see Figure 2 for an illustration). Hence the ring is reconnected at time $t+2(f+1)\delta+\tau$.

(c) The new failure strikes a node p that follows a segment I_i. Let q be the successor of p. If q is alive, it now follows a segment of size d_i+1. If q is the first dead node of segment I_j, let r be the node that follows I_j. Now r follows a segment of size d_i+d_j+1. In both cases, we conclude just as before. This completes the proof of Lemma 1. □

From Lemma 1, we easily derive by induction that

$$R(f) \leq f(f+1)\delta + f\tau$$

for all values of $f \leq \lfloor \log n \rfloor - 1$. During the ring reconnection, processes that discover a dead process initiate a broadcast of that information. We need to count, in the worst case, how many broadcasts are initiated to compute how long it takes for the information to be delivered to all nodes.

Lemma 2. *Let $p_i, 1 \leq i \leq f \leq \lfloor \log n \rfloor - 1$ be the $i-th$ process subject of a failure. In the worst case, at most $f-i+1$ processes can detect the death of p_i.*

Proof. A process p is discovered dead by process q in Task T3, if $\texttt{emitter}_q = p$. In that case, p is added to $\mathcal{D}_q$, and $\texttt{emitter}_q$ is re-computed using *FindEmitter*. That function cannot return any process in $\mathcal{D}_q$, and p is never removed from $\mathcal{D}_q$. Thus, q will never discover the death of p again. As long as q lives, no other process q' will execute the task T3 with $\texttt{emitter}_{q'} = p$, because q is an alive process between q' and p in the ring. Thus, q must fail after p, for p to be discovered once more. Since there are at most f faults, p_i, the $i-th$ failed process can thus be discovered dead by at most $f-i+1$ processes. □

We derive from Lemma 2 that at most $\sum_{i=1}^{f}(f-i+1) = \frac{f(f+1)}{2}$ broadcasts are initiated. Finally, the information on the f dead nodes must reach all alive nodes. For each segment I_i, there is a last failure after which the broadcast initiated by the observing process is not interrupted by new failures. That broadcast operation thus succeeds in delivering the list of newly discovered dead processes to all others ($d_i \leq \lfloor \log n \rfloor - 1$). In the worst case, that broadcast operation is the last to complete. As already mentioned, we conservatively consider that all the broadcast operations execute in sequence. Since there are at most $\frac{f(f+1)}{2}$ broadcast operations initiated , we obtain $T(f) \leq R(f) + \frac{f(f+1)}{2}B(n)$, which leads to the upper bound in Equation (1) and concludes the proof of Theorem 1. □

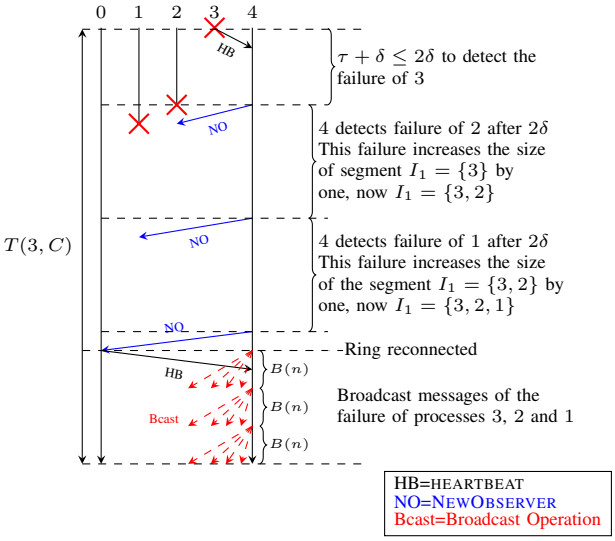

Fig. 2: From stable configuration C, growing segment I_1 of Figure 1: first failure on node 3, next two failures striking its ring predecessors.

The bound on $T(f)$ given by Equation (1) is quite pessimistic. We can identify three levels of complexity with their corresponding bounds on $T(f)$. In the most likely scenario, where the time between two consecutive faults is larger than $T(1)$, the system has time to return to a stable configuration before the second fault, in which case all faults can be considered as independent, and the average stabilization time is $T(1) = R(1) + B(n) = O(\log n)$. If the system suffers quickly overlapping faults, the location of impacted nodes becomes important. However, the larger the platform, the smaller the probability that successive faults strike consecutive nodes ($2/n$, where n is the number of alive nodes). Thus, on large platforms, overlapping failures are more likely to strike non consecutive nodes in the ring. If overlapping faults hit non consecutive nodes rapidly, i.e., faster than the time needed by the system to reach the next stable configuration, each error is detected once, but due to the one-port model, the upper bound on $T(f)$ becomes $R(1)+fB(n) = O(\log^2 n)$. Finally, in the unlikely scenario where f quickly overlapping faults hit f consecutive nodes in the ring, the Theorem 1 provides the upper bound for $T(f) \leq R(f) + \frac{f(f+1)}{2}B(n) = O(log^3 n)$.

