A scalable causal broadcast that tolerates dynamics of mobile networks
Daniel Wilhelm, Luciana Arantes, Pierre Sens

To cite this version:

HAL Id: hal-02652082
https://hal-upec-upem.archives-ouvertes.fr/hal-02652082v3
Submitted on 9 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A scalable causal broadcast that tolerates dynamics of mobile networks

Daniel Wilhelm  
Sorbonne University, CNRS  
Inria, LIP6, Paris, France  
daniel.wilhelm@lip6.fr

Luciana Arantes  
Sorbonne University, CNRS  
Inria, LIP6, Paris, France  
luciana.arantes@lip6.fr

Pierre Sens  
Sorbonne University, CNRS  
Inria, LIP6, Paris, France  
pierre.sens@lip6.fr

Abstract—Many distributed applications and protocols require causal broadcast. Various existing algorithms ensure causal order of broadcast messages, but they are either not scalable, or do not take into account the characteristics of mobile networks, such as nodes mobility, message losses, or limited capacity of nodes. This paper proposes a causal broadcast algorithm suitable for mobile networks since it copes with the dynamics, constraints, and specifications of such networks. Control information included in each message, and maintained on each node, is of small size and the algorithm handles message losses. Performance evaluation of experiments conducted on Omnet++ confirms the effectiveness of our causal broadcast protocol.

I. INTRODUCTION

Causal Broadcast is a fundamental group communication service used by many distributed applications, such as distributed databases, publisher/subscribe systems, collaborative applications, or distributed social networks. It ensures that messages are delivered to all nodes (processes) only once, preserving causal relation of broadcast messages, i.e., the delivery of broadcast messages must respect Lamport’s happened-before relationship [9]: if the broadcast of a message m precedes the broadcast of a message m′, then every process that delivers these two messages must deliver m before m′.

In this paper, we are particularly interested in providing a causal broadcast service for wireless mobile networks [8], composed of mobile nodes and reliable support stations. The dynamics of such networks where mobile nodes can move, leave/join the system, and fail, poses new challenges for the implementation of the group communication service. For instance, if a mobile node joins an executing distributed application where other nodes have already delivered and broadcasted some messages, the new node should not be blocked, waiting for these messages, if it will never receive them. Furthermore, the protocol must deal with message losses, the low memory capacity of mobile nodes, and, depending on the system, a high number of mobile nodes.

Many approaches have been proposed in the literature that guarantees and implements a message causal order. The two most well-known ones are (1) the piggybacking of causal per node information in each message, such as logical vector clocks [6][10], and (2) flooding through FIFO links [7], where messages are systematically forwarded at first reception. The dissemination pattern through FIFO links ensures that there exists no path between two nodes over which messages are sent out of causal order. The first approach is not suitable for tackling the dynamics and scalability issues of mobile networks, because the size of causal information depends on the number of nodes of the system. Therefore, we have chosen the second approach to implement our causal broadcast protocol. However, the latter, which was proposed by Friedman et al. [7], only offers causal order over static distributed systems where network topology does not change.

The authors in [13] have extended Friedman et al.’s broadcast protocol to dynamic systems, using data structures that do scale. On the other hand, their solution to cope with system dynamics does not address the issue of free mobile nodes movement since the network overlay must always be connected through links previously initialized by a particular handoff procedure. Therefore, a path of initialized links must always exist between each pair of nodes. Moreover, links are all supposed to be FIFO, reliable, and initialized in both directions. These characteristics are not realistic for mobile networks, which make [13] not suitable for such networks.

Our causal broadcast algorithm is designed for mobile networks, taking into account their intrinsic characteristics and constraints. Mobile nodes can join/leave the system, move, and temporarily fail. They are connected to support stations through a wireless network, which is neither reliable nor FIFO. Our algorithm renders them FIFO and reliable by applying message retransmission, sequence number assignment, and message reception acknowledgment. On the other hand, since support stations are connected by a wired network, existing protocols, such as TCP, ensure reliable FIFO communication among them. It is worth emphasizing that messages piggyback few control information, and memory usage complexity is low for mobile nodes, while, for support stations, it grows linearly with the number of local connected mobile nodes. Hence, our broadcast requires less control information than vector clocks and does not make the constraining assumptions of the flooding approach [7][13][12]. Performance evaluation results of experiments conducted over the simulator OMNeT++/INET [18] confirm the advantages of our solution.

