

RT-ByzCast: Byzantine-Resilient Real-Time Reliable Broadcast

David Kozhaya¹, Jérémie Decouchant² and Paulo Esteves-Verissimo², *Fellow, IEEE, ACM*

¹ABB Corporate Research, Baden-Dättwil, Switzerland

²Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg

Today's cyber-physical systems face various impediments to achieving their intended goals, namely, communication uncertainties and faults, relative to the increased integration of networked and wireless devices, hinder the synchronism needed to meet real-time deadlines. Moreover, being critical, these systems are also exposed to significant security threats. This threat combination increases the risk of physical damage. This paper addresses these problems by studying how to build the first real-time Byzantine reliable broadcast protocol (RTBRB) tolerating network uncertainties, faults, and attacks. Previous literature describes either real-time reliable broadcast protocols, or asynchronous (non real-time) Byzantine ones.

We first prove that it is impossible to implement RTBRB using traditional distributed computing paradigms, e.g., where the error/failure detection mechanisms of processes are decoupled from the broadcast algorithm itself, even with the help of the most powerful failure detectors. We circumvent this impossibility by proposing *RT-ByzCast*, an algorithm based on aggregating digital signatures in a sliding time-window and on empowering processes with self-crashing capabilities to mask and bound losses. We show that *RT-ByzCast* (i) operates in real-time by proving that messages broadcast by correct processes are delivered within a known bounded delay, and (ii) is reliable by demonstrating that correct processes using our algorithm crash themselves with a negligible probability, even with message loss rates as high as 60%.

Index Terms—real-time distributed systems; probabilistic losses; reliable broadcast; byzantine behavior; intrusion tolerant;

I. INTRODUCTION

Many of today's physical structures are governed by automated operations that are typically conducted and controlled by multiple sensing, computing, and communication devices. Generally, such physical structures including their distributed control are known as *cyber-physical systems* (CPS) [1]. In this paper, we address two problems relevant to modern CPS, found for example in large continuous process plants, manufacturing shop-floors, or power grid installations.

The increasing integration of sensors and actuators, networked and often wireless [2]–[5], as well as their scale, introduce uncertainties and faults in communication, which hamper the necessary synchronism to meet real-time deadlines, fundamental for various CPS applications [3]–[6].

Worse, most of those processes are in essence critical infrastructures. Hence, they are likely to be targeted by motivated attackers [7], [8], to impact the timeliness of communications — for example by maliciously worsening the impact of the above mentioned uncertainties and faults — and the sheer correctness of messages. These two facets create a completely new scenario, where the traditional approach to building real-time communication no longer works. The consequences of system failure can range from loss of availability to physical damage.

In this paper, we simultaneously address the three threats — uncertainty, faults, and attacks — introducing *RT-ByzCast*, to the best of our knowledge the first Byzantine-resilient real-time reliable broadcast protocol. Previous literature on fault-tolerance contains either real-time reliable broadcast protocols [2], [9]–[12], or asynchronous (non real-time) Byzantine ones [13]–[17]. On the one hand, real-time reliable broadcast protocols cannot tolerate attacks: such algorithms would fail, even if a single process is compromised. Using naive platform restarts or cryptography alone is insufficient, and hence does not solve the problem [18], [19]. On the other hand, existing asynchronous Byzantine solutions do not support real-time. Yet, critical CPS applications must have information

disseminated correctly, reliably and in real-time, e.g., as in power system control applications that need to open and close multiple distributed circuit breakers based on thresholds and sensor data [20].

We approach the aforementioned threats first by formally specifying the needed broadcast properties via an abstraction, which we call **Real-Time Byzantine Reliable Broadcast** (RTBRB). We then prove that in systems where processes can be Byzantine and communication can simultaneously fail, it is impossible to implement the RTBRB abstraction using traditional distributed computing assumptions and techniques, e.g., where the error/failure detection mechanisms of processes are oblivious of the broadcast algorithm. We also prove that this holds even with the help of sophisticated failure detectors that perfectly reveal the identities of all faulty and untimely processes.

We propose an algorithm, which we call *RT-ByzCast*, which can cope with intermediate timing violations and uncertainties while still providing a reliable end result in real-time. We specifically prove that our algorithm indeed implements all properties of the RTBRB abstraction. The main idea underlying our *RT-ByzCast* algorithm is the use of a temporal and spatial diffusion mechanism of signed messages over a sliding time-window. This diffusion mechanism is augmented with a scheme to aggregate signatures corresponding to those messages. Hence, the time-window constitutes a slack. During that slack, the aggregation and diffusion mechanism tries to mask unanticipated violations, e.g., those caused by communication losses, by having processes repeatedly aggregate signatures relative to the same message and disseminate that message again (temporal diffusion) and to other processes (spatial diffusion).

In order to ensure safety, when losses cannot be masked in time using our aggregation and diffusion scheme, our algorithm triggers processes suffering such losses to crash themselves. Doing so, *RT-ByzCast* bounds the amount of time a correct process can lose communication with a quorum of

the network. This self-crashing capability allows our algorithm to circumvent our impossibility result affecting traditional distributed assumptions and techniques: we use process self-crashing to make sure that the view of correct processes is consistent between the error/failure detection mechanisms and our broadcast algorithm. *RT-ByzCast* can affect which processes are correct and which are not can revive processes that have self-crashed themselves.

Our *RT-ByzCast* algorithm relies on quorums. The reader might wonder that our principle, albeit safe, may lead the system to a complete shutdown, by having “too many” processes crash themselves. To this end, we perform a thorough simulation of the probability of a process crashing itself in our *RT-ByzCast* algorithm. We demonstrate that even for high loss rates (60% loss rate), the probability that a process crashes itself in *RT-ByzCast* is negligible for systems with more than 5 processes. We show as well that as the system size increases, the probability of a process crashing itself becomes asymptotic to 0 for any loss rate; hence the system liveness is almost always guaranteed. This matches the typical applications of the target CPS systems we envisage, e.g., the smart grid, where computing sensing and communication devices are deployed in large scale. Moreover, we devise a method that allows systems using our algorithm to tolerate any number of correct processes crashing themselves without shutting down. Our method is based on a scheme for detecting correct processes that crash themselves and over provisioning the number of processes (replicas) used.

Another concern of the reader might arise regarding the expected performance of our algorithm. We show that *RT-ByzCast* can meet the timing constraints of a large class of typical CPS applications, mainly in SCADA and IoT areas: power system automation and substation automation applications (time constants ≤ 100 ms); slow speed auto-control functions (≤ 500 ms); continuous control applications (≤ 1 s); or (iii) operator commands of SCADA applications (≤ 2 s).

In summary, the main contributions of this paper are:

- 1) Theoretical proofs showing that when processes and links can fail, it is impossible to implement RTBRB under traditional distributed computing paradigms that consider processes’ correctness checks as oblivious of the broadcast algorithm. We prove that the impossibility holds even when the identity of all faulty and unresponsive processes is known.
- 2) *RT-ByzCast*, an algorithm that implements the RTBRB abstraction by aggregating and diffusing digital signatures in a sliding time-window to mask individual communication losses. When messages cannot be masked in time, *RT-ByzCast* forces processes experiencing those losses to crash themselves.
- 3) A thorough simulation of the reliability and availability of our *RT-ByzCast* algorithm under many communication loss rates and system sizes.
- 4) A performance evaluation showing *RT-ByzCast*’s latency and network overhead.
- 5) Ways to adapt *RT-ByzCast* to handle churn.

The rest of the paper is organized as follows. Section II details the system model. Section III defines the real-time Byzantine reliable broadcast (RTBRB) abstraction. Section IV studies the feasibility of implementing RTBRB under tradi-

tional distributed computing assumptions. Section V presents the intuitions and details of our *RT-ByzCast* algorithm. Section VI demonstrates the reliability nad performance of our *RT-ByzCast* algorithm by (i) evaluating the probability of a process crashing itself, (ii) computing the probability that the system shuts down, (iii) showing how to allow a system to tolerate the crash of correct processes without causing system shutdown, and (iv) showing the broadcast delivery latency under different system sizes and message loss rates. Section VII discusses how to revive self-crashed processes. Section VIII illustrates and details how our *RT-ByzCast* algorithm can be adapted to handle dynamic systems. Section IX discusses *RT-ByzCast* application domains and systems. Finally Section X and Section XI discuss existing related work and conclude the paper respectively. For presentation purposes, we defer proofs and additional evaluations to a companion dedicated appendix.

II. SYSTEM AND THREAT MODEL

A. System Model

a) Processes: We consider a distributed system consisting of a set of $n > 1$ processes, denoted by $\Pi = \{p_1, p_2, \dots, p_n\}$. Processes are synchronous, i.e., the delay for performing a local step has a fixed known bound, assumed to be negligible compared to communication delays.

Processes have access to local clocks. For presentation simplicity, we assume that these clocks are synchronized, with a bounded skew. Using these clocks, processes define *synchronous rounds* of the same fixed duration (time-triggered). Rounds are synchronized among all processes, i.e., the start and end of a round occur at all hosts at the same time (with a bounded skew). For a particular round, clocks can trigger an initialization signal at all processes, upon which processes are awakened and initialized as part of the system.

b) Communication: Every pair of processes is connected by two logical uni-directional links. Precisely, processes p_i and p_j are connected by links l_{ij} and l_{ji} . Links can abstract a physical bus or a dedicated network link/path.

We assume that links are reliable and within the maximum delay with high probability. This means that in any transmission attempt, where a message is sent over a link, there is a high probability that the message reaches its destination within a maximum delay d after being transmitted. However, there is a small probability that reliability and timeliness are violated. Such violations exist in networks, as arguably all communication is prone to random disturbances, e.g., bad channel quality, interference, collisions, and buffer overflows [4].

We consider both message losses and delays as omissions. That is, late messages (violating the d delay assumption) are simply dropped (ignored). This way we treat *timing faults* [21] as omissions. We define the duration of any synchronous round to be d , the upper bound on a reliable and timely transmission. As such, any message sent at the beginning of round r , if not omitted, is assumed to be received by the end of round r . We consider that processes always send their messages at the beginning of a round. In our model, it is sufficient that delay d is only known by the local clocks that define rounds and not processes. Processes in some round r consider late any received message with a round number $< r$.

B. Threat Model

a) *Clocks*: We assume that the local synchronized clocks of non-Byzantine nodes are secure and hence cannot be attacked. Previous work showed that such secure and synchronized clocks can be built in similar environments using trusted components [22] or GPS [23].

b) *Processes*: We assume that some processes can exhibit arbitrary, a.k.a. *Byzantine*, behavior. Byzantine nodes can abstract processes that have been compromised by attackers, or are executing the algorithm incorrectly, e.g., as a result of some fault (software or hardware). A Byzantine process can behave arbitrarily, e.g., it may crash, fail to send or receive messages, delay messages, send arbitrary messages, etc.

We recall that in every transmission attempt a link may (with some probability) violate reliability and timeliness by dropping the message or delivering it within a delay $> d$. In both cases (dropped and delayed), that message is omitted; hence a sender needs to re-transmit that message again and face yet another risk of transmission failure. Due to omissions (losses and delays in consecutive transmission attempts) and the required follow-up re-transmissions, the time it takes to send a message reliably from one process to another (measured from the time of the first transmission attempt) may be unbounded. So, despite links being reliable and timely with high probability, our communication system is no longer synchronous. Our system (not being synchronous) can tolerate at maximum $f = \lfloor \frac{n-1}{3} \rfloor$ Byzantine processes. Such an f is proved to be the maximum number of Byzantine processes that an asynchronous system with n processes can tolerate to implement any form of agreement [17], [24]. f is known to the processes.

A process that exhibits a Byzantine behavior is termed *faulty*. Otherwise the process does not deviate from the specification of the algorithm and is said to be *non-Byzantine*. If the algorithm specifies some non-Byzantine processes to crash themselves within the algorithm's execution, then such crashed processes are termed faulty as well (excluded by the algorithm). The rest of the processes (the non-faulty ones) are termed as *correct*. More formally, consider an algorithm \mathcal{A} that begins execution at some global time t for a period ΔT .

Definition 1. All non-Byzantine processes that do not crash themselves, in $[t, t + \Delta T]$ are correct wr.t. \mathcal{A} . All other processes (Byzantine and non-Byzantine that crashed themselves) are faulty.

c) *Communication*: Links are assumed to be faithful and authenticated, i.e., a link does not alter the content of messages. Hence, Byzantine processes cannot modify messages sent on a link connecting correct processes. Recall however, that links can, with a small probability, violate reliability and timeliness. A message transmitted on link l_{ij} , $\forall i \neq j$, at any time t has probability $0 < \epsilon_1 < P_{ij}^o(t) < \epsilon_2 \ll 1$ of getting lost or delayed.

In practice, losses and delays can be correlated and may usually come in bursts, i.e., periods where messages are consecutively lost/delayed. Burst lengths can be well anticipated (with high coverage)¹. Accordingly, we denote by k the

¹For example in control systems it is very unlikely that a network link has more than two consecutive message losses. We refer readers to studies like [10], [25] that elaborate on how to estimate link burstiness.

maximum burst lengths (in rounds) that are expected to be observed in our network (a.k.a. omission degree). Note that this does not mean the bursts of length $> k$ rounds will not occur, it rather indicates that they seldom do and the system should still account for them. Namely, a link l_{ij} , $\forall i \neq j$, at any time t has probability $P_{ij}^k(t)$ of experiencing bursts of length k . $P_{ij}^k(t)$ is essentially a function of $P_{ij}^o(t)$, i.e., $P_{ij}^k(t) = \mathcal{F}(P_{ij}^o(t))$.

III. REAL-TIME BYZANTINE RELIABLE BROADCAST

We now define the properties of the real-time Byzantine reliable broadcast (RTBRB). Roughly speaking, RTBRB is a communication primitive that allows information to be disseminated reliably and timely in a one-to-all manner.

For simplicity, we provide the definition of RTBRB as an abstraction that allows at most one message to be delivered/broadcast. This definition can be easily extended to allow processes to broadcast multiple messages, e.g., see [15], [18] that resolve this using message identifiers.

- 1) **RTBRB-Validity:** If a correct process broadcasts m , then some correct process eventually delivers m .
- 2) **RTBRB-No duplication:** Every correct process delivers at most once a message.
- 3) **RTBRB-Integrity:** If some correct process delivers a message m with sender p_i and process p_i is correct, then m was previously broadcast by p_i .
- 4) **RTBRB-Agreement:** If some message m is delivered by any correct process, then every correct process eventually delivers m .
- 5) **RTBRB-Timeliness:** There exists a known Δ such that if a correct process broadcasts m at real-time t , no correct process delivers m after real time $t + \Delta$.

RTBRB provides processes with two operations, namely RTBRB-broadcast() and RTBRB-deliver(). A process broadcasts a message by invoking RTBRB-broadcast(). Similarly, a process delivers a message by invoking RTBRB-deliver().

We next compare RTBRB properties to those of existing broadcast abstractions highlighting the differences.