C. Non Stabilization Risk Control

To guarantee convergence within $T(f)$ time units, Algorithm 1 assumes that $f \leq \lfloor \log(n) \rfloor - 1$. In order to evaluate the risk behind this assumption, consider that failures strike following an Exponential distribution of parameter λ. Let $P_T(f)$ be the probability of the event *"more than f failures strike within time T"*. Then $P_T(f) = 1 - \Sigma_{k=0}^{f} \frac{(\lambda T)^k}{k!} e^{-\lambda T}$.

Consider a platform of n nodes: if μ_{ind} is the MTBF of a single node, then $\lambda = \frac{n}{\mu_{\text{ind}}}$ [15]. Let $M = \lfloor \log(n) \rfloor - 1$, the assumption that there will not be more than M failures before stabilization is then true with probability $P_{T(M)}(M)$. In Figure 3, we represent this relation by showing the upper bound of δ to enforce $P_{T(M)}(M) < 10^{-9}$, at variable machines scale (n), and for different values of μ_{ind}, with a message time bound of $\tau = 1\mu s$. Figure 3 illustrates that for all values of δ lower than the bound shown for a given system size and individual

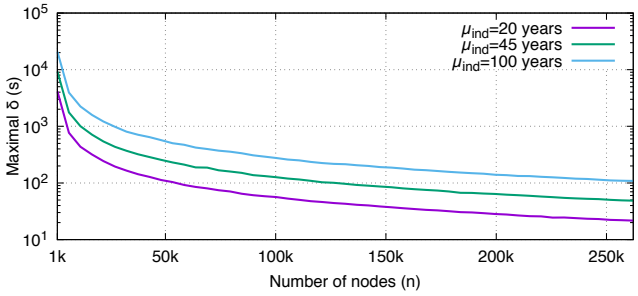

Fig. 3: Maximal value for δ to ensure that $P_{T(M)}(M) < 10^{-9}$ with $\tau = 1\mu s$ and $M = \lfloor \log_2(n) \rfloor$.

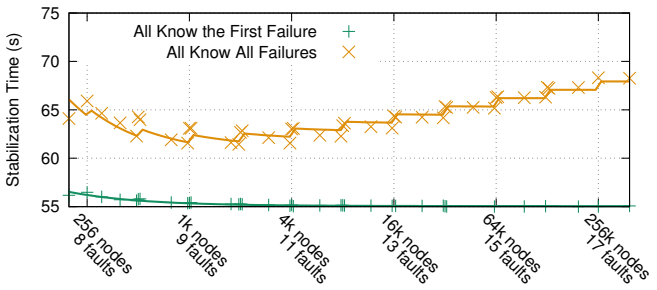

Fig. 4: Average Stabilization Time, when the maximal number of failures strike a platform of varying size in the scenario LOWNOISE ($\delta = 1min, \tau = 1\mu s, \eta = 10s$).

node reliability, the probability that failures strike fast enough to prevent Algorithm 1 from converging in $T(f)$ is negligible (less than 0.000000001). As already mentioned, this bound on δ is a loose upper bound, because the bound on $T(f)$ in Equation (1) is loose itself. Furthermore, it captures the risk that enough failures would strike during stabilization time to make the appearance of the worst case scenario *possible*, even though this worst case scenario has itself a very low probability to happen (as shown in Sections IV and V). Still, for the largest platforms with $n = 256,000$ nodes, we find that $\delta \leq 22s$ for the most pessimistic $\mu_{\text{ind}} = 20$ years, and $\delta \leq 60s$ if $\mu_{\text{ind}} = 45$ years results in timely convergence. With such large values, the detector generates negligible noise to the applications, as shown in Section V-C.

IV. SIMULATIONS

We conduct simulations and experiments to evaluate the performance of the algorithm under different execution scenarios and parameter settings. We instantiate the model parameters with realistic values taken from the literature. The code for all algorithms and simulations is publicly available [26], so that interested readers can build relevant scenarios of their choice. In this section, we report simulation results. See Section V for experiments.

A. Simulation Settings

The discrete-event simulator imitates how the protocol of Algorithm 1 would behave on a distributed machine of size n. Messages between a pair of alive nodes in this machine take a uniformly distributed time in the interval $(0, \tau]$. Failures are injected following an exponential law of parameter $\lambda = n/\mu_{\text{ind}}$ (see Section III-C). In order to generate a manageable amount of events, each heartbeat message and the corresponding timeouts are not simulated, but the simulator asserts that a timeout should have expired on the observer after the death of its emitter, if the observer is alive at that time (otherwise, the observer's observer is going to react, following the protocol).

The simulator computes (i) the average time to reach a stable configuration (all processes know all faults) starting from a configuration with a single failure injected at time 0, (ii) the average time to reach a configuration where all processes know about the initial failure, and (iii) the average number of failures striking during the time it takes to reach a stable configuration, over a set of 10,000 independent runs.

We consider two main scenarios for the simulations. In both scenarios, we target a large scale machine (up to 256,000 computing nodes) with a low latency interconnect ($\tau = 1\mu s$). In the scenario LOWNOISE, we set the failure detector so as to minimize the overhead in the failure free case: η is set to 10 seconds, and δ to 1 minute. We consider this case significant for platforms where nodes are expected to be reliable, or where alternative methods to detect most failures exist; the heartbeat mechanism is then used as a last resort solution, e.g. when special hardware providing a Baseboard Management Controller and controlled through a protocol like IPMI [30] is connected to the application notification system. We also considered a scenario LOWLAT, with the opposite assumptions, where active check through heartbeats is the primary method to detect failures, and a low latency of detection is required for the application: $\eta = 0.1s$, and $\delta = 1s$.