The rest of the paper is organized as follows. Section II gives some background on causal broadcast. Section III presents the system model. In Section IV, we describe our proposed causal broadcast protocol. Section V presents evaluation of results on OMNeT++. Section VI discusses related work and, finally, Section VII concludes the paper.
II. BACKGROUND

Mobile Networks are usually composed of a huge number of nodes, which render not sustainable full system membership knowledge by nodes. Instead, they have just a local partial view of the system, which usually contains much fewer nodes than the whole system, and only communicate with the nodes, denoted neighbors, that belong to this partial view. Messages are, therefore, disseminated transitively through an overlay network built with the local view of nodes: nodes send received messages to their respective neighbors, which, in their turn, also forward them.

In this work, we are interested in providing a group communication service which, besides the primitives for joining and leaving the system (Join() and Leave() respectively), offers to the application the primitives CoBroadcast(m), that broadcasts the message m to all nodes, and CoDeliver(m), that delivers m to the application, respecting the causal order of messages. Causal order ensures that sent messages are delivered while respecting the causal relation between them, based on the happened before relation [9] introduced by Leslie Lamport. (see Definition 1 below). Therefore, the delivery of received messages might be delayed until they respect causal order. We thus distinguish the reception of a message from its delivery. Note that due to re-transmissions a node might receive multiple times the same message, but the latter is delivered only once.

**Definition 1** (Happened before). The happened before relation, denoted \( \rightarrow \), partially orders events in a distributed system. Considering two events \( e_1 \) and \( e_2 \), \( e_1 \rightarrow e_2 \) iff: (a) \( e_1 \) and \( e_2 \) occurs on the same process and \( e_1 \) precedes \( e_2 \) or (b) for a message \( m \), \( e_1 = \text{send}(m) \) and \( e_2 = \text{deliver}(m) \) or (c) there exists an event \( e_3 \) such that \( e_1 \rightarrow e_3 \) and \( e_3 \rightarrow e_2 \) (transitivity).

Following (b) and (c) of Definition 1, causal order between two messages is formally defined as: \( \forall m, \text{send}(m) \rightarrow \text{send}(m') \Rightarrow \text{deliver}(m) \rightarrow \text{deliver}(m') \). By extending the above definition to broadcast and deliver of messages, we have: \( \forall m, \text{broadcast}(m) \rightarrow \text{broadcast}(m') \Rightarrow \text{deliver}(m) \rightarrow \text{deliver}(m') \).

We consider a dynamic system in which nodes can join/leave the system during execution. Furthermore, messages delivered by all nodes are discarded, and nodes that join the system will, therefore, never receive these messages. For this reason, we apply the following definition of causal broadcast in our work [11]:

**Definition 2** (Causal Broadcast). \( \forall \) messages \( m_1, m_2 \), 

\[ \text{broadcast}(m_1) \rightarrow \text{broadcast}(m_2) \Rightarrow \text{deliver}(m_2) \rightarrow \text{deliver}(m_1) \]

With such a definition, if \( \text{broadcast}(m_1) \rightarrow \text{broadcast}(m_2) \) and if \( m_1 \) is not available anymore in the system, then \( m_2 \) can still be delivered without blocking forever waiting for \( m_1 \). However, \( m_1 \) is never delivered after \( m_2 \).

For implementing causal order of broadcast messages, our algorithm exploits the principle of message forwarding over reliable FIFO links, as proposed in [7]. The scenario of Figure 1 explains such an approach. It consists of three nodes connected by reliable FIFO links. First, node A broadcasts \( m \) (a). Once B received \( m \), B delivers and sends it to C (b). Then B broadcasts \( m' \) (c). Finally, C delivers \( m \) and sends it to A. (d) shows that \( m' \) cannot be received before \( m \) by any node.