RTBRB Versus Existing Broadcast Primitives

a) *Byzantine Reliable Broadcast*: An asynchronous Byzantine reliable broadcast [13], [14] guarantees only a subset of the properties of RTBRB (all except RTBRB-Timeliness). In this sense, as opposed to asynchronous Byzantine reliable broadcast which delivers messages only eventually (i.e., with no predictability measures) our RTBRB primitive provides a timeliness bound: a message sent by correct processes is delivered within a known fixed duration after being broadcast.

b) *Real-Time Reliable Broadcast*: Synchronous (or real-time) reliable broadcasts [10], [26] provide known fixed bounds on delivering broadcast messages. However they only handle *fail-silent* process failures and hence cannot tolerate maliciousness.

c) *Atomic Broadcast*: Unlike liveness in the context of (asynchronous) reliable broadcast, the RTBRB-Timeliness property (a safety property) introduces a scent of (physical) ordering, since it stipulates, for each execution, a termination event to occur "at or before" some Δ on the time-line. This said, one may wonder to what extent does this go into

reaching a linear ordering, and thus in the direction of atomic broadcast. In RTBRB, the interleaving of broadcasts from multiple senders, namely when multiple broadcasts are issued in a period shorter than $\Delta - d$, might result in messages delivered to different processes in a different order. RTBRB-Timeliness ensures that a message m is delivered at some time in $[d, \Delta]$ after the broadcast.

IV. FEASIBILITY OF IMPLEMENTING RTBRB

We now study the feasibility of implementing RTBRB using traditional distributed assumptions and techniques.

A. RTBRB Using Failure Detectors

Traditionally, algorithms are designed to guarantee properties without affecting the correctness of processes. However, algorithms may rely on components, e.g., failure detectors (FDs) [27], which are external software blocks that provide hints about which processes are faulty. Failure detectors encapsulate synchrony assumptions. Thus, when added to systems with timing uncertainties (i.e., asynchronous), failure detectors make it possible to devise modular algorithms that solve, despite network uncertainties, difficult distributed computing problems, e.g., consensus [28].

In this section, we prove impossibility results showing that RTBRB cannot be implemented using traditional modular schemes relying on external software blocks, even when these blocks are the “most powerful” ones, such as perfect crash failure detectors, e.g., [27]. These impossibilities indicate that monitoring untimely processes in systems supported by real-time broadcast networks with timing uncertainties (timing and omission faults), should be done within the algorithm implementing RTBRB rather than externally. In short this means going against the modular failure detectors trend initiated with [27] and being more in line with the model of [26], hence featuring failure detectors and membership integrated with the broadcast protocol.

We give the intuition of why this is surprisingly so, and present proofs of the results. In this sense, we help establish, in this section, a clear understanding of the design constraints that should be followed/avoided when devising algorithms for distributed real-time systems exhibiting timing uncertainties, before delving into the system design. We first define a few essential terms.

Definition 2. A component is *oblivious of an algorithm* \mathcal{A} if that component does not adapt its behavior to the decisions and actions of \mathcal{A} . Such a component does not take any input from \mathcal{A} and hence any changes to \mathcal{A} result in no impact on the behavior of that component. Moreover, messages relative to that component are assumed to be independent of \mathcal{A} 's messages (i.e., messages are sent separately over the network).

Definition 3. A Perfect Crash Failure Detector (\mathcal{P}) [27] guarantees:

- 1) **Strong completeness:** Eventually every faulty process is permanently suspected by all correct processes.
- 2) **Strong accuracy:** A correct process is never suspected.

Assumption 1. Detector \mathcal{P} is oblivious of algorithm \mathcal{A} that implements RTBRB.

Assumption 2. Algorithm \mathcal{A} uses \mathcal{P} .

Assumption 3. Algorithm \mathcal{A} does not use any additional hardware or software components besides \mathcal{P} and the network.

Theorem 1. No algorithm \mathcal{A} can implement the RTBRB abstraction under Assumptions 1, 2 and 3.

The proof relies on the fact that even when knowing the identity of all correct processes, some correct processes might not be reachable in real-time (any predefined duration) through the network (due to probabilistic message losses). Traditional failure detectors [27] seek eventual reachability, while we seek real-time reachability. In other words, such an impossibility does not exist in time-free systems (a.k.a. asynchronous) but merely in those systems requiring a timeliness property. The reason follows from the fact that in time-free asynchronous systems, all correct processes are eventually reachable, even with our notion of links. Reachability changes from being a liveness property (eventual reachability) in time-free systems to being a safety property (timed reachability) in a real-time context. The detailed proof of Theorem 1 is subsumed by the proof of the more general result of Theorem 2.

We investigate next the feasibility of implementing RTBRB when failure detectors can detect, not only crashed processes, but also those processes that are not reachable in time. Perhaps surprisingly, we prove below that the impossibility still persists, if the failure detector remains oblivious of the algorithm implementing RTBRB.

B. RTBRB Using Proactive Reachability Failure Detectors

In order to determine which processes are alive and reachable in real-time distributed systems, a widely common practice is to use failure detectors that rely on heartbeats with a concrete timeliness specification [25], [29]–[32].

Definition 4. Consider some fixed interval of l rounds denoted by Δ_l . Process p_i is said to be **timely reachable** at round r' , if for all other correct processes p_j in the system, p_j has received some message from p_i within the past Δ_l interval from r' . Conversely, a process p_i is **unreachable**, if p_i is not timely reachable at some round by at least one other correct process.

Definition 5. A Proactive Reachability Failure Detector (\mathcal{PR}) guarantees the following properties:

- 1) **Strong timed completeness:** Every unreachable process is permanently suspected by all correct processes, at most Δ_m rounds after becoming unreachable.
- 2) **Strong timed accuracy:** No correct process that is timely reachable is suspected before crashing or becoming unreachable.

Note that the notion of reachability subsumes the notion of crashed (faulty). It is also important to observe that the notion of reachability in Definition 4 is the one that restricts the set of correct and reachable processes the most: the set of processes termed as unreachable is largest under Definition 4. In this sense, impossibilities proven with \mathcal{PR} using our definition of reachability are the strongest.

An example of a \mathcal{PR} implementation in our system is one that: (1) relies on timeouts (e.g. expiring every Δ_l rounds) and periodic message (heartbeat) exchange between processes, (2) defines timeouts on process p_i as a multiple of the period at which p_i sends heartbeats, and (3) “suspects” a process p_i only when some correct process cannot receive, within the specified timeouts, (direct or indirect) heartbeats sent by p_i .

Assumption 4. Detector \mathcal{PR} is oblivious of algorithm \mathcal{A} that implements RTBRB.

Assumption 5. Algorithm \mathcal{A} uses \mathcal{PR} .

Assumption 6. Algorithm \mathcal{A} does not use any additional software or hardware components besides \mathcal{PR} and the network.

Theorem 2. No algorithm \mathcal{A} can implement the RTBRB abstraction under Assumptions 4, 5 and 6.

Roughly speaking, the proof hinges on two main ideas:

- 1) The reachability uncovered by \mathcal{PR} has a different time-window than that of \mathcal{A} . Namely, suspected processes (by \mathcal{PR}) are processes that are not reachable up to the current moment, while \mathcal{A} would need to see the reachability from the current moment onward.
- 2) In the same network, due to probabilistic violations of reliability and timeliness, different oblivious applications (which send messages independently of each other) can reach different processes given specific time-constraints. Hence, processes considered to be reachable from \mathcal{PR} 's perspective might not be reachable in time by the broadcast algorithm.

The detailed proof of Theorem 2 can be found in Appendix A.

Corollary 1. Detecting unreachable processes using a failure detector is not sufficient to allow the implementation of the RTBRB abstraction if that failure detector is oblivious of the algorithm implementing RTBRB.

In short, in order to implement RTBRB in a probabilistically synchronous network, we need to (1) integrate the proactive reachability failure detector within the protocol implementing RTBRB and (2) allow this compound to decide which processes are reachable within the hard timing constraints. In other words, in order to implement RTBRB, the reachability from the perspective of the algorithm implementing RTBRB should be reconciled with that of \mathcal{PR} . We accordingly propose in what follows a solution that implements RTBRB by embedding the \mathcal{PR} functionality in the RTBRB protocol.

V. AN ALGORITHM IMPLEMENTING RTBRB

In this section, we propose an algorithm, which we call *RT-ByzCast* that implements the RTBRB abstraction by avoiding the traditional design pitfalls highlighted in Section IV. *RT-ByzCast* relies on three main things: processes monitoring each other, message diffusion and signature aggregation over a time-window, and processes capable of crashing themselves. We present the intuition behind each.

a) *Process monitoring*: Processes do not know beforehand when a broadcast might be invoked. In our system, omissions on any link may be of unbounded durations (even if with a small probability). Thus, there might exist a correct process p that stays unaware of a broadcast invoked by process q : p may lose for an unbounded amount of time messages relative q 's broadcast.

In order to provide any form of real-time guarantee, it is crucial to detect the occurrence of such timing violations and react to them. We develop the *proof-of-life* function (Function 1), which requires processes to periodically exchange heartbeats (and echo received ones) when they do not know of any broadcast. A process p that does not hear enough echoes

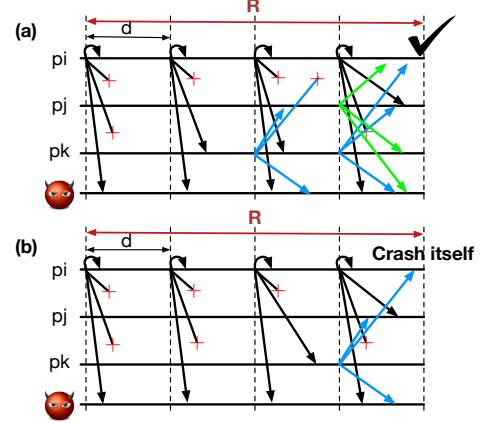


Figure 1. (a) p_i gathers $2f + 1$ signatures in \mathbb{R} rounds; (b) p_i does not gather $2f + 1$ signatures in \mathbb{R} rounds and crashes itself.

of its own heartbeats can thus suspect that it either has lost communication with many other processes or that it does not know about an invoked broadcast.

b) *Message diffusion and signature aggregation*: In our system, messages can be omitted, hence re-transmissions are needed to guarantee successful delivery. A digital signature of process p on some message m constitutes an unforgeable proof that p processed m , is aware of it, and has sent (relayed) m . Our algorithm requires processes to aggregate, over some pre-defined time-window, the different observed signatures on a received broadcast message (including their own signature). This way a process receiving a message m can know (from m 's aggregated signature) the identity of processes that have received m (within the pre-defined time-window). Besides aggregation our algorithm requires processes to diffuse a received message m (along with its aggregated signature) by re-transmitting m within the pre-defined time-window multiple times and to other processes. This diffusion helps mask network omissions by exploiting multiple network paths, e.g., relaying information via intermediate processes.

In fact, our time-window creates a slack period of “collective re-transmissions” in which, even if a message is omitted a few times, it still has a high chance (as shown in Section VI) of being successfully received by the end of that time-window (which can be viewed now as the new deadline). We denote by \mathbb{R} the duration of that time-window (in rounds). \mathbb{R} is a parameter of our algorithm (known to processes) and represents the delay we tolerate for a round-trip message delay. The larger the value of \mathbb{R} , the bigger the slack period, but the higher the chances of a message being received in that duration (see Section VI). In this sense, $\mathbb{R} = f(k)$, meaning that \mathbb{R} is a function of k (the anticipated omission degree), $\mathbb{R} \geq 2k + 2$ to account for a round-trip bound.

c) *Self-crash capabilities*: A message sent by a non-Byzantine process p might still fail to reach “enough” processes, despite diffusion and aggregation within a time-window of duration \mathbb{R} (assumptions on omission degree, despite their high coverage, might be violated). In that case, our algorithm triggers p to crash itself (we formally define when a process crashes itself in Definition 7). A non-Byzantine process p that “crashes itself” is no longer correct with respect to the algorithm's correctness properties.

Function 1 *proof-of-life*(Δ_r)@ process p_j

```

1: for every round  $r$  do
2:   Send  $\text{HB}(\{p_j, r\}; \Phi_{p_j})$  to all  $p \in \Pi$ 
3: end for
4: upon event <receive  $\text{HB}(\{p_i, r\}; \text{sig}_s)$  at some round  $r' >$  do
5: //  $\text{sig}_s$  designates the signatures on  $\{p_i, r\}$ .
6:   Verify message validity
7:   Set  $\text{sig}_s = \text{sig}_s \cup \Phi_{p_j}$ 
8:   for every round in  $[r' + 1, r + \Delta_r]$  do
9:     Send  $\text{HB}(\{p_i, r\}; \text{sig}_s)$  to all  $p \in \Pi$ 
10:    end for
11:
12: for every round  $r$  where  $r > \text{time}_w$  do
13:   Compute  $\mathcal{R}_{HB}(p_j, r)$ 
14: //  $\mathcal{R}_{HB}(p_j, r)$ : set of processes that received some heartbeat from  $p_j$  (directly or indirectly), during some round in  $[r - 1 - \Delta_r, r - 1]$ 
15:   if  $|\mathcal{R}_{HB}(p_j, r)| \leq 2f$  then
16:     Transition to Dead-State
17:   end if
18: end for

```

A. *RT-ByzCast Overview*

We present now a high-level view of how our algorithm works in systems where $n = 3f + 1$ processes.

1) Behavior of Processes Unaware of a Broadcast

Every process that did not receive any message relative to a broadcast, executes the *proof-of-life* function (Function 1), where a process periodically sends a signed heartbeat message to all other processes. When process p_j receives p_i 's heartbeat at round r , p_j : (i) aggregates signatures relative to p_i 's heartbeats (signatures of heartbeats sent by p_i in $[r, r + \mathbb{R}]$), (ii) appends its own signature to the formed aggregate, and (iii) periodically echoes p_i 's heartbeat, with the aggregated signatures, to all other processes (see Figure 1). A process p executing the *proof-of-life* crashes itself when p does not receive in any time-window of duration \mathbb{R} , $2f + 1$ signatures on its own heartbeats (including its own signature).

2) Behavior of Processes Knowledgeable of a Broadcast

In this case, a process echoes periodically the broadcasted value, only if that value is indeed sent by the process that originally issued the broadcast. If multiple values are heard relative to a single process, the first heard value is the one to be echoed. Processes continue to execute the proof-of-life function, however, piggybacking proof-of-life messages to the echo messages of the broadcast.

Similar to the proof-of-life function, processes aggregate and append signatures relative to a given broadcast value in a window of \mathbb{R} rounds (from the current round). Upon receiving a value that is signed by more than $2f$ processes, a process delivers that value. A process that does not receive more than $2f$ signatures in a \mathbb{R} time-window after being aware of the broadcast crashes itself (see Definition 7).