B. Simulation Results

In Figure 4, we force the simulator to inject the maximum number of failures tolerated by the algorithm for a given platform size ($\lfloor \log_2(n) \rfloor - 1$) in a very short time, inferior to δ, in order to evaluate the average stabilization time in the most volatile environment. Varying the system size (n), and the number of injected failures simultaneously, we evaluate the time taken for the first failure to be notified to all processes, and for all the processes to be notified of all the failures that struck since the last stable configuration.

The figure considers scenario LOWNOISE. Points on the graph show times reported by the simulator, while lines represent functions fitted to these points, $O(\frac{1}{\eta} + \lfloor \log_2(n) \rfloor)$ for *all know all failures* (orange lines), and $O(\frac{1}{n})$ for *all know the first failure* (green lines).

In average, the first failure, striking at time 0, is detected $\delta - \frac{\eta}{2}$ seconds later, and this is the observed base line for detecting the first failure at all nodes. The reliable broadcast overhead in this case is negligible, because $\tau << \delta$ and η. There are a few executions in which, within the first δ seconds, another failure hits the observer of the first failure, introducing another δ delay to actually detect the first failure and broadcast it. As the size of the machine increases, this probability decreases. Such overlapping failure cases contribute to a longer detection and notification time that can be fitted with a function inversely proportional to the platform size, but have a low probability to happen, introducing a measurable but small

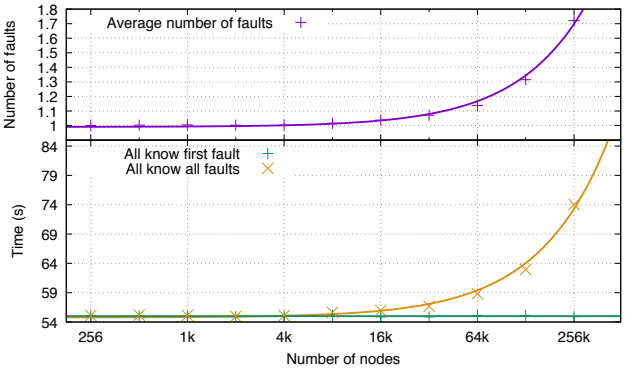

Fig. 5: Average Stabilization Time, with random overlapping failures in scenario LOWNOISE ($\delta = 1min, \tau = 1\mu s, \eta = 10s$), with $\mu_{\text{ind}} = 1year$.

overhead at small scale. For general stabilization, where all processes need to know all failures, the reliable broadcast remains as fast as for the initial failure. However, if any failure strikes before that broadcast phase is complete, this delays reaching stabilization by another δ followed by a logarithmic phase. As we observe in both figures, this shows at large scale, where failures have a high probability to strike successively, each introducing a constant overhead. The fitting function thus shows the same *inversely proportional* property in the beginning, then the logarithmic behavior starts to dominate at large scale.

We conducted the same set of simulations on the LOWLAT scenario, but cannot include them for lack of space. The evaluation presents the exact same characteristics, shifted by the ratio between the two values for δ.

We then consider the average case, when failures are not forced to strike quasi-simultaneously. We set the MTBF of independent components to a very pessimistic value ($\mu_{\text{ind}} = 1year$), making the MTBF of the platform decrease to a couple of minutes at 256,000 nodes. Although we do not expect such a pessimistic value in real platforms, we evaluate this case in order to ensure that failures may occur before the initial one is detected and broadcast (or stabilization would be reached immediately after). Figure 5 presents the average number of failures observed at different scales, the average time for all nodes to know about the first failure, and the average time for all nodes to know about all failures. Points represent values given by the simulator, while lines represent fitting functions: $O(1)$ for the time for all to know the first failure, $O(n)$ for the average number of failures and the average time for all to know all failures. We present here the scenario LOWNOISE, although the result also holds for scenario LOWLAT, at a different scale.

This figure shows that, on average, and even with extremely low MTBFs, the probability that two independent failures hit the system in an overlapping manner (before the first failure is known by all nodes) is very low. This happens when the MTBF of the system becomes comparable to δ. In that case, the first failure still takes close to a constant time to be notified to all. This is because $\tau \log_2(n)$ remains very small compared to δ, and once the broadcast is initiated, it completes in $\tau \log_2(n)$. The successive failures may strike anytime between $[0, \delta]$, delaying the time to reach the stable configuration by another $\delta + \tau \log_2(n)$. On average, at 256,000 nodes, this happens in the middle of the initial failure detection interval, delaying the completion by $\delta/2$. Each failure, however, is independent in that case, and each is detected almost δ time units after it strikes.