III. MODEL

We consider a mobile network composed of Mobiles Hosts (\( h \) nodes) and Mobile Support Stations (\( s \) nodes). Nodes communicate exclusively through message passing. Group communication primitives are called by applications running on \( h \) nodes while \( s \) nodes deal with message loss and guarantee that messages reach their destination. Every \( h \) and \( s \) node is uniquely identified by an \( id \).

All \( s \) nodes are reliable and static, i.e., they do not move, join, or leave the system, neither fail. They are connected by a high speed wired network, whose links are reliable and FIFO, and over which we build a static logical — tree-based — overlay network. They communicate with each other exclusively through this overlay by using the TCP protocol. Every \( s \) node antenna has the same fixed transmission range, which defines its respective cell to which \( h \) nodes, close to it, connect themselves. Furthermore, \( s \) nodes hold most of the consistency and causal order information of the protocol since they have much more memory and computing power than \( h \) nodes, and no energy limitation.

On the other hand, \( h \) nodes can move, join, or leave the system, and are subject to temporary failures. The latter happens when a node crashes and then re-joins the system, recovering its last saved state. \( h \) nodes communicate with the \( s \) nodes of their respective cells through a wireless network where interferences can lead to message losses, but not message corruption. A \( s \) node acts as a relay, forwarding the broadcasted messages of the \( h \) nodes of its cell. Wireless links are not supposed reliable, nor FIFO. Note that a \( h \) node may be temporarily disconnected from the system if no cell covers its position. Furthermore, an \( h \) node can be within the transmission range of two \( s \) nodes simultaneously, but it is connected to at most one \( s \) node at a given moment, which is generally the closest one. Finally, unlike \( s \) nodes, \( h \) nodes have computing and energy limitations, thus maintaining only a small data structure.
We should point out that cells must overlap in order to ensure the covering of the whole area of the system, as shown in Figure 2.

Figure 2 shows an example of a network topology. Cells are represented by dashed circles. It is worth remarking that node \( m \) is within no cell, and that node \( a \) is within both \( s_6 \)’s and \( s_7 \)’s cells.

The same network of Figure 2 is represented in Figure 3 but \( s \) and \( h \) nodes are logically organized in a tree-based overlay. The wired and wireless networks are respectively represented by solid and dashed lines. Some wired links from Figure 2 have been removed, and \( h \) nodes are connected to at most one \( s \) node (e.g., \( a \)). \( h \) nodes are leaves of the tree since they only communicate with their respective \( s \) node. The latter can also be a leaf, provided that no \( h \) node is connected to it. Note that \( m \) that is within no cell, is temporarily disconnected from the tree.

IV. CAUSAL BROADCAST ALGORITHM

Our causal broadcast algorithm, presented in Algorithm 1 to 4, consists of three parts: the dissemination of application messages, the handling of join/leave, and mobility operations of \( h \) nodes.

Each line is preceded by a symbol (*, #, or +), corresponding to the part of the algorithm to which the line is related. Lines preceded by * are those related to the dissemination of application messages, those preceded by # are those related to the join and leave of \( h \) nodes, and those preceded by + are those related to the mobility of \( h \) nodes.

Algorithm 1 presents the tasks executed by \( h \) nodes. Algorithm 2 to 4 present the tasks executed by \( s \) nodes. Algorithm 2 handles the reception of messages sent by \( h \) nodes, Algorithm 3 the periodical sending of ack messages by \( s \) nodes, and Algorithm 4 handles the reception of messages sent by \( s \) nodes.

An application running on a \( h \) node can call the following four functions provided by the algorithm:

- **Join()**: whenever the \( h \) node wants to join the system.
- **Leave()**: whenever the \( h \) node wants to leave the system.
- **CoBroadcast(m)**: for broadcasting the message \( m \).
- **CoDeliver()**: delivering a message, if available.

A. Data structures and message types

Specific structures are kept by \( h \) and \( s \) nodes to guarantee the causal delivery of application messages.

First, \( h \) nodes maintain variables to identify messages, to manage application messages, and to manage the connection with their cell’s \( s \) node.