3) RT-ByzCast Latency

Our *RT-ByzCast* algorithm delivers messages within an upper bound of $3\mathbb{R}$ rounds. The following example briefly highlights the intuition behind requiring $3\mathbb{R}$ (the detailed proof can be found in Appendix B). Assume a correct process issues a broadcast at round r . By round $r + \mathbb{R}$, at least $2f$ other processes would have seen that broadcast (otherwise the broadcasting process kills itself). In the worst case, those processes would see the broadcast at round $r + \mathbb{R} - 1$. Assume that the remaining f processes in the system have not yet received the broadcast message. By round $r + 2\mathbb{R}$ these f processes should have heard about the broadcast or otherwise

Function 2 *aggregate-sig* _{p_j} ($v, p_i, \Phi_{p_x}, \dots, \Phi_{p_z}, p_k$)

```

1: if  $(v, \dots) \notin \text{Msg}[p_i][p_k]$  then
2:   Set  $\text{sig}_s = \Phi_{p_j} \cup \Phi_{p_x}, \dots, \Phi_{p_z}$ , after verifying that  $p_x, \dots, p_z$  indeed signed  $v$ .
3:   Add  $(v, \text{sig}_s)$  to  $\text{Msg}[p_i][p_k]$ ,
4: end if
5: if  $(v, \dots) \in \text{Msg}[p_i][p_k]$  then
6:   Consider that  $\text{sig}_s = \dots$ , then update  $(v, \text{sig}_s)$  in  $\text{Msg}[p_i][p_k]$  such that  $(v, \text{sig}_s)$  includes as well the signatures relative to  $p_x, \dots, p_z$ , i.e.,  $\text{sig}_s = \text{sig}_s \cup \Phi_{p_x}, \dots, \Phi_{p_z}$ .
7: end if
8: return  $\text{sig}_s$ ;

```

Function 3 *deliver-message* _{p_j} (p_i, v, sig_s) @ round r

```

1: Deliver  $v$  (if not already delivered)
2: Initialize  $\mathcal{R}_{deliver}(p_i, r) = p_j$ 
3: Stop sending any Echo()
4: Send  $\text{Deliver}_{p_j}((p_i, v, \text{sig}_s); \Phi_{p_j})$  at the beginning of every round in  $[r + 1, r + 1 + 2\mathbb{R}]$ , where  $\text{sig}_s$  contains signatures of all processes in  $\mathcal{R}_{echo}(p_i, r_0, v)$ .
   //  $\Phi_{p_j}$  is  $p_j$ 's signature on the new payload:  $(p_i, v, \text{sig}_s)$ 
5: At the beginning of round  $r + 2 + 2\mathbb{R}$ :
6: if (no Echo() or RTBRB-broadcast() is being sent) then
7:   Execute proof-of-life( $\mathbb{R}$ ) (non-piggyback mode).
8: end if

```

they would kill themselves (processes that already received the broadcast send heartbeats only via echo messages relative to the broadcast). After receiving the broadcast, these f processes have an additional \mathbb{R} rounds to collect enough signatures and hence will deliver the broadcast message at worst by $r + 3\mathbb{R}$.

B. Description of RT-ByzCast

We now describe our algorithm formally and in more details. Using our *RT-ByzCast* algorithm, non-Byzantine processes can be in one of two states: the *Alive-State* or the *Dead-State*. Non-Byzantine processes in the *Alive-State* follow the algorithm faithfully. Processes in the *Dead-State*, however, refrain from sending any messages and do not respond to any higher-level application requests. A process in the *Dead-State* appears to have crashed as in *fail-stop* failures [16]. For presentation simplicity, we consider that once a process is in the *Dead-State*, it remains in the *Dead-State* forever. In Section VII, we discuss how processes can be revived, by returning to the *Alive-State* and participating in the algorithm.

Initially, after receiving an initiation signal from the clock, all processes are awakened and execute in the *Alive-State*. Every process p_i defines a time-window of a known fixed size of $\mathbb{R} > 1$ rounds and $\text{Msg}[]^n[]^n$, an initially empty array to store the values broadcast relative to each process, with the corresponding signatures on these values. $\text{Msg}[]^n[]^n$ is the local data structure that stores relative to each process that issues a broadcast and each process that echoes it, the broadcast values and their aggregated signatures. For example, $\text{Msg}[p_i][p_k]$ stores all messages with aggregated signatures that are broadcast by some process p_i and that are echoed by a process p_k . After initialization, every process in the *Alive-State* executes a *proof-of-life* function with parameter $\Delta_r = \mathbb{R}$.

A process p_i that executes the *proof-of-life* function sends a heartbeat $\text{HB}(\{p_i, r\}; \Phi_{p_i})$ to all processes at the beginning of every round r , where Φ_{p_i} denotes p_i 's signature over the heartbeat. A process p_j receiving $\text{HB}(\{p_i, r\}; \text{sig}_s)$ at some round r' verifies that the heartbeat has been indeed seen by the processes that signed it (sig_s is the set of

signatures on $\{p_i, r\}$). Then p_j appends its signature to $sigs$ and sends $\text{HB}(\{p_i, r\}; sigs)$ to all other processes in all rounds $\in [r'+1, r+\Delta_r]$ (lines 7-10). A process p_i computes at the beginning of every round $r : r > \Delta_r$ the set $\mathcal{R}_{HB}(p_i, r)$, which is the set of distinct processes that were able to receive some heartbeat from p_i (directly or indirectly), during some round in $[r - 1 - \Delta_r, r - 1]$.

Precisely, let us denote by M the set of $\text{HB}(\{p_i, r''\}; *)$ heartbeats that p_i received. Then $\mathcal{R}_{HB}(p_i, r)$ contains all processes that have their signature in any heartbeat in M such that round $r'' \in [r - 1 - \Delta_r, r - 1]$. For any round after Δ_r rounds elapse since system initialization (i.e., \forall rounds $r : r > \Delta_r$), if ever $|\mathcal{R}_{HB}(p_i, r)| \leq 2f$, then process p_i transitions to the *Dead-State* (lines 12-18).

RT-ByzCast: The Algorithm (Algorithm 4)

After system start-up, every process p_i begins by executing the proof-of-life function with $\Delta_r = \mathbb{R}$. A process p_i that wishes to broadcast a value v at round $r-1$ initializes an empty set denoted by $\mathcal{R}_{echo}(p_i, r, v)$. Then p_i signs $(p_i, r-1, v)$ with an unforgeable signature Φ_{p_i} and produces the tuple $\{(p_i, r-1, v); \Phi_{p_i}\}$. This tuple is made of two parts: a *payload* and a *signed-by* part. For example, the tuple $\{(p_i, r-1, v); \Phi_{p_i}\}$ has $(p_i, r-1, v)$ as payload and Φ_{p_i} as the signed-by. Afterwards process p_i sends at the beginning of round $r-1$, the message $\text{RTBRB-broadcast}((p_i, r-1, v); \Phi_{p_i})$ to all processes. When a process p_j (that is executing the proof-of-life function) first receives an $\text{RTBRB-broadcast}((p_i, r', v); \Phi_{p_i})$ message in round $r \geq r'$ from p_i , process p_j initializes an empty set denoted by $\mathcal{R}_{echo}^{p_j}(p_i, r+1, v; \Phi_{p_i})$. Afterwards p_j retrieves the tuple $\{(p_i, r', v); \Phi_{p_i}\}$, verifies that p_i indeed sent v , and produces the tuple $\{(p_i, r', v; \Phi_{p_i}); \Phi_{p_j}\}$ by appending p_j 's signature to the signed-by field.

Process p_j starts sending at the beginning of every round (as of round r onward) $\text{Echo}_{p_j}((p_i, r', v; \Phi_{p_i}); \Phi_{p_j})$ to all other processes. Process p_j continues to execute the *proof-of-life* function however by piggybacking its heartbeats on $\text{Echo}_{p_j}((p_i, r', v; \Phi_{p_i}); \Phi_{p_j})$.

When process p_j receives a (valid) $\text{Echo}_{p_k}((p_i, r', v; \Phi_{p_i}); \Phi_{p_x}, \dots, \Phi_{p_z})$ at some round r , p_j behaves as follows.

A. If p_j is not sending any $\text{Echo}()$ (lines 14-27): then if p_j has not already delivered a message from p_i , p_j aggregates signatures in the variable $sigs$. It does so using the function $\text{aggregate-sig}_{p_j}(\dots)$ (detailed Function 2) that combines all observed signatures on messages sent by p_k containing (p_i, r', v) . Then if signatures in $sigs$ do not add up to $2f+1$ processes, p_j sends $\text{Echo}_{p_j}((p_i, r+1, v; \Phi_{p_i}); \Phi_{p_j})$ and initializes a set ($\mathcal{R}_{echo}(p_i, r+1, v)$) that tracks the number of signatures collected on the $\text{Echo}_{p_j}((p_i, r+1, v; \Phi_{p_i}); \dots)$. Otherwise, i.e., if signatures in $sigs$ amount to at least $2f+1$ processes, p_j sets $\mathcal{R}_{echo}(p_i, r, v) = sigs$ and delivers the message by executing $\text{deliver-message}(v, sigs)$. Any process p_j that executes $\text{deliver-message}(p_i, v, sigs)$ delivers message $(p_i, v, sigs)$ and only sends $\text{Deliver}_{p_j}((p_i, v, sigs); \dots)$ for $2\mathbb{R}$ rounds. p_j should collect a least $2f+1$ signatures during the first \mathbb{R} rounds. This requirement ensures that there is a full \mathbb{R} -duration during which at least $f+1$ correct processes are

sending only such deliver messages².

If no $\text{Echo}()$ or $\text{RTBRB-broadcast}()$ is being sent/received, p_j executes the proof-of-life function with parameter \mathbb{R} (without piggybacking).

Algorithm 4 RT-ByzCast

```

1: Init:  $\text{Msg}[]^n[]^n = \emptyset$ 
2: Execute proof-of-life( $\mathbb{R}$ );
3: upon event  $< p_i \text{ wants to broadcast a value } v >$  do
4:   Execute proof-of-life function in piggyback mode
5:   Initialize  $\mathcal{R}_{echo}(p_i, r, v) = \emptyset$ 
6:   Send periodically starting from the current round
       $\text{RTBRB-broadcast}((p_i, r_{current}, v); \Phi_{p_i})$  to all  $p \in \Pi$ 
7:
8: upon event  $< \text{receive RTBRB-broadcast}((p_i, r', v); \Phi_{p_i}) \text{ in round } r - 1 \geq r' \text{ for the first time} >$  do
9:   Execute proof-of-life function in piggyback mode
10:  Initialize  $\mathcal{R}_{echo}^{p_j}(p_i, r, v) = p_i; sigs = \Phi_{p_j}$ 
11:  Send  $\text{Echo}_{p_j}((p_i, r', v; \Phi_{p_i}); sigs)$  to all  $p \in \Pi$  at rounds  $\geq r$ 
12:
@ process  $p_j$ :
13: upon event  $< \text{receive Echo}_{p_k}((p_i, r', v; \Phi_{p_i}); \Phi_{p_x}, \dots, \Phi_{p_z}) \text{ at round } r >$  do
14:   if  $p_j$  is not sending any  $\text{Echo}()$  then
15:     Set  $sigs = \text{aggregate-sig}_{p_j}(v, p_i, \Phi_{p_x} \dots \Phi_{p_z}, p_k)$ 
16:   if  $p_j$  has not already delivered a message relative to  $p_i$  then
17:     Execute proof-of-life function in piggyback mode.
18:     Initialize  $\mathcal{R}_{echo}(p_i, r+1, v) = \emptyset$ 
19:   if  $sigs \leq 2f$  then
20:     Send at the beginning of every cycle (as of round  $r+1$  onward)  $\text{Echo}_{p_j}((p_i, r+1, v; \Phi_{p_i}); \Phi_{p_j})$  if  $k \neq j$ 
21:   end if
22: end if
23: if  $sigs > 2f$  (for the first time) then
24:   Set  $\mathcal{R}_{echo}(p_i, r, v) = sigs$ 
25:   Execute deliver-message( $p_i, v, sigs$ )
26: end if
27: end if
28: if  $p_j$  is sending an  $\text{Echo}_{p_j}((p_i, r', v; \Phi_{p_i}); \dots)$  then
29:   Set  $sigs = \text{aggregate-sig}_{p_j}(v, p_i, \Phi_{p_x} \dots \Phi_{p_z}, p_k)$ 
30:   if  $sigs > 2f$  (for the first time) then
31:     Set  $\mathcal{R}_{echo}(p_i, r, v) = sigs$ 
32:     Execute deliver-message( $p_i, v, sigs$ ).
33: end if
34: if  $sigs \leq 2f \wedge k = j$  then
35:   Set  $\mathcal{R}_{echo}(p_i, r, v) = sigs$ .
36: end if
37: if  $sigs \leq 2f$  then
38:   Send at the beginning of every cycle (as of round  $r+1$  onward)  $\text{Echo}_{p_k}((p_i, r', v; \Phi_{p_i}); sigs)$  to all  $p \in \Pi$ 
39: end if
40: end if
41: if  $p_j$  is sending  $\text{Echo}_{p_j}((p_i, r'', v'; \Phi_{p_i}); \dots) : v' \neq v$  then
42:   Set  $sigs = \text{aggregate-sig}_{p_j}(v', p_i, \Phi_{p_x} \dots \Phi_{p_z}, p_k)$ 
43:   if  $sigs > 2f$  (for the first time) then
44:     Set  $\mathcal{R}_{echo}(p_i, r, v) = sigs$ 
45:     Execute deliver-message( $p_i, v', sigs$ )
46:   end if
47: end if
48:
49: upon event  $< \text{receive Deliver}_{p_k}((p_i, v, sigs); \Phi_{p_x} \dots \Phi_{p_z}) \text{ at round } r >$  do
50:   if  $((p_i, v)$  is not delivered yet then
51:     Deliver  $v$ 
52:     Stop sending any  $\text{Echo}()$ 
53:     Initialize set  $\mathcal{R}_{deliver}(p_i, r) = \{p_x, \dots, p_z\}$ 
54:   else
55:      $\mathcal{R}_{deliver}(p_i, r') = \mathcal{R}_{deliver}(p_i, r') \cup \{p_x, \dots, p_z\}$ .
56:   end if
57:   Send  $\text{Deliver}_{p_j}((p_i, v, sigs); \text{signatures})$  to all  $p \in \Pi$  at every round in  $[r+1, r+1+2\mathbb{R}]$ , signatures contains the signatures of all processes in  $\mathcal{R}_{deliver}(p_i, \dots)$ .
58:   Execute same commands as lines 5-8 of deliver-message $p_j$ (...)
59:

```

²This requirement ensures agreement on delivering the same message among all correct processes, when broadcasting nodes are malicious.

B. If p_j is sending an $\text{Echo}_{p_j}((p_i, r', v; \Phi_{p_i}); \dots)$ (lines 30-42): In this case, if p_j has not delivered a message from p_i , p_j aggregates signatures in $sigs$ (using the function $\text{aggregate-sig}_{p_j}(\dots)$). Then p_j checks if $sigs$ contains more than $2f$ signatures, in which case p_j sets $\mathcal{R}_{\text{echo}}(p_i, r, v) = sigs$ and delivers (p_i, r', v) by executing $\text{deliver-message}(p_i, v, sigs)$. Otherwise, i.e., if $|sigs| \leq 2f$, p_j sends echoes of (p_i, r', v) with the new aggregated signatures, if $(k = j)$.

C. If p_j is sending an $\text{Echo}_{p_j}((p_i, r'', v'; \Phi_{p_i}); *)$ where $v' \neq v$: p_j aggregates signatures in $sigs$ (using the function $\text{aggregate-sig}_{p_j}(v', p_i, \Phi_{p_x} \dots \Phi_{p_z}, \Phi_{p_k})$). Then p_j checks if $sigs$ contains more than $2f$ signatures, in which case p_j delivers (p_i, r'', v') by executing $\text{deliver-message}(p_i, v', sigs)$.