V. EXPERIMENTAL EVALUATION

This section presents an experimental evaluation of an operational implementation of the proposed failure detector on the Titan ORNL supercomputer. We have implemented the failure detection and propagation service in the reference implementation of the User-Level Failure Mitigation (ULFM) draft MPI standard [5], provided by OPEN MPI. ULFM is an extension of the MPI standard that empowers MPI users (that is, applications, library developers, or parallel programming languages) to provide their own fault tolerant strategy. The general design of ULFM relies on local semantics: failures are notified to the user only in MPI calls that involve a failed process, and a correct ULFM implementation will try to make all operations succeed, if it can complete locally. Although this relaxed design eases the implementation requirements and delivers higher failure-free performance, the fact that a failure is guaranteed to be detected only after an active reception from the dead process can lead to an increase of latency during failure recovery operations, because the same process failures may be detected sequentially by multiple processes, possibly at a much later time than when they were first reported. Moreover, several routines imply necessarily a communicator-wide knowledge on failures: operations like `MPI_COMM_AGREE` and `MPI_COMM_SHRINK` need to build consistent knowledge on (sub)sets of acknowledged failures; a pending point-to-point reception from any source must eventually raise an error, if it cannot complete because of the death of a processor. Therefore, the addition of the failure detection and propagation service provides an acceleration to such scenarios, by eliminating delayed local observation of the failure, which can then be immediately reported to the upper-level, which can then act upon it quickly.

A. Implementation

The failure detector is composed of two components: the observation ring, and the propagation overlay. The components operate on a group of processes, which must be MPI consistent (that is, identical at all ranks). The propagation topology is implemented at the Byte Transport Layer (BTL) level, which provides the portable low-level transport abstraction in OPEN MPI.

The propagation overlay takes advantage of the Active Message behavior of the OPEN MPI BTL's. Each message, with a size lesser than the "eager" protocol switch point, contains the index of the callback function to be analyzed by upon reception. This approach provides independence from the MPI semantic (including matching). Upon the reception of a propagation message, the message is forwarded according to two possible algorithms. In the case where the overlay is not corrected to incorporate the knowledge about failed processes, thus the group can be considered as an invariant

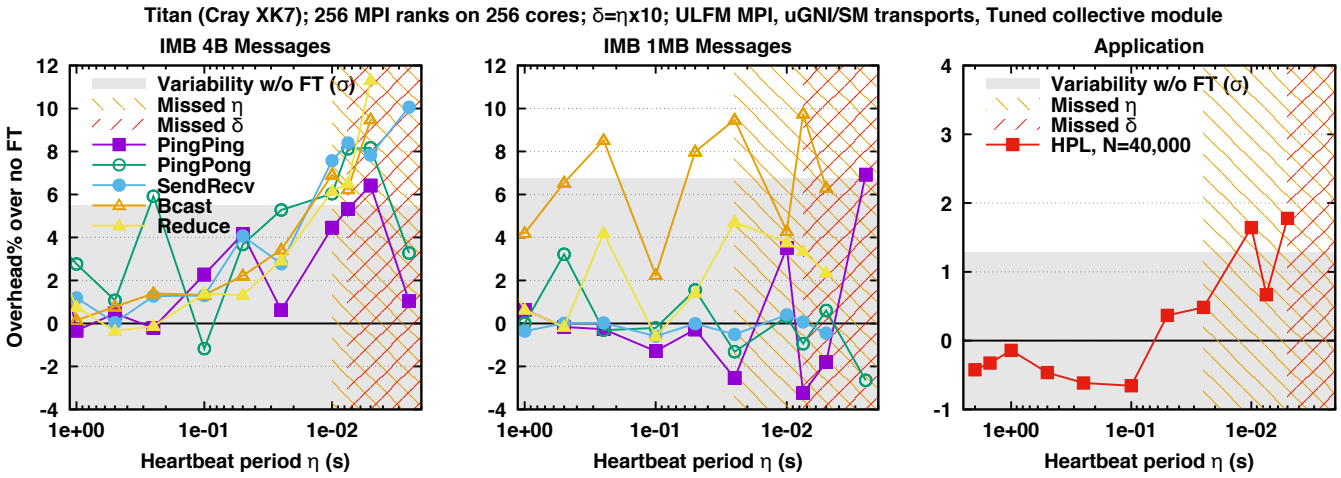

Fig. 6: Sensitivity to noise resulting from the failure detector activity for varied workloads.

during the entire execution, the message is forwarded as is through the propagation topology which is constructed every time a broadcast is initiated, according to the algorithm presented in Section II, in order to guarantee the logarithmic propagation delay. When the upper level declares, through a runtime parameter, that it repairs its communicators after every stabilization phase, the reliable propagation overlay can reduce the size of the messages to include only the latest detected failures, and the overlay is then built considering all processes of the group.