A \( h \) node piggybacks its id, \( id_h \), on sent messages to identify them. It also checks that received messages come from its cell by comparing their attached id to its cell’s one, \( id_c \), since \( h \) nodes may be in reach of several cells.

Moreover, a \( h \) node maintains some variables to manage application messages: two sequence number counters, \( seq_h \) and \( seq_C \), the first to stamp new broadcasted application messages, the latter to contain the sequence number of the next application message to deliver. Additionally, two buffers, \( SBuffer \) and \( RBuffer \), the first stores unacknowledged sent application messages, the second contains received application messages until they are FIFO ordered.

Finally, a \( h \) node uses some variables to manage its connection: Two session number counters are used to identify connections, \( Ses \) and \( Ses\_LC \), the first contains the current session number, the latter the session number of the latest session in which the connection was established, i.e., where the node received a reply from the \( s \) node to which it tried to connect, confirming the reception of the connection request. A variable \( state \) is also used to identify the node’s current connection state. A node can be in four states: \( \text{init} \), \( \text{join} \), \( \text{conn} \), \( \text{estab} \).

Secondly, \( s \) nodes also maintain variables to manage their messages, as well as a structure for each connected \( h \) node \( h_i \).

A \( s \) node maintains its cell’s id, \( id_c \), a sequence number counter for new broadcasted messages \( seq_C \), and a buffer which stores unacknowledged broadcasted application messages, denoted \( SBuffer \).

On the one hand, the structure associated with \( h_i \) contains some variables to manage the connection. The structure associated to \( h_i \) is identified with \( h_i \)’s id, \( id_h \). Two sequence
number counters are also used, \(seq_h\) and \(seq_{ACK}\), the first contains sequence number of the next message of \(h_i\) to re-broadcast, the second the sequence number of the most recent application message \(h_i\) acknowledged. The session number of the connection with \(h_i\) is stored in \(Ses\). A buffer, \(RBuffer\), contains received application messages of \(h_i\) until they are FIFO ordered.

On the other hand, some variables of the structure are only used during the handoff process. \(H_{lock}\) locks the structure if a handoff is in progress, to ensure that handoffs concerning \(h_i\) are done sequentially. \(seq_{CO}\) saves the state of \(SBuffer\). \(m_{nd}\) contains messages discarded by the \(s\) node which \(h_i\) did not delivered at the previous \(s\) node to which it was connected. \(CoRequest\) stores the most recent pending \(Req_1\) request message received during the current handoff.

Messages are divided into three groups: the first handles the dissemination and acknowledgment of application messages, the second the join/leave of \(h\) nodes, and the third the mobility of \(h\) nodes.

The first group contains:

- Application messages from node type A to B:
  - \(<App_{a,s}, data, idh, seq_h, seq_i, idc, M_d>\)
  - \(<App_{s,s}, data, idh, seq_h, seq_i>\) and \(<App_{s,h}, data, seq_h, idc, M_d>\)

- Acknowledge messages of \(h\) nodes \(<ack_{h, idh, seq_C, Ses}>\) and of \(s\) nodes \(<ack_{C, idc, vSeq}>\).

The second group contains:

- \(<join, idh, Ses>\) sent by \(h\) nodes to connect to the system.
- \(<init_{ACK}, idh, seq_h, seq_{C, Ses}>\): sent by \(s\) nodes to conclude the connection phase.
- \(<leave, idh, Ses>\) sent by \(h\) nodes to leave the system.
- \(<leave_{ACK}, idh, Ses>\) acknowledge the reception of \(leave\).
- \(<Delete, idh, Ses>\) sent to \(s\) nodes to delete the \(h\) node \(id_h\) if registered.

The third group contains:

- \(<init, id_h, seq_{C, Ses}, Ses_{LC}>\) sent by \(h\) nodes to change cell.
- Messages exchanged between \(s\) nodes during handoffs to ensure causal order for moving \(h\) nodes:
  - \(<Req_1, id_h, seq_{C, Ses}, Ses_{LC}>\), \(<Rsp_1, id_h, seq_{C, M_d}, Ses_{LC}>\), \(<Req_2, id_h, msg_{req}, Ses>, <Rsp_2, id_h, msg, msg_{rev}, Ses>\)
  - \(<AppC, data, idc, seq_c, idh, Ses>\) application messages sent during the connection phase to a specific \(h\) node.