If p_j receives $\text{Deliver}_{p_k}((p_i, v, sigs); \Phi_{p_x} \dots \Phi_{p_z})$ for the first time at round r (lines 49-58), p_j delivers $(p_i, v, sigs)$ by executing $\text{deliver-message}(p_i, v, sigs)$. If p_j has previously seen such a $\text{Deliver}()$ message, p_j aggregates in the variable signatures all the signatures it has on $\text{Deliver}_{p_j}((p_i, v, sigs); \dots)$ messages. Then p_j sends $\text{Deliver}_{p_j}((p_i, v, sigs); \text{signatures})$ at the beginning of every round $\in [r + 1, r + 1 + 2\mathbb{R}]$.

At the beginning of round $r + 2 + 2\mathbb{R}$, if p_j is not sending any $\text{Echo}()$ or RTBRB-broadcast(), then p_j executes the proof-of-life function with parameter \mathbb{R} (without piggybacking).

Definition 6. Any RTBRB-broadcast(p_i, \dots) or $\text{Echo}(p_i, \dots)$ message is termed invalid when it possesses any incorrect signature. Any $\text{Deliver}((p_i, \dots, sigs); \dots)$ is termed invalid when it possesses any incorrect signature, or when $sigs$ has less than $2f + 1$ correct signatures. If a process receives invalid messages, it ignores them.

Remark 1. A process sending RTBRB-broadcast(), $\text{Echo}()$ or $\text{Deliver}_{p_j}()$, executes proof-of-life by piggybacking heartbeats to the sent RTBRB-broadcast(), $\text{Echo}_{p_j}()$ and $\text{Deliver}_{p_j}()$; otherwise heartbeats are sent in individual messages. This is represented in lines (4), (9), and (17) of Algorithm 4.

Definition 7. A process p_j crashes itself and transitions to the Dead-State at the end of round r_{cur} if any of the four cases below is satisfied.

Case 1. A process p_j transitions to the Dead-State if both conditions below hold.

- 1) $\exists r': |\mathcal{R}_{\text{echo}}(p_i, r', v)| \leq 2f, \forall v \wedge r_{\text{cur}} \geq r' + \mathbb{R}$
- 2) If p_j has not discovered a lie. A process p_j is said to “discover a lie”, if p_j can verify that two different values have been sent by the process issuing the broadcast.

Case 2. p_j transitions to the Dead-State if both conditions hold: $\exists r': |\mathcal{R}_{\text{deliver}}(p_i, r')| \leq 2f$, and $r_{\text{cur}} \geq r' + \mathbb{R}$.

Case 3. Process p_j transitions to the Dead-State if p_j does not hear from at least $2f + 1$ processes in any \mathbb{R} window.

Case 4. Process p_j transitions to the Dead-State if p_j in some duration \mathbb{R} does not see echoes of its messages from at least $2f + 1$ processes.

We give now a brief intuition behind each of the four cases. In Case 1, the first condition ensures that processes, which cannot communicate with a Byzantine quorum, crash themselves. The second condition avoids the denial of service

attack that a Byzantine process can launch. For Case 2 and Case 3, consider a setting where a correct process p aggregates $2f + 1$ signatures, delivers a message and directly loses communication with other processes. Now assume that other processes detect that the broadcasting process is lying. These other processes do not kill themselves (by Case 1). If p stays alive this would violate agreement. Case 2 eliminates this violation. Case 3 eliminates a violation in agreement when the processes that detect a lie are fewer than $f + 1$ and get disconnected temporarily from the other correct processes that deliver some v . Case 4 makes sure that any process whose messages are not seen by (at least) $2f + 1$ processes in any duration of value \mathbb{R} crashes itself. This case 4 is used to allow the detection of processes that crash themselves (see Section VI-C).

We present in Appendix B the proof that *RT-ByzCast* implements the RTBRB abstraction.

VI. EVALUATION

We now present an analytic performance study of *RT-ByzCast*. We evaluate first the probability of having a correct process in *RT-ByzCast*, enter the *Dead-State*, i.e., crash itself.

A. Probability of a Correct Process Shutdown

The probability of a correct process crashing itself is crucial, as it may hinder the reliability of the whole system, namely the system’s liveness. For instance, in systems where $n = 3f + 1$ (which are a subset of systems with $f < \frac{n}{3}$), having a single correct process crash itself can result in a cascade of self-induced crashes leading the whole system to shutdown. This cascade is caused by the fact that one correct process crashing itself may prevent correct processes ever gathering the required $2f + 1$ quorum, e.g., if in the worst case all f processes are Byzantine in addition to the correct process that crashed itself.

In our simulation setting, we first assume that all messages can be lost/omitted independently and with the same probability, i.e., we assume that $P_{ij}^o(t) = p \forall t, i \neq j$. In Appendix C, we study the effect of having correlated losses/omissions that may result in bursts and show that a large enough slack window masks losses/omissions regardless of correlations and preserves drawn conclusions under the independent assumption. We describe the simulation setting in which we perform our evaluation.

Let us denote by \mathcal{C} , the minimum set of non-Byzantine processes, i.e., $|\mathcal{C}| = n - f$. We run our simulations for $|\mathcal{C}| \in \{5, 10, 20, 30, 40, 50, 100, 200\}$ processes. For each value of $|\mathcal{C}|$, we consider various probability values with which a sent message can be lost/omitted. Namely, we consider the probability of losing/omitting a message sent at any point in time to be $p \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$. For a given value of $|\mathcal{C}|$ and p , we invoke a broadcast at one of the processes and record, after \mathbb{R} rounds of communication, if any process does not receive $|\mathcal{C}|$ signatures on the value being broadcast. We repeat such an instance 10^6 times. We report our results showing:

$$\frac{\text{num. of instances in which some correct process crashes itself}}{10^6},$$

for $\mathbb{R} \in \{5, 6, 10, 15\}$ rounds respectively. We select these values of \mathbb{R} to show that a well chosen fixed value regardless

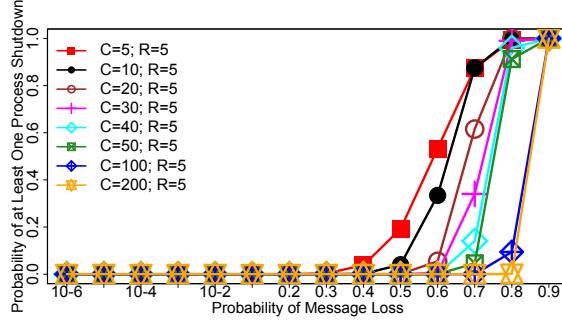


Figure 2. The probability of a correct process crashing itself in a system after 10 communication rounds ($\mathbb{R} = 5$).

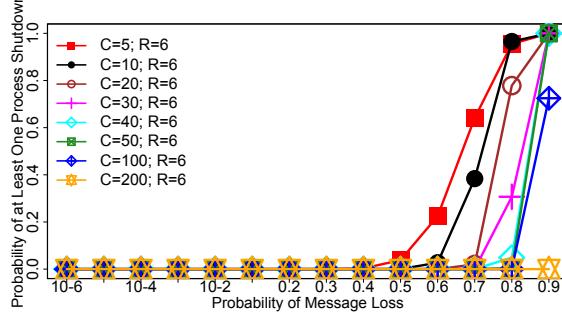


Figure 3. The probability of a correct process crashing itself in a system after 10 communication rounds ($\mathbb{R} = 6$).

of system size and loss probability allows our *RT-ByzCast* algorithm to be reliable.

Our results in Figure 2, Figure 3, Figure 4, and Figure 5 show that our *RT-ByzCast* algorithm achieves a higher reliability, i.e., a smaller probability of having a correct process crash itself, as the number of non-Byzantine processes, $|\mathcal{C}|$, increases. This aspect of our algorithm is fundamental to the typical applications of the target CPS systems we envisage, e.g., the smart grid, where computing sensing and communication devices are deployed on a large scale.

Moreover, from a different perspective, namely by fixing $|\mathcal{C}|$ and varying the window size (\mathbb{R}), we show in Appendix D that increasing the number of communication rounds (i.e., the value of \mathbb{R}) increases the reliability of our *RT-ByzCast* algorithm for any number of correct processes. In fact, with $\mathbb{R} = 10$ the probability of a process crashing itself becomes negligible even with up to 60% losses/omissions rate.

B. Tolerating A Correct Process Self-Crash

Despite the small probability of having a process crash itself, we provide in this section a means to allow the system using *RT-ByzCast* to tolerate processes self-crashing while still being available. Let us respectively denote by $Prob(p_{shutdown})$ and $Prob(sys_{shutdown})$ the probability that a process kills itself and the probability that the whole system shuts down. A system where $n = 3f + 1$ shuts down, in the worst-case, as soon as a single process crashes itself. Hence, an upper bound on the probability that the whole system shuts down can be formally expressed by:

$$Prob(sys_{shutdown}) = 1 - (1 - Prob(p_{shutdown}))^{2f+1}. \quad (1)$$

A system using our *RT-ByzCast* algorithm can be made more reliable while tolerating the maximum number of Byzant-

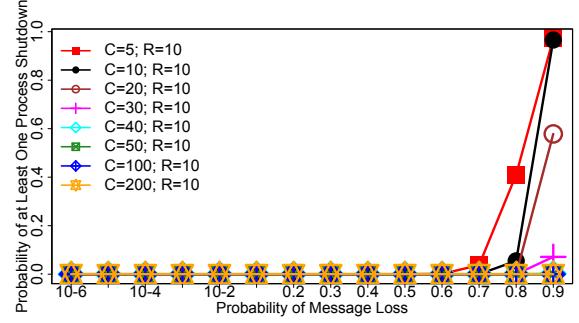


Figure 4. The probability of a correct process crashing itself in a system after 10 communication rounds ($\mathbb{R} = 10$).

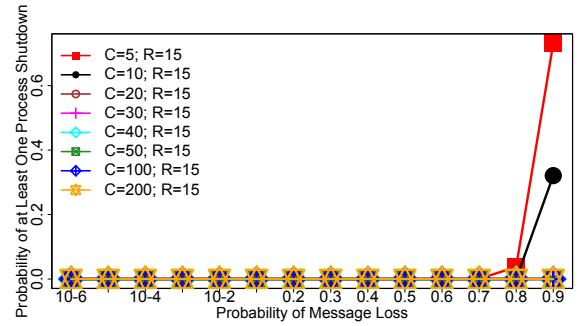


Figure 5. The probability of a correct process crashing itself in a system after 15 communication rounds ($\mathbb{R} = 15$).

tine processes. Precisely, we can allow the system to tolerate a single correct process crashing itself, after which any additional correct processes that crashes itself would trigger the whole system to shutdown. To do so, for any desired value of f , the maximum number of tolerable Byzantine processes, we choose the total number of processes to be $n = 3f + 3$. Note that selecting n as such means that the Byzantine quorum now becomes of size $2f + 2$ rather than $2f + 1$. Correct processes can gather such a quorum even if one correct process crashes itself and all f processes are Byzantine.

With this optimization, the probability that whole system shuts down takes a new value, which is expressed by:

$$Prob(sys_{shutdown}) = 1 - \sum_{k=0}^1 \binom{n-f}{k} (Prob(p_{shutdown}))^k (1 - Prob(p_{shutdown}))^{n-f-k}. \quad (2)$$

From Figure 4, a correct process crashes itself with a probability $< 10^{-6}$ in systems using more than 5 correct processes, $\mathbb{R} = 10$ rounds, and experiencing network omission rate of up to 60%. Accordingly, we perform a numerical analysis of the probability that the whole system shuts down as expressed in (1), and (2) and we show our results in Figure 6.

The graphs of Figure 6 highlight the increase in reliability between the two system design choices showing that a system implementing *RT-ByzCast* operates with a very high reliability (negligible probability of system shutdown).

Enhancing the reliability of the system further, by allowing it to tolerate more correct processes crashing themselves, is possible. However, doing so requires the system to have a mechanism of globally detecting aborted processes that crash themselves. We further elaborate on that in what follows.

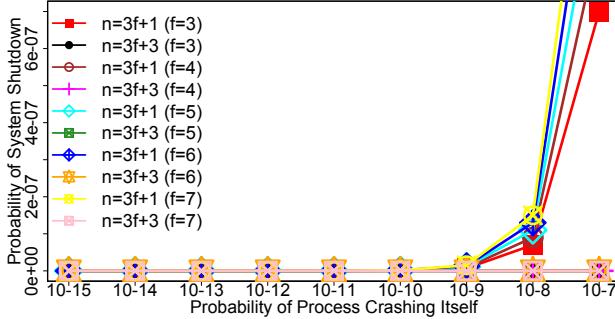


Figure 6. The probability of the whole system shutting down, when $\mathbb{R} = 10$, w.r.t. different values of the probability of a correct process crashing itself under 60% network loss rates (taken from Figure 4).

C. Tolerating Any Number of Processes Self-Crashing

In this section, we introduce a way that allows us to design, using *RT-ByzCast*, systems that can tolerate any number of correct processes crashing themselves. The solution is based on two main things: (i) over-provisioning and (ii) self-crash detection. To over-provision in system design, namely means to include additional redundant process replicas. In other words, to tolerate f Byzantine processes and rep simultaneous process self-crashes in the system ($\forall f, rep$) we choose the total number of processes, n , to satisfy $n = 3f + rep + 1$.

Processes then will execute our *RT-ByzCast* using Byzantine quorums depending on the value of n , which can only monotonically decrease (due to processes crashing themselves). Designing a system as such means that the system can tolerate up to rep correct processes crashing themselves simultaneously (for any value of rep).

However, in order to support system liveness for up to rep correct processes crashing themselves, the system would need to detect, within \mathbb{R} rounds after having crashed, every process that crashed itself. This detection allows other correct processes to update n and the needed quorums, without resulting in a cascade of self crashes leading to system shutdown. To this end, we describe in what follows a mechanism to detect crashed processes within \mathbb{R} rounds after crashing.

Detecting a Correct Process in the Dead-State

We recall that when a process p_i executes the proof-of-life function as well as when that process issues a broadcast, p_i needs to issue a message that should be seen by $2f + 1$ processes within a window of \mathbb{R} rounds; otherwise p_i would crash itself (case 4 of Section V-B). In this sense, having any set of $2f + 1$ processes not hearing from p_i in a window of \mathbb{R} rounds indicates that p_i has crashed itself (or is Byzantine). Note that a process p_j can hear from p_i either directly or via other processes.