The observation ring is also built at the BTL level. The emission of the heartbeats poses a particular challenge in practice. The timely activation and delivery of heartbeats is of critical importance to enforce the perfection of the detector, and the bound on τ. Missing its η emission period deadlines puts the emitter process at risk of becoming suspected by its observer, even though it is still alive. If the heartbeats are emitted from the application context, they can only be sent when the application enters MPI routines, and consequently, a compute intensive MPI application would often miss the η period. In our implementation, the heartbeats are emitted from within a separate, library internal thread, in order to render their emission independent from the application's communication pattern. For ease of implementation, the `MPI_THREAD_MULTIPLE` support is enabled by default when the detector thread is enabled; however, future software releases will drop this requirement. An intricate issue also arises from a negative interaction between the emission and the reception of heartbeat messages. In order to check the liveliness of the emitter process (after the δ timeout), the observer has to check if it has received heartbeats. From an implementation perspective, if the heartbeats are sent through the "eager" channel, the detector thread, which is the receive thread in this case, has to be active and poll the BTL engine for progress. However, if the application has posted operations on large messages, the poll operation may start progressing these (long) operations before returning control to the detector thread, leading to an unsafe delay in the emission of heartbeats from that same thread. To circumvent that difficulty, the detector thread emits heartbeats using the "RDMA put" channel. Heartbeats are thus directly deposited by raising a flag in the registered memory at the receiver, using hardware accelerated put operations that do not require active polling. The observer can then simply check that the flag has been raised during the last δ period with a local load operation, and reset the flag with a local store, which are mostly impervious to noise and do not delay the η period. This approach also allows the observer to miss δ periods without endangering the correctness of the protocol (only increasing the time to detect and notify the failure, but no triggering a false positive).

B. Experimental Conditions

The experiments are carried out on the Titan ORNL Supercomputer [27], a Cray XK7 machine with 16-core AMD Opteron processors and the Cray Gemini interconnect. The ULFM MPI implementation is based on a pre-release of Open MPI 2.x (`r#6e6bbfd`), which supports the optimized uGNI and shared-memory transports (without XPmem), and uses the Tuned collective module. The MPI implementation is compiled with the `MPI_THREAD_MULTIPLE` support. Every experiment is repeated 30 times and we present the average. The benchmarks are deployed with one MPI rank per core, and all threads of an MPI process are bound to that same core (application, detector, and driver threads when applicable, i.e., the detector thread does **not** require exclusive compute resources).

C. Noise and Accuracy

The first set of experiments investigate the noise generated by the detector and its accuracy for different workloads when η and δ vary, in a method similar to [20] that focused exclusively on measuring the noise generated by different failure detection strategies. The η and δ periods are set so that $\delta = 10 \times \eta$. If the test is successful (that is, no failure was detected, since none was injected in this experiment), then η is reduced, and the experiment is repeated, until a false positive is reported. We also collect the number of times an η deadline was missed, even when the δ timeout is still respected. We first considered a non-communicative, compute-only MPI application where each rank calls LAPACK `DGEMM` operations on local matrices, without calling MPI routines for extended periods of time.

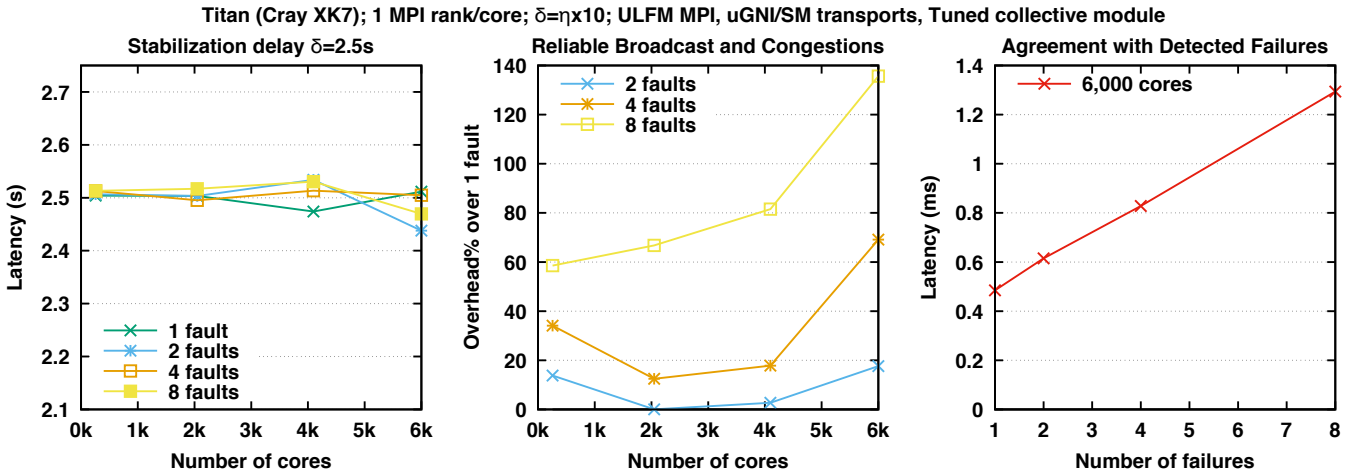

Fig. 7: Detection and propagation delay, and impact on completion time of fault-tolerant agreement operation.

Without the detector thread, the non-communicative benchmark reports false detections for all considered values of η. With the detector thread, this non-communicative benchmark succeeds until η is set to one millisecond. However, starting from $\eta < 5$ milliseconds, messages indicating a missed η deadline are occasionally issued (although the δ timeout is still respected). These observations are consistent with the scheduling time quantums (`sched_min_granularity` is set to 3ms), and indicate that the thread scheduling latency is an absolute for the minimum η period. Smaller periods could be achieved with a real time scheduler, but such capabilities require administrative privileges, which is undesirable.