We define several functions to make the algorithm more easily readable. Messages are sent with the \(broadcast(type,...)\) function, whose behavior and arguments change in function of the message type.

\(h\) nodes use the \(broadcast\) function to sent messages (\(App_{h,s}, ack, join, leave, init\)) on the wireless network to their cell’s \(s\) node.

When used by \(s\) nodes, the behaviour of the \(broadcast\) function changes according to the message type:

- \(App_{s,h},\ ac_{C, init_{ACK}}, App_{C, leave}\) and \(leave_{ACK}\) messages are sent on the wireless network.
- \(App_{s, Req}\) messages are forwarded on the wireless network and to the neighbor \(s\) nodes, except the one which sent them.
- \(Req_1, Rsp_1, Req_2, Rsp_2\) messages are forwarded if the \(s\) node is not the destination of them. In this case, they are forwarded to the neighbor \(s\) nodes, except the one which sent them.
- \(Delete\) messages are forwarded to the neighbor \(s\) nodes, except the one which sent them.

Moreover, we define some other functions. \(chooseC()\) returns the position of the nearest \(s\) node, and \(minSeq()\) returns the sequence number of the oldest message of \(SBuffer\) (or \(seq_{C}\) if \(SBuffer\) is empty).

### B. Dissemination of application messages

Similarly to [7] and [13], the dissemination mechanism is based on flooding over an overlay network. Nodes are logically organized in a tree, like the one of Figure 3.

A \(h\) node calls \(CoBroadcast(m)\) (Algo1.11-13) in order to broadcast an application message \(m\). All \(h\) nodes of the system should deliver \(m\), respecting causal order of messages. On the other hand, \(s\) nodes are responsible for the dissemination of application messages (Algo2.44-47). A \(s\) node re-broadcasts to the \(h\) nodes of its cell every application message it receives from a \(h\) node of its cell (Algo2.2-8). It also sends the message to its \(s\) node neighbors of the overlay through the wired network. An application message received by a \(s\) node, sent by a second \(s\) node, is forwarded in the same way, except that it is not sent back to the sender (Algo4.54).

A \(h\) node includes in every application message it broadcasts both its id, \(id_h\), and the id of the current cell to which it is connected, \(idc\). Since wireless links are neither FIFO nor reliable, our protocol needs to detect out of order messages as well as losses. To this end, a \(h\) node associates a sequence number value \(seq_h\) to every new message it broadcasts, by keeping a local sequence number counter variable which is incremented at every new broadcast. A \(s\) node also has its own sequence number counter variable, \(seq_{C}\), used to timestamp every new message it re-broadcasts. It increments the counter at every re-broadcast of a different message and controls, for each connected \(h\) node \(h_i\), which is the \(seq_{C}\) of the last message that \(h_i\) has delivered.

Both \(h\) and \(s\) nodes maintain two types of local buffers: (1) \(RBuffer\), which stores received application messages until they are FIFO ordered (2) \(SBuffer\) which keeps pending messages sent over the wireless network, i.e., those messages that have not been acknowledged yet by all the receiver(s) \(h\) nodes.

A \(s\) node keeps one \(RBuffer\) per connected \(h\) node. It uses the \(seq_h\) value included in application messages sent by a given \(h\) node, \(h_i\), to detect out of order message receptions: if a message \(m\) with \(seq_{hm}\) sent by \(h_i\) is received by \(s\) but the latter has not received yet all messages from \(h_{i}\) whose \(seq_{h}\) value is smaller than \(seq_{hm}\), then \(m\) is inserted in the \(RBuffer\) that \(s\) associates to \(h_i\) (Algo2.7). Similarly, \(h_i\) uses the sequence number values (\(seq_{C}\)) of the messages received from \(s\) to order them, i.e., temporally keeping those which are out of order messages in its \(RBuffer\) (Algo1.18). \(s\) nodes (resp. \(h\) nodes)
Algorithm 1: Tasks of $h_i$