The algorithm for detecting process self-crashes (and Byzantine processes mimicking a self-crash behavior) works as follows. Every process now includes in its message the round number during which it has sent the message. Every process p_i appends to every message it sends a list \mathcal{L}_{p_i} of size n , relative to the n processes in the system. \mathcal{L}_{p_i} contains, for every process $p_j \in \Pi$, the time-stamp relative to last time that p_j sent a message (based on p_j 's messages that p_i received). Note that a stored time-stamp relative to process p_j should have an associated proof of validity, that being the signature of

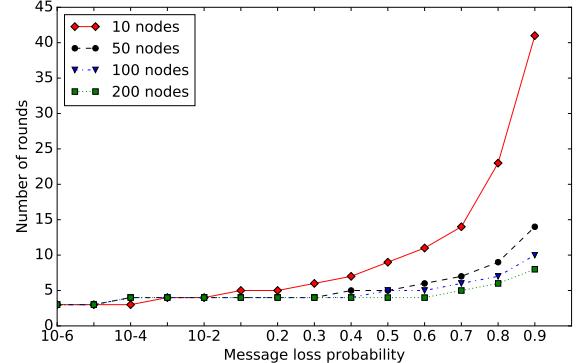


Figure 7. The size of the time window (\mathbb{R}) depending on the system size and the message loss probability.

p_j on a message holding that time-stamp (invalid time-stamps are ignored). Every process merges all the \mathcal{L} lists received such that the time-stamp with largest valid value for each process is kept. In every round $r : r > \mathbb{R}$, if a process p_i receives $2f + 1$ lists such that the time-stamp relative to some process p_j is less than the current round minus \mathbb{R} in all received \mathcal{L} , then p_i does the following: (1) claims p_j as crashed, (2) updates n to $n - 1$ and updates as well the new corresponding Byzantine quorum size, and (3) ignores all messages relative to p_j that may arrive at a later time.

D. Performance of *RT-ByzCast*

In addition to evaluating the reliability of *RT-ByzCast*, we evaluate how fast it can deliver messages. In Appendix E, we also show the incurred cost that *RT-ByzCast* has on the network bandwidth. *RT-ByzCast* has a latency of at most $3\mathbb{R}$ to deliver a message to all correct processes. Recall that we define \mathbb{R} as a slack period of collective re-transmissions of some message m sent by process p . The value of \mathbb{R} should be chosen such that p can know with high probability that at least $2f + 1$ processes received m .

We now simulate the value of \mathbb{R} (in communication rounds) as follows. We initially set $\mathbb{R} = 1$ round and run our reliability experiments for 2×10^5 times. We gradually increase the value of \mathbb{R} until no node aborts in all 2×10^5 repetitions. In our experiments, we assumed the worst case situation where all “ f ” nodes are Byzantine and can obstruct our algorithm by not forwarding messages’ signatures. Figure 7 represents the value of \mathbb{R} for 10, 50, 100 and 200 nodes, depending on the probability of message loss. First, as one may expect the empirical value of \mathbb{R} increases with the probability of message loss. Second, we observe that increasing the system’s size decreases the value of \mathbb{R} . This may appear unexpected, however, it is due to the fact that nodes forward messages in an all-to-all manner where messages can be probabilistically lost. Therefore, message losses have a lower impact on the broadcast delay in large scale systems. For example, with 90% message losses, \mathbb{R} is 41 (7) rounds with 10 (200) nodes resp.

To obtain the latency of *RT-ByzCast*, we empirically evaluate the worst case computational delay (d_{max}) needed per node per round. This delay includes the time a node needs to process all received messages and prepare the corresponding messages to be sent. Nodes use the ecdsa-256 encryption scheme [33], which is known for being fast and for having relatively short signatures. The overall worst case latency of

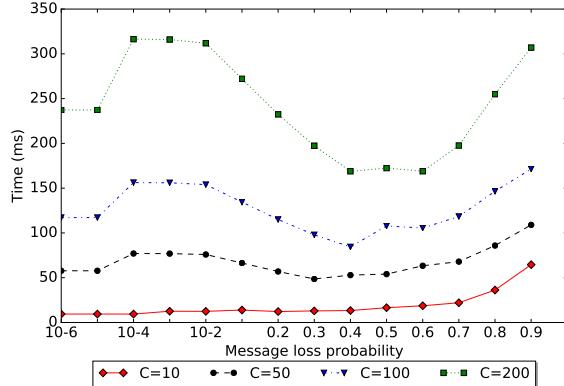


Figure 8. Total broadcast time depending on the system size and the message loss probability.

RT-ByzCast would then be computed as $3 \times \mathbb{R} \times d_{max}$. We illustrate this overall latency in Figure 8. We observed from our experiments that having more Byzantine nodes in the system puts a higher computing load on correct nodes. The variations in Figure 8 for a given system size result from two main factors: (i) the value of \mathbb{R} increases with message losses (Figure 7), and (ii) the value of d_{max} decreases with higher message losses, depending on the maximum number of simultaneously received messages per round. With low loss rates (from 10^{-6} to 10^{-4} in Figure 8), a slight increase in the value of \mathbb{R} increases the broadcast time noticeably. As message losses become more frequent (e.g., from 10^{-4} to 0.5) the value of d_{max} decreases, reducing the broadcast delay, until the value of \mathbb{R} increases sufficiently to compensate and increase again the broadcast time (e.g., from 0.5 to 0.9).

Our numbers in Figure 8 show that without any performance optimizations, and despite message losses, *RT-ByzCast* can meet the timing constraints of various applications, such as (i) power system automation and substation automation applications (IEC 61850-5 standard [34]), e.g., transfer of automation functions (TT3 class messages) ≈ 100 ms (systems size ≤ 50), slow speed auto-control functions, time-tagged alarms, event records, set-values read/write operations (TT4 class messages) ≤ 500 ms, (ii) continuous control (e.g. temperature-driven) applications ≤ 1 s [20], and (iii) operator commands of SCADA applications (≤ 2 s) [35].

However, we believe that *RT-ByzCast*'s latency has a substantial margin for improvement, whose avenues we discuss in the conclusions.

VII. REVIVING PROCESSES IN THE *Dead-State*

In order to allow processes to leave the *Dead-State*, we propose a modification to our *RT-ByzCast* algorithm. A process p_i that issues a broadcast at some round r sends now $\text{RTBRB-broadcast}(p_i, r, v; \Phi_{p_i})$ instead of $\text{RTBRB-broadcast}(p_i, v; \Phi_{p_i})$. In other words, p_i includes the round in which the broadcast is issued (first sent) in the payload of the $\text{RTBRB-broadcast}()$ message that p_i sends. The behavior of our *RT-ByzCast* algorithm of Section V-B remains the same.

A process in the *Dead-State* mimics the behavior of a crashed process in the *fail-stop* model [16]. However, unlike an actual crashed process, a process p_i in the *Dead-State* continues to listen and process received messages. In other

words, a process p_j that is in the *Dead-State* executes our *RT-ByzCast* algorithm, however with the exception that p_j refrains from sending any message. If process p_j (in the *Dead-State*) detects that some value v relative to process p_i has been signed by at least $2f + 1$ processes (i.e., $|\mathcal{R}_{echo}(p_i, r, v)| > 2f$), then p_j performs a check: if the current round, r_{cur} , satisfies $r_{cur} - 2\mathbb{R} > r$, where r is the round in which v was issued, then p_j delivers the message v (if it has not done so), shifts to the *Alive-State*. Then p_j transmits during every round in $[r_{cur} + 1, r_{cur} + 1 + \mathbb{R}]$ the message $\text{Deliver}((p_i, r, v, sigs); *)$ and resumes as in Section V-B.

VIII. THE CASE OF A DYNAMIC SYSTEM

So far in the paper, we have discussed systems where the maximum number of processes n is static and known to all processes. In this section, we consider a dynamic system where the maximum number of processes in the system is unknown, as processes can randomly join and leave. For consistency, as the system now allows processes to join at any point in time, we assume the following: a process that transitions to the *Alive-State* at round r is not required to deliver any messages that were broadcast at rounds $< r$.

We consider that correct processes now can exist, at any point in time, in one of three possible states: *Alive-State*, *Dead-State*, and *Pending-State*. Only those processes that exist in the *Alive-State* are the ones that should guarantee the RTBRB properties. We assume that initially $x > 3$ processes exist and are awakened by the clock initialization signal. We assume that these processes constitute a *trusted pool* and that their identity is known to any process that wishes to join the system later. Some processes within the trusted pool, at most $\lfloor \frac{x-1}{3} \rfloor$, can be Byzantine. Our trusted pool can be viewed as a trusted authority, which is a widely adopted assumption in Byzantine group communication protocols [36], [37]. In fact, in the absence of a centralized admission control scheme, a trusted entity becomes a necessity for allowing joins in an intrusion-tolerant system [38]. Processes in the trusted pool constitute the initial system and execute our *RT-ByzCast* algorithm.

Hence, the total number of processes, n , is initially equal to x (n however is not fixed). n is the number of processes that exist in either the *Alive-State* or the *Dead-State*. We consider that at any point in time at most $f = \lfloor \frac{n-1}{3} \rfloor$ can be Byzantine.

We assume that every process maintains a list \mathcal{ID} that contains the ids of the processes in the system. Initially the \mathcal{ID} list contains the ids of all processes in the trusted pool. A process p_j that wants to join the system at round r , initializes its state to the *Pending-State*, sets $n = x$, and defines \mathcal{ID} as the list of ids of all processes in trusted pool.

p_j begins to send at the beginning of every round in $[r + 1, r + 1 + \mathbb{R}]$ the heartbeat message, $\text{HB}(\{p_j, key_j, r\}; \Phi_{p_j})$, to the processes in the trusted pool. Upon the receipt of a $\text{HB}(\{p_j, key_j, r\}; \Phi_{p_j})$ in round $r' : r' \in [r + 1, r + 1 + \mathbb{R}]$ a process p_i in the trusted pool executes the following, if p_i is not sending some message for another potential joining process. p_i verifies the received heartbeat. Then p_i creates a $\text{Join}(\{p_j, key_j, r\}\{n, \mathcal{ID}\}; \Phi_{p_j}, \Phi_{p_i})$ message and sends this message in rounds $r'' : r'' \in [r' + 1, r + 1 + \mathbb{R}]$ it to all processes in the trusted pool, and to p_j as well.

If a process p_k in the trusted pool receives $\text{Join}(\{p_j, key_j, r\}\{n, \mathcal{ID}\}; \Phi_{p_j}, \dots, \Phi_{p_i})$ at some round $r' : r' \in [r + 1, r + 1 + \mathbb{R}]$, p_k verifies the message

received and validates that the values of n and \mathcal{ID} match with its local values. Then, considering that $\mathcal{SIG}(p_k)$, contains the signatures that p_k received so far relative to $\{p_j, key_j, r\} \{n, \mathcal{ID}\}$ (p_k 's signature is included in $\mathcal{SIG}(p_k)$), process p_k executes $\mathcal{SIG}(p_k) = \mathcal{SIG}(p_k) \cup \{\Phi_{p_j}, \dots, \Phi_{p_i}\}$. If $\mathcal{SIG}(p_k)$ contains signatures from at least $\lfloor \frac{x-1}{3} \rfloor + 1$ processes, then p_k sends in rounds $r'' : r'' \in [r'+1, r+1+\mathbb{R}]$ $\text{Join}(\{p_j, key_j, r\} \{n, \mathcal{ID}\}; \mathcal{SIG}(p_k))$ to all processes in the trusted group and to p_j as well.

The joining process p_j also executes the same steps when receiving $\text{Join}(\{p_j, key_j, r\} \{n, \mathcal{ID}\}; \Phi_{p_j}, \dots, \Phi_{p_i})$ messages. If at some round $r' \leq r + 1 + \mathbb{R}$, $\mathcal{SIG}(p_j)$ relative to some (n, \mathcal{ID}) contains signatures from at least $2\lfloor \frac{x-1}{3} \rfloor + 1$ processes in the trusted pool, p_j transitions to the *Alive-State*, updates its parameters accordingly (to be equivalent to the new values of $n + 1$ and $\mathcal{ID} \cup p_j$), and invokes an RTBRB-broadcast ($\text{join}(p_j, key_j); \Phi_{p_j}$). Otherwise, i.e., if p_j fails to collect enough signatures on some values of n and \mathcal{ID} by round $r + 1 + \mathbb{R}$, then p_j remains in the *Pending-State*, quits sending any messages, and tries to join the system again after a random duration elapses.

All processes that take part of the RTBRB-broadcast ($\text{join}(p_j, key_j); \Phi_{p_j}$) do not count p_j 's signature as part of any quorum (but simply use it for verifying the validity of the message sent). Any process that RTBRB-delivers ($\text{join}(p_j, key_j); *$), except p_j , updates the set of public keys to include key_j and the local parameters to $n = n + 1$ and $f = \lfloor \frac{n-1}{3} \rfloor$, $\mathcal{ID} \cup p_j$. That concludes the joining procedure for processes. A process p_l that wants to leave the system, issues an RTBRB-broadcast ($\text{leave}(p_l, key_l); \Phi_{p_l}$). p_l remains in the *Alive-State* until p_l delivers the message relative to its leave request broadcast, after which p_l shifts to the *Dead-State*. If it wishes, p_l can join the system later.

IX. RT-BYZCAST POTENTIAL APPLICATION DOMAINS

The problem we address — disseminating information in a real-time Byzantine-resilient manner given communication synchrony disruptions — is of fundamental importance to any critical distributed system seeking dependability and security in the CPS context. RTBRB is important to, e.g., (1) achieve a real-time common view of the state of the system by independent processes, (2) distribute sensory information in a coherent way or (3) help ensure the consistency of replicas driving a same end-component, e.g., an actuator.

The difficulty of the problem lies in ensuring predictability and resilience despite communication uncertainties, faults, and attacks that hinder the synchronism needed to meet real-time deadlines. In this sense, *RT-ByzCast* showcases an algorithmic design solution to circumvent the impossibilities that prevent traditional modular distributed computing approaches from achieving real-time Byzantine-resilient information dissemination. As a well-contained primitive, *RT-ByzCast* can be integrated in implementations of existing systems and frameworks, for example those related to distributed control systems (DCSs) [39] which include monitoring and control applications for factory automation, substation automation and smart grids. In fact, real-time execution platforms for DCSs, such as FASA [31], [40], have architectures that adhere to the assumptions considered in this paper; thus *RT-ByzCast* can be of substantial use to both the underlying system-monitoring

algorithms as well as to the control applications being run (e.g., power control applications [20]). Other application systems include ship-board DCSs requiring distributed real-time data services [41], [42], traffic control and agile manufacturing that require “fresh” data reflecting real-world status [41]–[44], and multi-gateway *SmartData* construct protocols for CPSs on Wireless Sensor Networks or on the Internet of Things [45].

X. RELATED WORK

The work in this paper has evolved from both [46] and [26], in what concerns the timing/synchrony aspect. In [46], it was assumed that all non-faulty processes remain connected synchronously, regardless of any process and network failures. This strong assumption about the network was too ideal, in terms of scale and timing behaviour. This resulted in a poor performance in practice (latency of ≈ 20 s) that limit [46]'s application areas. The limited performance is mainly attributed to deterministic network model (which in turn affected the solution design). Moreover, the system model in [46] did not allow processes that malfunction (violates assumptions) to know that they are being treated as faulty by the model. In contrast, *RT-ByzCast* provides latencies in the range of milliseconds (from few to few hundreds depending on system size and message loss). We achieve this performance gain by forcing processes to operate within the needed delays: we exclude processes that are incapable of meeting the desired timing requirements. Moreover, in our model, processes that violate timeliness assumptions transition to the *Dead State* and hence are aware of their “non-correctness” (in terms of time).