Next, in Figure 6 we present the noise incurred on a variety of communication, and computation workloads, provided by the Intel MPI Benchmark (version 4.1), and HPL (version 2.2), respectively. Accuracy results are similar overall in the communicative benchmarks. All tests of the IMB-MPI1 suite can run without false detection for $\eta \geq 10ms$. Notably, point-to-point only benchmarks can succeed with η value as low as $2.5ms$ but occasionally report false suspicions. Collective communication benchmarks are more sensitive and report occasional heartbeat emission deadline misses until $\eta \geq 25ms$, due to contentions on the access to hardware network resources.

The latency performance (left graph) and bandwidth performance (center graph) are barely affected by low frequencies of heartbeat emissions. For higher frequencies, the overhead generated by the noise can reach approximately 10%. The bandwidth performance is less impacted overall than the latency, especially for point-to-point bandwidth, which remains unchanged for all but the most extreme values of η. The application performance (Linpack, right graph) exhibits no observable performance degradation for $\eta \geq 100ms$. For higher frequencies, the performance degradation remains contained under 2%.

D. Failure Detection Time

Figure 7 presents the behavior observed when injecting failures. The first graph (left) presents the time to reach a stable state when injecting 1 to 8 failures for a varying number of nodes. After synchronizing, the desired number of MPI processes (whose ranks are chosen at random) simulate a failure. All other processes post an *any-source* reception. When the reception raises a process failure exception (the only possible outcome for this non-matched any-source reception), the process counts the number of locally known failed processes, and if it does not contain all injected failures, repeats the reception. The time at which all failures have been locally observed is reported at each rank. We observe that for small scales, the reported delay is consistently close to δ. If emitters were sending heartbeats to their observer at random starting time, we would expect the detection time to be closer to $\delta - \eta/2$; however, as all processes start to sending heartbeats to their observer at the end of the `MPI_Init` function, they are almost synchronized, and for all runs we observe a consistent delay at small scale. At larger scale, processes leave `MPI_Init` at a more variable date, and the average starts to converge toward the theoretical bound. This observation matches the model, considering that in this scenario all failures are "simultaneous", and that the random allocation of failures has a low probability of hurting observer/emitter pairs. Consequently, the detection and propagation of each of these failures progresses concurrently and do not suffer from the cumulative effect of detecting multiple predecessors' failures on the ring.

The second experiment (center in Figure 7) investigates the effect of collisions on the reliable broadcast propagation delay. The benchmark is similar to the previous experiment, except that before a process simulates a failure, it sends its observer a special "trigger heartbeat", which initiates an immediate propagation reporting it dead, without waiting for the δ timeout. The rest of the observation protocol remains unchanged (i.e., heartbeats are exchanged between live processes with an η period, and the observer of the injection process switches to observing the predecessor). We then present the increase in the average duration of the reliable broadcast when multiple broadcasts are progressing concurrently. To simplify the proof of the upper bound on stabilization time (Theorem 1), we have considered that successive broadcasts are totally sequential. This is an admittedly pessimistic hypothesis, and indeed, performing two concurrent propagations does not significantly increase the delay, as the two reliable broadcasts can actually

overlap almost completely. However, starting from 4, and, more prominently, for 8 concurrent broadcasts, the average completion time is significantly increased. Considering the small size of the messages, the bandwidth requirements are small, and contention on port access is indeed the major cause of the imperfect overlap between these concurrent broadcasts, therefore vindicating the importance of considering a port-limited model during the design of the failure detector and propagation algorithms.

The last experiment (right in Figure 7) presents the performance of the agreement algorithm after failures have been injected. The authors of [14] presented a similar performance result for their agreement algorithm. In their results, the agreement performance was severely impacted when failure were discovered during the agreement (with the failure free performance of 80μs increasing to approximatively 80ms), an effect the authors claim is due to failure detection overhead. In their work, failure detection was delegated to an ORTE based RAS service, responsible for detecting and propagating failures. In this experiment, we strive to recreate as closely as possible this setup, except that we deploy our failure detector in lieu of the ORTE RAS service. We consider the same implementation of the agreement, on 6,000 Titan cores (the same number of cores they deployed on the generally similar Cray XC30 Darter system). Some in-band detection capabilities are active, in particular, failure of shared-memory sibling ranks are reported by the node's local operating system. With the replacement of the ORTE RAS service by our failure detector algorithm, the time to completion of the agreement algorithm decreases to below 1.5ms (a 50x improvement). This is due to the faster propagation of failure knowledge among the agreement participants: instead of waiting for (long) in-band timeouts or ORTE RAS notification, a process whose parent or children have failed can observe the condition much earlier, and start the on-line mending of the fan-in/fan-out tree topology at an earlier date. Interestingly, previously hidden performance issues become visible, as failure detection is not the dominant cost anymore: we observe that the performance of the agreement decreases linearly with the number of detected failures, a behavior that can be attributed to the agreement algorithm performing a linear scanning of the group when a failure is reported.

VI. Related work

In this section, we survey related work on failure detectors and then on fault-tolerant broadcast algorithms.

A. Failure detectors

A number of failure detection (FD) algorithms have been proposed in the literature. Most current implementations of FDs are based on an all-to-all communication approach where each node periodically sends heartbeat messages to all nodes. Because they consider a fully connected set of known nodes that communicate in an all-to-all manner, these implementations are not appropriate for platforms equipped with a large number of nodes. Several efforts have been made towards scaling up failure detectors implementations [2], [22]. An alternative approach for implementing scalable failure detectors is to use gossip-like protocols where nodes randomly choose a few other nodes with whom they exchange their failure information [29], [12], [13], [18], [28]. Targeting HPC computations at scale, a scalable failure detector is propsoed in [19], based on observing random nodes and gossiping information. In their protocol, each ping message transmits information on all currently known failures, either via a liveness matrix or in compressed form.

Practically, gossip approaches bring along redundant failure information which degrades their scalability. Furthermore, the randomization used by gossip protocols makes the definition of timeout values difficult, since the monitoring sets change often over time. In order to eventually avoid false detections, these techniques tend to oversize their timeouts, which results in longer detection times. Theoretically, gossip approaches introduce random detection and propagation times, whose worst-case with a prescribed risk factor are hard to bound[3]. In contrast, our algorithm follows a deterministic detection and propagation topology with (i) constant-size heartbeats and well-defined delays, (ii) a single observer, (iii) a logarithmic-time propagation, and (iv) a guaranteed worst-time to stabilization, thereby achieving all the goals of randomized methods with a deterministic implementation.

B. Fault-Tolerant Broadcast

Fault-tolerant broadcasting algorithms have been extensively studied, and we refer the reader to the surveys in [24], [16]. A key-concept is the fault-tolerant diameter of the interconnection graph, which is defined as the maximum length of the longest path in the graph when a given number of (arbitrarily chosen) nodes have failed [21]. The main objective in this context is to identify classes of overlay networks whose fault-tolerant diameter is close to their initial (fault-free) diameter, even when allowing a number of failures close to their minimal degree (allowing more failures than the minimal degree could disconnect the graph). Furthermore, these overlay networks should provide enough vertex-disjoint paths for broadcast algorithms to resist that many failures.

Research has concentrated on regular graphs (where all vertices have the same degree): hypercubes [21], [25], [11], binomial graphs [1] or circulant networks [23]. For all these graphs, efficient broadcast algorithms have been proposed. These algorithms tolerate a number of failures up to their degree minus one, and execute within a number of steps (in the one-port model) that does not exceed twice their original diameter. However, to the best of our knowledge, such algorithms require the number of nodes in the graph to be a power of two, or a constant times a power of two, while we need an algorithm for an arbitrary number of nodes. This motivates our solution based upon a double diffusion (see Section II).

[3]Absolute worst-case times are infinite, as some nodes could be observed only after an unbounded delay. To give a simple example, after an observation round with n nodes randomly selecting their targets, in expectation, n/e nodes will not be observed (where $e = 2.718$ is Euler's number).

VII. Conclusion

Failure detection is a critical service for resilience. The failure detector presented in this work relies on heartbeats, timeouts, and communication bounds to provide a reliable solution that works at scale, independently of the type of faults that create permanent node failures. Our study reveals a complicated tradeoff between system noise, detection time, and risks: a low detection time would demand a low latency in the detection of failures, thus a tight approximation of the communication bound, increasing the risk of a false positive, and a frequent emission of heartbeat messages, increasing the system noise generated by the failure detector. We proposed a scalable algorithm capable of tolerating high frequency failures, and proved a theoretical upper bound to the time required to reconfigure the system in a state that allows new failures to strike; therefore the algorithm can tolerate an arbitrary number of failures, provided that they do not strike with higher frequency. The algorithm was implemented in a resilient MPI distribution, which we used to assess its performance and impact on applications at large scale. The performance evaluation shows that for reasonable values of detection time, the ring strategy for detection introduces a negligible or non-measurable amount of additional noise in the system, while the high performance reliable broadcast strategy for notification allows for quickly disseminating the fault information, once detected by the observing process.

Implementation considerations lead us to advocate that the detection part of the service should be provided at a lower levels of the software stack, either inside the operating system, or inside the interconnect hardware: active heartbeats to probe the activity of remote nodes could be handled by these lower levels without measurable noise, and with tighter bounds, since the other levels of the software stack would not introduce additional components to the noise. Future work should focus on providing this capability, and on evaluating the approach to address the tradeoff between detection time and risk.

References

[1] T. Angskun, G. Bosilca, and J. Dongarra. Binomial graph: A scalable and fault-tolerant logical network topology. In I. e. a. Stojmenovic, editor, *Parallel and Distributed Processing and Applications ISPA*, pages 471–482. Springer, 2007.

[2] M. Bertier, O. Marin, and P. Sens. Performance analysis of a hierarchical failure detector. In *International Conference on Dependable Systems and Networks*, pages 635–644, San Francisco, CA, 2003.

[3] P. B. Bhat, C. S. Raghavendra, and V. K. Prasanna. Efficient collective communication in distributed heterogeneous systems. *J. Parallel and Distributed Computing*, 63(3):251–263, 2003.

[4] W. Bland, A. Bouteiller, T. Herault, G. Bosilca, and J. Dongarra. Post-failure recovery of MPI communication capability: Design and rationale. *International Journal of High Performance Computing Applications*, 27(3):244–254, 2013.

[5] W. Bland, A. Bouteiller, T. Herault, J. Hursey, G. Bosilca, and J. J. Dongarra. An evaluation of user-level failure mitigation support in MPI. *Computing*, 95(12):1171–1184, 2013.