In [26], the timeliness problem was addressed by what the authors called weak-fail-silence: despite the capability of the transmission medium to deliver messages reliably and in real-time, the protocol should not be agnostic of potential timing or omission faults (even if sporadic). This notion was embedded in a bounded omission specification that weakens the basic fail-silence (crash) hypothesis. In our paper, we make a significant advance, by taking a further step: providing these reliable real-time communication guarantees, in environments with much higher uncertainty levels (faults and attacks). In fact, the bounded omissions assumption of [26] could not be taken as is, if we were to tolerate such higher and more uncertain fault sets (as we consider): it could easily lead to system unavailability in faulty periods.

The rest of the literature on broadcast primitives, to the best of our knowledge, either does not take into account timeliness and maliciousness or addresses them separately. We next summarize few such related works.

A. Byzantine Reliable Communication

In order to withstand unpredictable security threats and software unreliability arising within networks, various works have investigated how to provide a reliable broadcast capable of tolerating arbitrary process behavior. A celebrated algorithm in distributed computing is that of Bracha and Toueg [14], which implements reliable broadcast in an asynchronous system of n processes with at most $f < \frac{n}{3}$ Byzantine processes (proved to be an upper bound on the number of tolerable Byzantine processes [13], [14]). The broadcast algorithm of [14], similar to the one proposed in this paper, relies on echoing enough messages before delivering a value. Later works tried to further optimize this algorithm, e.g., in terms of the total

number of asynchronous communication rounds needed for termination [15]. On a different level Obenshain et al. [47] design and construct an intrusion-tolerant overlay capable of tolerating Byzantine actions, based on the key understanding that no overlay node should be trusted or given preference. The authors use a maximal topology with minimal weights to prevent routing attacks at the overlay level and rely on source routing augmented with redundant dissemination methods to limit the effect of compromised forwarder processes.

Unlike our *RT-ByzCast* algorithm, the solutions of [13]–[15], [47] do not provide timeliness guarantees on message delivery, even when applied to weakly synchronous networks such as the one we consider in this paper.

B. Timeliness Communication Guarantees

Given the need for predictable responsiveness in various applications today, many efforts have been made to devise timely broadcast (or routing/communication) algorithms, i.e., communication protocols with known fixed delays.

Timeliness on web communication, for example, has been addressed in content distribution networks [48], [49] such as Akamai [50]. The concept behind such distribution networks is to place data in a close geographical proximity from its consumers and hence minimize latency. In a slightly different context, timeliness has been also addressed at the level of data center networks [51], [52]. The proposed solutions in this area mainly advocate modifications of network devices. Bessani et al. propose Jiter [9], an application-layer routing, as a means to provide message latency and reliability assurances for control traffic in wide-area IP networks. Jiter routes deadline constrained control messages using an overlay network created on top of multi-homed communication infrastructure. Babay et al. [10] present an overlay transport service that can provide highly reliable communication while meeting stringent timeliness guarantees. Their scheme relies on an analysis of real-world network data, upon which they develop timely dissemination graphs and specify targeted redundant transmissions to timeliness and reliability.

Existing efforts have also investigated providing end-to-end guarantees in networks. Jacob et al. [2] propose an approach to integrate a wireless real-time communication protocol into CPS. Their approach decouples communication from application tasks. They devise a protocol based on dynamically establishing contracts between source/destination devices and the networking protocol. Another set of protocols known as resource reservation protocols (RSVP) combine flow specification, resource reservation, admission control, and packet scheduling to achieve end-to-end QoS [53]. Relying on software-defined networks, Kumar et al. [54] propose a framework that synthesizes network paths which can meet the requisite delay requirements for real-time flows.

On a different level Guerraoui et al. [20] devise an algorithm that allows processes to communicate in timely fashion via a real-time distributed shared memory (DSM). Using their DSM, write operations of a correct process become consistently visible to all alive processes within a known fixed delay. Their approach relies on bundling data messages with control traffic, namely traffic relative to the failure detector component used for monitoring processes in distributed control systems.

However, all above solutions [2], [9], [10], [20], [48]–[54], unlike *RT-ByzCast*, cannot handle malicious process behavior and hence would fail if any process is compromised.

XI. CONCLUSION

This paper studied how to realize the RTBRB abstraction: a real-time Byzantine-resilient reliable broadcast in the presence of unbounded communication losses and delays. We first showed that implementing RTBRB is challenging and in fact is impossible under traditional paradigms that consider failure detection mechanisms working independently of the distributed algorithms. To circumvent it, we proposed *RT-ByzCast*, an algorithm that deploys temporal and spatial diffusion of messages, and signature aggregation in a sliding time-window to mask losses and track process connectivity. *RT-ByzCast* bounds network timing uncertainties by reconciling two mechanisms: cooperative round-based message retransmission involving all processes; and proactive self-crashing of misbehaving processes.

We proved that *RT-ByzCast* indeed implements the desired RTBRB abstraction and we showed that it does so robustly and efficiently: we evaluated the reliability of our algorithm, showing that it can tolerate quite high loss rates ($\approx 60\%$) whilst still delivering the real-time service and ensuring a negligible probability that any correct process crashes itself, hence guaranteeing system survivability. We also showed that *RT-ByzCast* meets the timing constraints expected by a wide range of CPS and IoT applications, whose latency requirements are normally inversely proportional to the system size: RTBRB delays are in the order of the dozens of milliseconds on average, ranging from few milliseconds to hundreds of milliseconds, as system size goes from 10 to 200 nodes. Furthermore, *RT-ByzCast* design and minimal environment assumptions, simplify its integration in existing architectures.

By providing the first solution to the RTBRB problem in environments where synchrony constraints co-exist with maliciousness, the objective of this paper has been met: discovering and proving impossibility results countering common beliefs about failure detection, which will guide future research in these environments, as well as developing algorithms to circumvent them which, whilst non-optimised, already provide suitable performance for real world CPS/IoT applications.

However, we believe *RT-ByzCast*'s latency has a substantial margin for improvement, to meet more demanding timeliness needs, still in large scale system deployments. We plan to address these optimisations in future work, such as using gossip paradigms to decrease the nodes' transmission costs, and using multisignature/aggregate signature schemes to amortize the costs of signatures.

XII. ACKNOWLEDGMENTS

This work is in part supported by the Fonds National de la Recherche Luxembourg through PEARL grant FNR/P14/8149128.

REFERENCES

- [1] J. R. Moyne and D. M. Tilbury, “The emergence of industrial control networks for manufacturing control, diagnostics, and safety data,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 29–47, 2007.
- [2] R. Jacob, M. Zimmerling, P. Huang, J. Beutel, and L. Thiele, “End-to-end real-time guarantees in wireless cyber-physical systems,” in *RTSS*, 2016, pp. 167–178.
- [3] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, “Foundations of control and estimation over lossy networks,” *Proceedings of the IEEE*, vol. 95, no. 1, pp. 163–187, 2007.
- [4] D. Dzung, R. Guerraoui, D. Kozhaya, and Y.-A. Pignolet, “To transmit now or not to transmit now,” in *SRDS*, 2015, pp. 246–255.

- [5] DLC+VIT4IP, "D1.1 scenarios and requirements specification," Tech. Rep., 2010. [Online]. Available: <http://www.dlc-vit4ip.org/wb/media/Downloads/D1.1-V0.5-20100910-team.pdf>
- [6] C. Kim, M. Sun, S. Mohan, H. Yun, L. Sha, and T. F. Abdelzaher, "A framework for the safe interoperability of medical devices in the presence of network failures," in *ICCPs*, 2010, pp. 149–158.
- [7] M. M. Patel and A. Aggarwal, "Security attacks in wireless sensor networks: A survey," in *ISSP*, 2013, pp. 329–333.
- [8] F. Janurio, C. Carvalho, A. Cardoso, and P. Gil, "Security challenges in scada systems over wireless sensor and actuator networks," in *ICUMT*, 2016, pp. 363–368.
- [9] A. Bessani, N. F. Neves, P. Veríssimo, W. Dantas, A. Fonseca, R. Silva, P. Luz, and M. Correia, "Jiter: Just-in-time application-layer routing," *Computer Networks*, pp. 122 – 136, 2016.
- [10] A. Babay, E. Wagner, M. Dinitz, and Y. Amir, "Timely, reliable, and cost-effective internet transport service using dissemination graphs," in *ICDCS*, 2017, pp. 1–12.
- [11] D. F. Gordon, W. M. Spears, O. Sokolsky, and I. Lee, "Distributed spatial control, global monitoring and steering of mobile agents," in *IISA*, 1999, pp. 681–688.
- [12] A. Quagli, D. Fontanelli, L. Greco, L. Palopoli, and A. Bicchi, "Designing real-time embedded controllers using the anytime computing paradigm," in *ETFA*, 2009, pp. 1–8.
- [13] G. Bracha and S. Toueg, "Asynchronous consensus and broadcast protocols," *J. ACM*, vol. 32, no. 4, pp. 824–840, 1985.
- [14] G. Bracha, "Asynchronous byzantine agreement protocols," *Inf. Comput.*, vol. 75, no. 2, pp. 130–143, 1987.
- [15] D. Imbs and M. Raynal, "Simple and efficient reliable broadcast in the presence of byzantine processes," *ARXIV*, 2015. [Online]. Available: <http://arxiv.org/abs/1510.06882>
- [16] R. D. Schlichting and F. B. Schneider, "Fail-stop processors: An approach to designing fault-tolerant computing systems," *ACM Trans. Comput. Syst.*, vol. 1, no. 3, pp. 222–238, 1983.
- [17] D. Dolev, "The byzantine generals strike again," Stanford University, CA, USA, Tech. Rep., 1981.
- [18] C. Cachin, R. Guerraoui, and L. Rodrigues, *Introduction to Reliable and Secure Distributed Programming*. Springer-Verlag New York, 2011.
- [19] F. Abad, R. Mancuso, S. Bak, O. Dantsker, and M. Caccamo, "Reset-based recovery for real-time cyber-physical systems with temporal safety constraints," in *ETFA*, 2016, pp. 1–8.
- [20] R. Guerraoui, D. Kozhaya, and Y. A. Pignolet, "Right on time distributed shared memory," in *RTSS*, 2016, pp. 315–326.
- [21] C. Almeida and P. Verissimo, "Using light-weight groups to handle timing failures in quasi-synchronous systems," in *RTSS*, 1998.
- [22] P. Veríssimo and A. Casimiro, "The timely computing base model and architecture," *IEEE Trans. Comput.*, vol. 51, no. 8, pp. 916–930, 2002.
- [23] S. Viswanathan, R. Tan, and D. K. Y. Yau, "Exploiting power grid for accurate and secure clock synchronization in industrial iot," in *RTSS*, 2016, pp. 146–156.
- [24] T. Distler, C. Cachin, and R. Kapitz, "Resource-efficient byzantine fault tolerance," *IEEE Trans. Comput.*, vol. 65, no. 9, pp. 2807–2819, 2016.
- [25] H. Kopetz and G. Grunsteidl, "Ttp - a time-triggered protocol for fault-tolerant real-time systems," in *FTCS*, 1993, pp. 14–23.
- [26] P. Veríssimo, L. Rodrigues, and M. Baptista, "Amp: A highly parallel atomic multicast protocol," in *SIGCOMM*, 1989, pp. 83–93.
- [27] T. D. Chandra and S. Toueg, "Unreliable failure detectors for reliable distributed systems," *J. ACM*, vol. 43, no. 2, pp. 225–267, 1996.
- [28] L. Lamport, "The part-time parliament," *ACM Trans. Comput. Syst.*, vol. 16, no. 2, pp. 133–169, 1998.
- [29] T. Abdelzaher, A. Shaikh, F. Jahanian, and K. Shin, "Rtcast: lightweight multicast for real-time process groups," in *RTAS*, 1996, pp. 250–259.
- [30] R. Guerraoui, D. Kozhaya, M. Oriol, and Y.-A. Pignolet, "Who's on board? probabilistic membership for real-time distributed control systems," in *SRDS*, 2016, pp. 167–176.
- [31] M. Oriol, T. Garner, T. de Gooijer, M. Wahler, and E. Ferranti, "Fault-tolerant fault tolerance for component-based automation systems," in *ISARCS*, 2013, pp. 49–58.
- [32] R. Barbosa, A. Ferreira, and J. Karlsson, "Implementation of a flexible membership protocol on a real-time ethernet prototype," in *PRDC*, 2007, pp. 342–347.
- [33] D. Johnson, A. Menezes, and S. Vanstone, "The elliptic curve digital signature algorithm (ecdsa)," *International journal of information security*, vol. 1, no. 1, pp. 36–63, 2001.
- [34] IEC, "Iec 61850-5 international standard," in <https://www.sis.se/api/document/preview/565562/>, 2003.
- [35] A. Moga, T. Sivanthi, and C. Franke, "Os-level virtualization for industrial automation systems: Are we there yet?" in *SAC*, 2016.
- [36] M. Correia, L. C. Lung, N. F. Neves, and P. Veríssimo, "Efficient byzantine-resilient reliable multicast on a hybrid failure model," in *SRDS*, 2002, pp. 2–11.
- [37] M. K. Reiter, "Secure agreement protocols: Reliable and atomic group multicast in rampart," in *CCS*, 1994, pp. 68–80.
- [38] G. Danezis, C. Lesniewski-Laas, M. F. Kaashoek, and R. Anderson, "Sybil-resistant dht routing," in *ESORICS*, 2005, pp. 305–318.
- [39] B. Galloway and G. P. Hancke, "Introduction to industrial control networks," *IEEE Communications Surveys Tutorials*, vol. 15, no. 2, pp. 860–880, 2013.
- [40] M. Oriol, M. Wahler, R. Steiger, S. Stoeter, E. Vardar, H. Koziolek, and A. Kumar, "Fasa: A scalable software framework for distributed control systems," in *ISARCS*, 2012, pp. 51–60.
- [41] Y. Wei, S. H. Son, J. A. Stankovic, and K. D. Kang, "Qos management in replicated real-time databases," in *RTSS*, 2003, pp. 86–97.
- [42] P. Peddi and L. C. DiPippo, "A replication strategy for distributed real-time object-oriented databases," in *ISORC*, 2002, pp. 129–136.
- [43] A. Aslinger and S. H. Son, "Efficient replication control in distributed real-time databases," in *AICCSA*, 2005, pp. 34–38.
- [44] H. Zou and F. Jahanian, "A real-time primary-backup replication service," *IEEE Trans. Parallel Distrib. Syst.*, vol. 10, pp. 533–548, 1999.
- [45] D. Resner and A. A. Fröhlich, "Design rationale of a cross-layer, trustful space-time protocol for wsns," in *ETFA*, 2015, pp. 1–8.
- [46] F. Cristian, H. Aghili, H. R. Strong, and D. Dolev, "Atomic broadcast: From simple message diffusion to byzantine agreement," *Inf. Comput.*, vol. 118, no. 1, pp. 158–179, 1995.
- [47] D. Obenshain, T. Tantillo, A. Babay, J. Schultz, A. Newell, M. E. Hoque, Y. Amir, and C. Nita-Rotaru, "Practical intrusion-tolerant networks," in *ICDCS*, 2016, pp. 45–56.
- [48] J. Apostolopoulos, T. Wong, W. tian Tan, and S. Wee, "On multiple description streaming with content delivery networks," in *INFOCOM*, vol. 3, 2002, pp. 1736–1745.
- [49] G. Pallis and A. Vakali, "Insight and perspectives for content delivery networks," *Commun. ACM*, vol. 49, no. 1, pp. 101–106, 2006.
- [50] J. Dilley, B. Maggs, J. Parikh, H. Prokop, R. Sitaraman, and B. Weihl, "Globally distributed content delivery," *IEEE Internet Computing*, vol. 6, no. 5, pp. 50–58, 2002.
- [51] B. Vamanan, J. Hasan, and T. Vijaykumar, "Deadline-aware datacenter tcp (d2tcp)," in *SIGCOMM*, 2012, pp. 115–126.
- [52] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowstron, "Better never than late: Meeting deadlines in datacenter networks," in *SIGCOMM*, 2011, pp. 50–61.
- [53] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, "Rsvp: a new resource reservation protocol," *IEEE Network*, vol. 7, no. 5, pp. 8–18, 1993.
- [54] R. Kumar, M. Hasan, S. Padhy, K. Evchenko, L. Piramanayagam, S. Mohan, and R. B. Bobba, "Dependable end-to-end delay constraints for real-time systems using sdns," *arXiv*, 2017.
- [55] N. Nayyar, Y. Gai, and B. Krishnamachari, "On a restless multi-armed bandit problem with non-identical arms," *49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 369 – 376, sept. 2011.
- [56] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell Syst. Tech. J.*, vol. 42, no. 5, pp. 1977–1997, 1963.
- [57] H. S. Wang and N. Moayeri, "Finite-state markov channel-a useful model for radio communication channels," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 163–171, Feb. 1995.
- [58] Q. Zhang and S. Kassam, "Finite-state markov model for rayleigh fading channels," *Communications, IEEE Transactions on*, vol. 47, no. 11, pp. 1688–1692, Nov 1999.
- [59] L. Kanal and A. R. K. Sastry, "Models for channels with memory and their applications to error control," *Proceedings of the IEEE*, vol. 66, no. 7, pp. 724–744, July 1978.
- [60] H.-S. Wang and N. Moayeri, "Finite-state markov channel-a useful model for radio communication channels," *Vehicular Technology, IEEE Transactions on*, vol. 44, no. 1, pp. 163–171, Feb 1995.
- [61] D. Zhang and K. Wasserman, "Transmission schemes for time-varying wireless channels with partial state observations," in *INFOCOM 2002*, vol. 2, June 2002, pp. 467–476.
- [62] A. Laourine and L. Tong, "Betting on gilbert-elliot channels," *Wireless Communications, IEEE Transactions on*, vol. 9, no. 2, pp. 723–733, February 2010.
- [63] C. Tang and P. K. McKinley, "Modeling multicast packet losses in wireless lans," in *MSWIM*, 2003, pp. 130–133.
- [64] J. pierre Ebert, A. Willig, D. ing Adam Wolisz, and T. Berlin, "A gilbert-elliot bit error model and the efficient use in packet level simulation," Tech. Rep., 1999. [Online]. Available: <http://www2.tkn.tu-berlin.de/publications/papers/tknreport02.pdf>
- [65] D. Dzung and Y.-A. Pignolet, "Dynamic selection of wireless/powerline links using markov decision processes," *IEEE Conference on Smart Grid Communication (SmartGridComm)*, pp. 277–282, 2013.
- [66] O. Hohlfeld, R. Geib, and G. Hasslinger, "Packet loss in real-time services: Markovian models generating qoe impairments," in *IWQoS*, 2008, pp. 239–248.
- [67] G. Hasslinger and O. Hohlfeld, "The gilbert-elliot model for packet loss in real time services on the internet," in *MMB*, 2008, pp. 1–15.
- [68] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," *J. ACM*, vol. 58, no. 1, pp. 1–50, Dec. 2010.



David Kozhaya is a Scientist at ABB Corporate Research, Switzerland. He received his PhD degree in Computer Science in 2016, from EPFL, Switzerland, where he was granted a fellowship from doctoral school. His primary research interests include reliable distributed computing, real-time distributed systems, and fault- and intrusion-tolerant distributed algorithms. His past work experiences span across interdisciplinary domains ranging from research, teaching programming languages and computer literacy, financial and market analysis, and management of non-profit organizations.



Jérémie Decouchant is a Research Associate at SnT, University of Luxembourg. He received his Ph.D in Computer Science from the University of Grenoble-Alpes, France. Before that he obtained an engineering degree (MSc) from the Ensimag engineering school, in Grenoble. His research focused on the design and analysis of mechanisms to protect distributed collaborative systems against selfish or Byzantine behaviors. More recently, he has been designing distributed and privacy preserving genomic information processing workflows.



Paulo Esteves-Veríssimo is a Professor and FNR PEARL Chair at the University of Luxembourg (UL), and head of the CritiX research group at UL's SnT Centre (<http://wwwen.uni.lu/snt>). He is adjunct Professor of the ECE Dept., Carnegie Mellon University. Previously, he has been a Professor of the Univ. of Lisbon. Veríssimo is Fellow of IEEE and of ACM. He is Chair of the IFIP WG 10.4 on Dependable Computing and Fault-Tolerance and vice-Chair of the Steering Committee of the DSN conference. He is associate editor of the IEEE Transactions on Computers. He is interested in secure and dependable distributed architectures, middleware and algorithms for: resilience of large-scale systems and critical infrastructures, privacy and integrity of highly sensitive data, and adaptability and safety of real-time networked embedded systems. He is author of over 180 peer-refereed publications and co-author of 5 books.

APPENDIX A PROOF OF THEOREM 2

Intuitively, Theorem 1 follows from Theorem 2, so we prove in this section Theorem 2. We prove that it is impossible for an algorithm \mathcal{A} that can obtain \mathcal{L} (the list of non-suspected processes) to implement RTBRB-Validity and RTBRB-Timeliness, when \mathcal{A} cannot change which processes are in \mathcal{L} .

By contradiction assume that an algorithm \mathcal{A} implements RTBRB-Validity and RTBRB-Timeliness. The Δ of the RTBRB-Timeliness is a fixed duration; hence any process can send a finite number of messages within that Δ , e.g., M messages. We now compute the probability that a correct process $p_i \in \mathcal{L}$ loses all M messages sent to another process in \mathcal{L} . Recall that $P_{ij}(t)$ is the probability that link l_{ij} loses a message at time $t \forall i \neq j$. Let $P_{ij}(t \cap t')$ be the probability that link l_{ij} loses the messages (if any is sent) at time t and time t' . Since $0 < P_{ij}(t) \leq 1 \forall t$, then

$$0 < P_{ij}(t) = \frac{P_{ij}(t \cap t')}{P_{ij}(t'|t)} \leq 1 \forall t', t. \quad (3)$$

By (3), $P_{ij}(t'|t) > 0$ (and $0 < P_{ij}(t \cap t') \leq 1$). By induction, we have $P_{ij}(t'|t, t_1, \dots, t_x) > 0 \forall t' > t, t_x$. Denote by $B_{ij}(t)$ the event that link l_{ij} loses all messages (if any is sent) during the interval $t + \Delta$. Let t_x denote the times at which p_i sends a message in $[t + \Delta]$. The probability of $B_{ij}(t)$ happening is:

$$\begin{aligned} Pr(B_{ij}(t)) &= P_{ij}(t_1 \cap t_2 \cap \dots \cap t_M) \\ &= P_{ij}(t_1) \times P_{ij}(t_2|t_1) \times \dots \times P_{ij}(t_M|t_1, \dots, t_{M-1}) > 0. \end{aligned}$$

Given $0 < P_{ij}(t) \leq 1 \forall t$ and $P_{ij}(t'|t, t_1, \dots, t_x) > 0 \forall t' > t, t_x$, then we have $0 < Pr(B_{ij}(t)) \leq 1$; there is a positive probability that the link connecting p_i to p_j ($p_i, p_j \in \mathcal{L}$) loses all messages sent by p_i in any finite duration. Assuming independence between links, then there is a positive probability that all messages sent by p_i in any duration Δ and to any finite number of processes are lost. Assume that p_i is the process that invokes a broadcast. By RTBRB-Validity p_i delivers its own message. However, no other process in \mathcal{L} can deliver with probability 1, p_i 's message within any fixed duration after being broadcast.

APPENDIX B *RT-ByzCast* PROOF OF CORRECTNESS

We prove in what follows that our *RT-ByzCast* algorithm guarantees all RTBRB properties (defined in Section III).

A. RTBRB-Validity

Consider that a correct process p_i broadcasts message m at round r .

Lemma 1. At round $r + \mathbb{R}$, either: (i) p_i can verify that at least $2f + 1$ processes received its broadcast message, or (ii) p_i crashes itself.

Proof. p_i is a non-Byzantine process and hence it does not send different messages (relative to some broadcast instance) to different processes. By Case 1 of Definition 7, p_i crashes itself in any round $r_{curr} \geq r + \mathbb{R}$, if p_i has received $\leq 2f$ distinct signatures (including its own) on messages echoed from its broadcast. Therefore, at round $r + \mathbb{R}$, p_i either crashes

itself (hence p_i is no longer correct) or has at least $2f + 1$ process signatures. \square

Validity is a property that concerns a correct sender, hence the case when p_i does not crash itself. Following from Lemma 1, if p_i does not crash itself by round $r + \mathbb{R}$, then p_i can verify that at least $2f + 1$ processes received its broadcast message, in which case (lines 25-27 of Algorithm 2) p_i executes Function 4. Accordingly p_i delivers m and begins to send deliver messages as of round $r + \mathbb{R} + 1$ (lines 1-4 of Function 4).

Following from Case 2 of Definition 7, p_i remains correct (does not crash itself) at round $r + 2\mathbb{R} + 1$ if at least $2f + 1$ processes sign its deliver message. In other words, message m which is broadcast by a correct process p_i (that does not crash itself within $[r, r + 3\mathbb{R}]$) is eventually delivered by some correct process (in this case p_i itself), proving validity.

B. RTBRB-No Duplication

A process p_j in our *RT-ByzCast* algorithm delivers a message m broadcast by p_i , if and if only if p_j has not previously delivered any m relative to p_i . This is ensured by line 1 of Function 4.

C. RTBRB-Integrity

Assume that p_j is a correct process that delivers a message m relative to a correct process p_i . By lines 25-27 of Algorithm 2, this is possible only if p_j received on some $\text{Echo}((p_i, m) \dots)$ message signatures relative to at least $2f + 1$ processes.

This means at least $f + 1$ correct processes are transmitting some $\text{Echo}((p_i, m) \dots)$ messages. In our algorithm, a process p_k transmits an echo message with sender p_i and message m in two cases: (a) if p_k receives a RTBRB-broadcast $((p_i, m), \dots)$ message from p_i itself (lines 8-11 of algorithm 2), or (b) if p_k receives some $\text{Echo}((p_i, m) \dots)$ message signed by p_i (lines 13-24 of Algorithm 2).

In both cases (a) and (b), p_k is sure that p_i did indeed broadcast m .

D. RTBRB-Agreement

Lemma 2 (No two correct processes deliver different messages). *If a correct process p delivers message m and a correct process q delivers message m' , then $m = m'$.*

Proof. By lines (32-34 and 45-48 of Algorithm 2), any process that executes the *deliver-message()* function to deliver a message m of p_i should have m signed by at least $2f + 1$ processes. In our algorithm a correct process appends its signature to m (signs m) if it is not echoing any message other than m . In other words, a correct process never appends its signature to two messages m and m' such that $m \neq m'$. Any set of $2f + 1$ processes has at least one correct process in common. This means that no two correct processes deliver different messages m and m' (where $m \neq m'$) of p_i . \square

Lemma 3. *If some correct process p_j delivers message m then all correct processes eventually deliver some message m' .*

Proof. Let p_j be the only process that delivers message m , say at round r . From Function 4, p_j starts to send only $\text{Deliver}_{p_j}((m))$ messages from round $r + 1$ till round $r + 1 + 2\mathbb{R}$. At the end of round $r + 1 + \mathbb{R}$ process p_j crashes itself if it does not receive $\text{Deliver}_{\dots}((m))$ signed by at least $2f + 1$ processes (Case 2 of Definition 7). Since p_j is considered to be correct (hence does not crash itself), then this means that at least $2f + 1$ processes have received $\text{Deliver}_{\dots}((m))$ by round $r + 1 + \mathbb{R}$. By lines 51-60 of Algorithm 2 any correct process that receives $\text{Deliver}_{\dots}((m))$ delivers (has delivered) some message. this means that from round $r + 1 + \mathbb{R}$ till round $r + 1 + 2\mathbb{R}$ (included) at least $f + 1$ correct processes send only $\text{Deliver}_{p_j}(\dots)$ messages. Thus, all other correct processes can gather at most $2f$ signatures on any messages sent during $[r + 1 + \mathbb{R}, r + 1 + 2\mathbb{R}]$ (in which case they crash themselves by Case 4 of Definition 7), unless they receive some $\text{Deliver}_{p_j}((m))$. If they receive $\text{Deliver}_{p_j}((m))$ they deliver message m . Hence all correct processes (non-Byzantine processes that do not crash themselves) deliver some message. \square

The proof follows then from Lemma 2 and Lemma 3.

E. RTBRB-Timeliness

Consider that a non-Byzantine process p_i broadcasts message m at round r . and Assume that p_i is correct, that is p_i does not crash itself in $[r, r + \Delta]$.

Lemma 4. *If a correct process p_i broadcasts a message m at round r , then at least $f + 1$ correct processes receive m by round $r + \mathbb{R}$.*

Proof. Following from Lemma 1, at round $r + \mathbb{R}$, p_i can verify that at least $2f + 1$ processes have received its broadcast, otherwise p_i crashes itself. In other words, if p_i is still alive at round $r + \mathbb{R}$, then this means that at least $2f + 1$ processes received p_i 's broadcast message out of which at least $f + 1$ are correct. \square

Since process p_i at round $r + \mathbb{R}$ collected $2f + 1$ signatures, it delivers m as indicated by Function 4 (lines 30-35 of Algorithm 2).

Lemma 5. *If a correct process p_i broadcasts a message m at round r , then at least $f + 1$ correct processes deliver m by round $r + 2\mathbb{R}$.*

Proof. By lines (2-4) of Function 4 p_i sends only $\text{Deliver}(m)$ messages in all rounds $\in [r + 1 + \mathbb{R}, r + 1 + 3\mathbb{R}]$. For p_i not to crash itself (Case 2 of Definition 7) p_i needs to collect $2f + 1$ signatures on its $\text{Deliver}(m)$ message by round $r + 1 + 2\mathbb{R}$. This means if p_i is does not crash itself by end of round $r + 1 + 2\mathbb{R}$, then at least $f + 1$ correct processes have received and signed p_i 's $\text{Deliver}(m)$ message by round $r + 2\mathbb{R}$. By lines (51-59) of Algorithm 2 all such correct processes deliver m at lastest by round $r + 2\mathbb{R}$. \square

Following from Lemma 5, there exists at most f other correct processes at round $r + 2\mathbb{R}$ that may have not delivered m .