[6] W. Bland, H. Lu, S. Seo, and P. Balaji. Lessons Learned Implementing User-Level Failure Mitigation in MPICH. In *Proc. CCGrid*. IEEE, 2015.

[7] T. D. Chandra and S. Toueg. Unreliable failure detectors for reliable distributed systems. *J. ACM*, 43(2):225–267, 1996.

[8] W. Chen, S. Toueg, and M. K. Aguilera. On the quality of service of failure detectors. *IEEE Trans. Comput.*, 51(5):561–580, 2002.

[9] I. P. Egwutuoha, D. Levy, B. Selic, and S. Chen. A survey of fault tolerance mechanisms and checkpoint/restart implementations for high performance computing systems. *The Journal of Supercomputing*, 65(3):1302–1326, 2013.

[10] K. B. Ferreira, P. Bridges, and R. Brightwell. Characterizing application sensitivity to OS interference using kernel-level noise injection. In *Proc. SC '08*. IEEE Computer Society Press, 2008.

[11] P. Fraigniaud. Asymptotically optimal broadcasting and gossiping in faulty hypercube multicomputers. *IEEE Trans. Computers*, 41(11):1410–1419, 1992.

[12] I. Gupta, T. D. Chandra, and G. S. Goldszmidt. On scalable and efficient distributed failure detectors. In *Proceedings of the Twentieth Annual ACM Symposium on Principles of Distributed Computing*, PODC '01, pages 170–179, New York, NY, USA, 2001. ACM.

[13] N. Hayashibara, A. Cherif, and T. Katayama. Failure detectors for large-scale distributed systems. In *21st Symposium on Reliable Distributed Systems (SRDS 2002), 13-16 October 2002, Osaka, Japan*, pages 404–409, 2002.

[14] T. Herault, A. Bouteiller, G. Bosilca, M. Gamell, K. Teranishi, M. Parashar, and J. Dongarra. Practical scalable consensus for pseudo-synchronous distributed systems. In *Proc. SC'15*. IEEE Computer Society Press, 2015.

[15] T. Hérault and Y. Robert, editors. *Fault-Tolerance Techniques for High-Performance Computing*, Computer Communications and Networks. Springer Verlag, 2015.

[16] M.-C. Heydemann. Cayley graphs and interconnection networks. In G. Hahn and G. Sabidussi, editors, *Graph Symmetry: Algebraic Methods and Applications*, pages 167–224. Springer, 1997.

[17] T. Hoefler, T. Schneider, and A. Lumsdaine. Characterizing the influence of system noise on large-scale applications by simulation. In *Proc. SC '10*. IEEE Computer Society Press, 2010.

[18] Y. Horita, K. Taura, and T. Chikayama. A scalable and efficient self-organizing failure detector for grid applications. In *Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing*, GRID '05, pages 202–210, Washington, DC, USA, 2005. IEEE Computer Society.

[19] A. Katti, G. Di Fatta, T. Naughton, and C. Engelmann. Scalable and fault tolerant failure detection and consensus. In *Proc. EuroMPI '15*. ACM, 2015.

[20] K. Kharbas, D. Kim, T. Hoefler, and F. Mueller. Assessing HPC Failure Detectors for MPI Jobs. In *Proc. PDP '12*. IEEE Computer Society, 2012.

[21] M. Krishnamoorthy and B. Krishnamurthy. Fault diameter of interconnection networks. *Computers & Mathematics with Applications*, 13(5-6):577–582, 1987.

[22] M. Larrea, A. Fernández, and S. Arévalo. Optimal implementation of the weakest failure detector for solving consensus. In *19th IEEE Symposium on Reliable Distributed Systems, SRDS'00, Nürnberg, Germany, October 16-18, 2000, Proceedings*, pages 52–59, 2000.

[23] S.-C. Liaw, G. J. Chang, F. Cao, and D. F. Hsu. Fault-tolerant routing in circulant networks and cycle prefix networks. *Annals of Combinatorics*, 2(2):165–172, 1998.

[24] A. Pelc. Fault-tolerant broadcasting and gossiping in communication networks. *Networks*, 28(3):143–156, 1996.

[25] P. Ramanathan and K. G. Shin. Reliable broadcast in hypercube multicomputers. *IEEE Trans. Comput.*, 37(12):1654–1657, 1988.

[26] Repository. Hidden for double-blind review, 2016.

[27] Titan. Oak Ridge National Laboratory, https://www.olcf.ornl.gov/titan/, 2016.

[28] Y. Tock, B. Mandler, J. E. Moreira, and T. Jones. Design and implementation of a scalable membership service for supercomputer resiliency-aware runtime. In *Euro-Par 2013 Parallel Processing - 19th International Conference, Aachen, Germany, August 26-30, 2013. Proceedings*, pages 354–366, 2013.

[29] R. van Renesse, Y. Minsky, and M. Hayden. A gossip-style failure detection service. In *Proceedings of the IFIP International Conference on Distributed Systems Platforms and Open Distributed Processing*, Middleware '98, pages 55–70, London, UK, UK, 1998. Springer-Verlag.

[30] D. S. Wung. *Intelligent platform management interface (IPMI)*. PhD thesis, SLAC National Accelerator Laboratory, 2009.