Lemma 6. *If a correct process p_i broadcasts a message m at round r , then all correct processes that have not delivered m by round $r + 2\mathbb{R}$ deliver m by round $r + 3\mathbb{R}$.*

Proof. Any process that delivers m at some round r' , sends only `Deliver(m)` messages in all rounds $\in [r', r' + 2\mathbb{R}]$ (lines 51-59 of Algorithm 2 and lines 2-4 of Function 4). Hence following from Lemma 1 and Lemma 5, in all rounds $[r + 1 + 2\mathbb{R}, r + 1 + 3\mathbb{R}]$, at least $f + 1$ correct processes only send `Deliver(m)` messages. Hence the rest of the correct processes (at most f) that have not delivered m yet would (i) crash themselves at $r + 1 + 3\mathbb{R}$ by Case 3 of Definition 7, or (ii) deliver m since they heard a message from some of the $f + 1$ correct processes that already delivered m . \square

As such, correct processes (that do not crash themselves) deliver message m from p_i by round $r + 3\mathbb{R}$. Hence, *RTBRB-Timeliness* is satisfied with $\Delta = 3\mathbb{R}$.

APPENDIX C CORRELATED LINK LOSSES

We now study the effect of correlated losses/omissions. To this end, we assume that links now may exhibit time-correlated losses/omissions, which might result in bursts.

We specifically simulate links following the Gilbert-Elliott (GE) model [55], [56]. The GE model, consisting of two states (see Fig. 9), is a simple non-trivial finite state Markov chain (FSMC) [57], established to capture well message loss behavior [58]–[60]. In fact, the GE model has been empirically verified, by a large body of work [56], [58], [61]–[64], as a good approximation of message losses in real-life communication scenarios. The GE model, for instance, has been used to model losses in wireless media IEEE 802.11 [64], wired power line networks [65] and other hybrid networks [66], [67]. The two states of the GE model (Fig. 9), noted by *good* and *bad*, can for example abstract the following: the communication link between a pair of processes occupies the bad state when the packet success-rate drops below a certain “unacceptable” threshold or when communication delays become slower than expected (e.g., $> d$); the link occupies the good state otherwise. At any round, the link can be in exactly one of the two states: the good state or the bad state. A link in the good state delivers messages (if any is sent) in a reliably and timely fashion; however, if the link is in the bad state then messages are either lost or delayed ($> d$). In other words messages sent on a link, which is in the bad state, are omitted. Every link *transitions* across rounds, i.e., at the beginning of every round a link moves to its new state, which can be the same state it existed in or the other state.

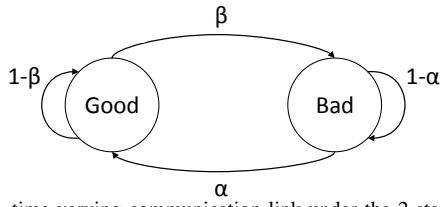


Figure 9. A time-varying communication link under the 2-state GE model

For example, given the link is in the good state at some round, it will remain in the same state at the next round with probability $1 - \beta$ and will move to the bad state with probability β . Similarly if the link state is bad at some round, it will remain bad at the next round with probability $(1 - \alpha)$ and will shift to good with probability α . Since links transition according to the transition probabilities (i.e., α and

β), different links can exist in different states in a given round, even if all links start from the same state and are governed by the same α and β .

We run our simulations again given the same $|\mathcal{C}| \in \{5, 10, 20, 30, 40, 50, 100, 200\}$ processes but now considering the various transition probabilities $(\alpha, \beta) \in \{(0.8, 10^{-6}); (0.7, 10^{-5}); (0.6, 10^{-3}); (0.5, 10^{-2}); (0.4, 0.1); (0.3, 0.4); (0.2, 0.6); (0.1, 0.7)\}$. Namely, we assume that all links have the same transition probabilities and start from the good state. Every link then transitions to a new state at the beginning of every round. We select values of α and β satisfying the positive correlation condition (a.k.a. bursty condition) [68], which requires that $(1 - \beta) > \alpha$. In fact, we show in Figure 10 the probability of having bursts of various lengths under our selected values of transition probabilities.

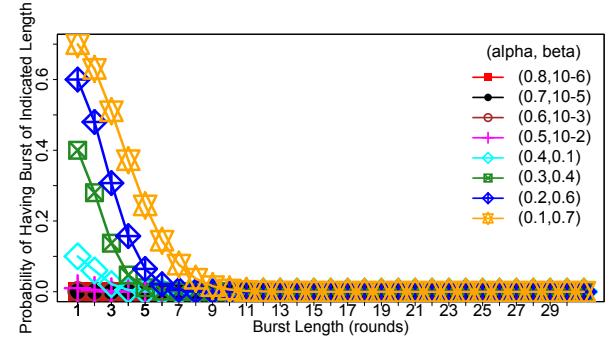


Figure 10. The probability of bursts of variable lengths under different transition probabilities (α, β).

For a given value of $|\mathcal{C}|$ and (α, β) , we invoke a broadcast at one of the processes and record, after \mathbb{R} rounds of communication, if any process does not receive $|\mathcal{C}|$ signatures on the value being broadcast. We repeat such an instance 10^6 times.

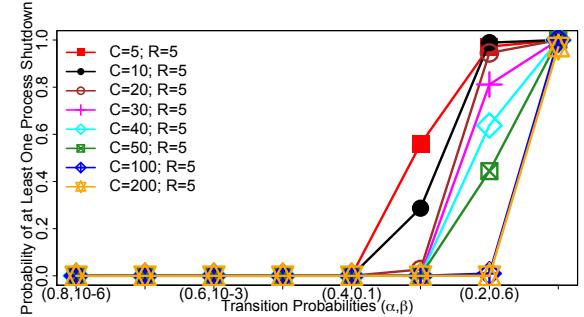


Figure 11. The probability of a correct process crashing itself in a system after 10 communication rounds ($\mathbb{R} = 5$) under correlated link losses.

We report our results showing:

$$\frac{\text{num. of instances in which some correct process crashes itself}}{10^6},$$

for $\mathbb{R} \in \{5, 6, 10, 15\}$ rounds respectively. We select these values of \mathbb{R} to show that, despite loss correlation, a well chosen value of the slack window still allows our *RT-ByzCast* algorithm to provide the intended reliability.

Our results in Figure 11, Figure 12, Figure 13, and Figure 14 show similar conclusions as the case of independent message loss. Namely, the probability of a correct process crashing itself decreases (1) as the number of non-Byzantine processes in the system increases, and (2) as the number of communication rounds (i.e., the value of \mathbb{R}) increases.

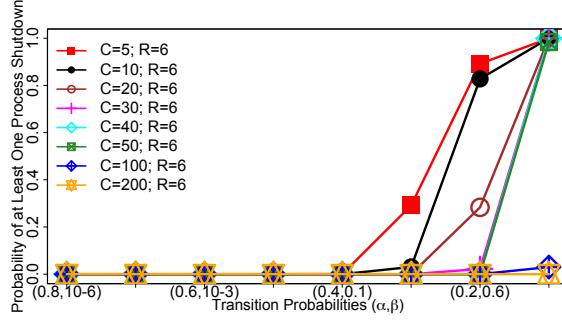


Figure 12. The probability of a correct process crashing itself in a system after 10 communication rounds ($\mathbb{R} = 6$) under correlated link losses.

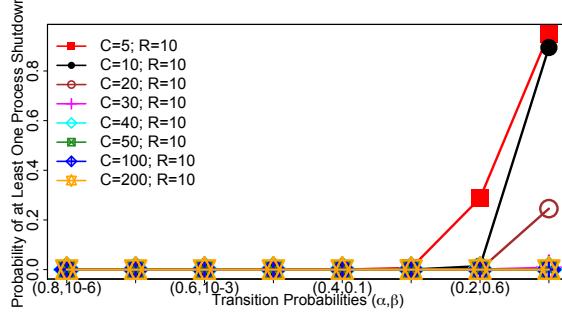


Figure 13. The probability of a correct process crashing itself in a system after 10 communication rounds ($\mathbb{R} = 10$) under correlated link losses.

APPENDIX D RELIABILITY EVALUATION

In this section we illustrate the probability of a correct process shutting down. Unlike the graphs in Section VI.A, we fix $|C|$ and vary the window size (\mathbb{R}). Our results in Figure 15, Figure 16, Figure 17, and Figure 18 show that increasing the number of communication rounds (i.e., the value of \mathbb{R}) increases the reliability of our *RT-ByzCast* algorithm for any number of correct processes. In fact, with $\mathbb{R} = 10$ the probability of a process crashing itself becomes negligible even with up to 60% losses/omissions rate.

APPENDIX E *RT-ByzCast* BANDWIDTH CONSUMPTION

Finally, we evaluated the protocol’s peak bandwidth consumption, both in terms of reception and in emission. Figure 19 and Figure 20 respectively present the peak bandwidth consumptions during the broadcast of 128 bits and 1 Mbits messages, depending on the message loss probability. We

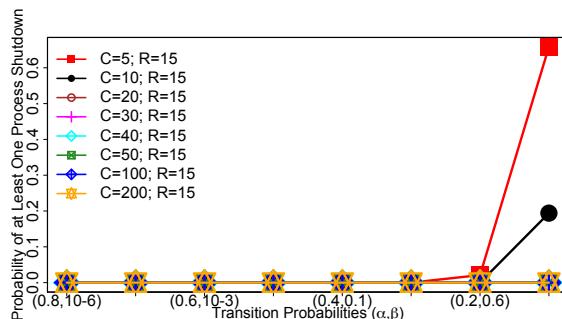


Figure 14. The probability of a correct process crashing itself in a system after 15 communication rounds ($\mathbb{R} = 15$) under correlated link losses.

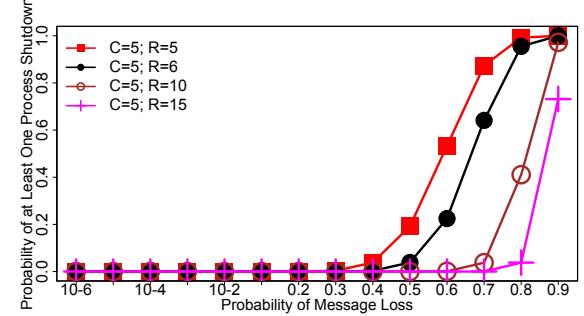


Figure 15. The probability of a correct process crashing itself in a system of $|C| = 5$ w.r.t. a varying number communication rounds.

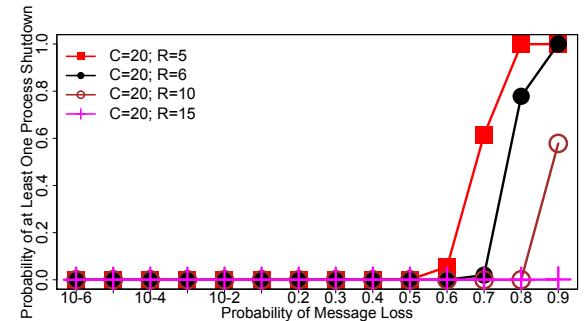


Figure 16. The probability of a correct process crashing itself in a system of $|C| = 20$ w.r.t. a varying number communication rounds.

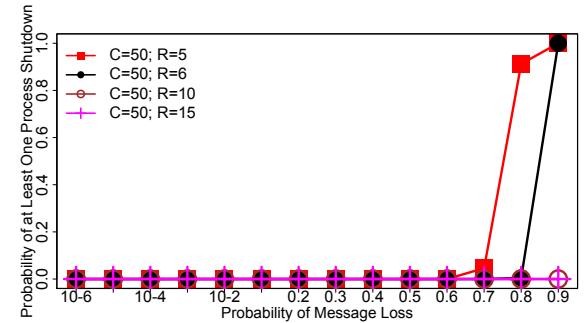


Figure 17. The probability of a correct process crashing itself in a system of $|C| = 50$ w.r.t. a varying number communication rounds..

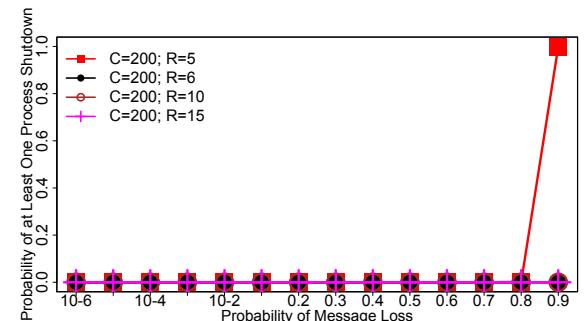


Figure 18. The probability of a correct process crashing itself in a system of $|C| = 200$ w.r.t. a varying number communication rounds..

observed that correct nodes are using more bandwidth when there are Byzantine nodes in the system. We therefore present the bandwidth consumptions when the system contains the maximum number of Byzantine nodes tolerated. We use plain lines to represent the nodes' bandwidth usage in reception, and dashed lines for their bandwidth usage in emission. For a given system size, the difference between those two lines is due to message losses, and is accentuated by the presence of Byzantine nodes.

From these figures, we could make two observations. First, as expected, the nodes' bandwidth consumption increases with the system's size and with the message's size. Second, the bandwidth consumptions of nodes tend to decrease when the message loss probabilities increase. For example, Figure 19 shows that broadcasting a 128 bits message in a system of 100 nodes consume around 1.9 Mbits in emission, and less than 1.1 Mbits in reception. Comparatively, broadcasting a larger 1 Mbits message in a system of 100 nodes consume less than 100 Mbits in emission, and less than 75 Mbits in reception. Those numbers were obtained without any optimizations, and we leave for future works the improvement of *RTByzCast*'s bandwidth consumption. Possible ideas include reducing the nodes' fanout, decreasing the bandwidth's footprint of digital signatures, and the avoidance of redundant messages.

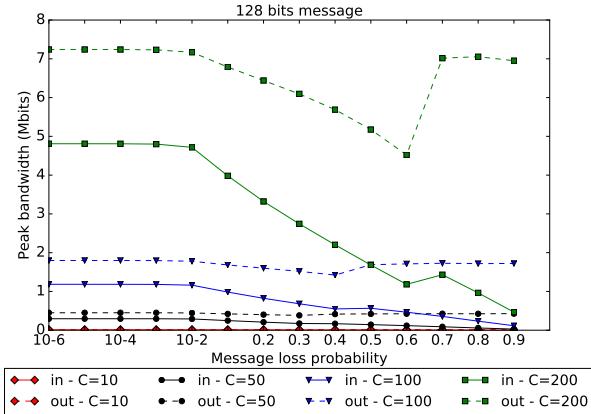


Figure 19. Total broadcast delay for a 128 bits-message depending on the number of nodes and the probability of message losses.

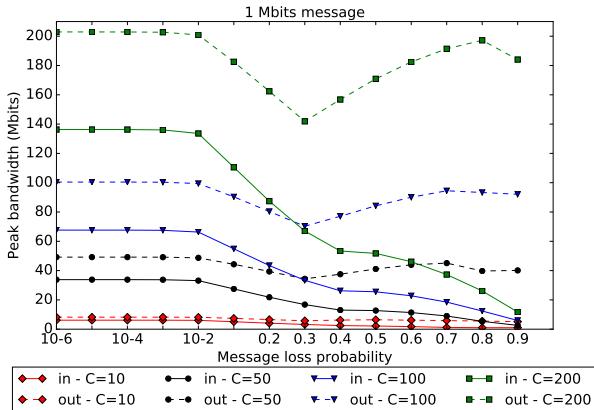


Figure 20. Total broadcast delay for a 1 Mbits-message depending on the number of nodes and the probability of message losses.