Join

1# seq$_h$=seqC, $\text{Ses}_i$=Ses$_{LC_i}$=0
2# SBuffer$_i$=RBuffer$_i$=
3# id$_C_i$=chooseC()
4# state$_i$=join
5# broadcast(<join.id$_h$,id$_{S_i}$>)

Upon changing cell

6+ id$_C_i$=chooseC()
7+ state$_i$=(state$_i$==join ? join : init)
8+ Ses$_i$++
9+ stop(ackTimeout)
10+ broadcast(<state$_i$,id$_h$,seq$_C_i$,id$_C_i$,id$_{S_i}$,Ses$_{LC_i}$>)

Upon calling CoBroadcast(m)

11* msg=<App$_S$.data,id$_h$,seq$_C_i$++>
12* broadcast(msg)
13* SBuffer$_i$.insert(msg)

Leave

14# broadcast(<leave,id$_h$,Ses$_i$>)

upon reception of m=<type,...> from a s node

15* switch (m)
16* case <App$_S$.data,id$_h$,seq$_C_i$.M$_d$> :  
17* if seq$_i$==seq$_C_i$ then
18* RBuffer$_i$.insert(m)
19* else if seq$_C_i$=seq$_C_i$ then
20* seq$_C_i$++
21* if id$_h$ $\notin$ M$_d$ then
22* deliver(data)
23* FIFODeliver()
24* case <App$_C$.data,id$_h$,seq$_C_i$.id$_{D_i}$,Ses$_i$>:
25* if id$_h$==id$_h$ $\land$ Ses$_i$==Ses$_i$ then
26+ if state$_i$==init then
27+ seq$_C_i$=0 ; Ses$_{LC_i}$=Ses$_i$ ; state$_i$=connecting
28+ if seq$_i$==seq$_C_i$ then
29+ RBuffer$_i$.insert(m)
30+ else if seq$_C_i$==seq$_C_i$ then
31+ seq$_C_i$++
32+ deliver(data)
33+ FIFODeliver()
34* case <ackC,vSeq>:
35* SBuffer$_i$ = \{ $m' \in$ SBuffer$_i$.m $\land$ seq$_i$<vSeq[id$_i$]\}
36* seq$_C_i$=Seq[id$_i$]
37* case <initACK,id$_h$,seq$_h$,seq$_C_i$.Ses$_i$>:
38+ if id$_h$==id$_h$ $\land$ state$_i$!=estab $\land$ Ses$_i$==Ses$_i$ then
39* if !ackTimeout then
40* StartACKtimeout()
41* if initTimeout then
42* stop initTimeout
43* seq$_C_i$=seq$_C_i$ $\land$ Ses$_{LC_i}$=Ses$_i$ ; state$_i$=estab
44+ clear(SBuffer$_i$.seq$_h$
45+ FIFODeliver()
46* case <leaveACK.id$_h$>:
47+ if id$_h$==id$_h$ then
48+ leave()

upon expiration of M’s timeout

49* if M==ack$_h$ then
50* broadcast(<ack$_h$.id$_h$,seq$_C_i$.Ses$_i$>)
51* else
52* broadcast(<M>)
53* setTimeout(M,calcTimeout())

Algorithm 2: upon reception of m=<type,...> from $h_i$ at $S_j$

1* switch (m)
2* case <App$_h$.s.data,id$_h$,seq$_h$>:
3* if seq$_h$==seq$_h$ then
4* Disseminate(data,id$_h$,seq$_h$)
5* seq$_h$++
6* FIFODisseminate()
7* else if seq$_h$ < seq$_h$ then
8* RBuffer$_{j_i}$.insert(m)
9* case <ack$_h$.id$_h$,seq$_C$.Ses$_i$>:
10* if h[id$_h$] $\land$ Ses$_i$==Ses$_i$ then
11* seq$_C_i$=seq$_C$
12* clear(SBuffer$_i$.seq$_C_i$)
13+ if seq$_C_i$<minSeq() then
14+ clear(m$_{nd_i}$)
15+ if m$_{nd_i}$ $\neq$ then
16+ seq$_C_i$=calcSeq(id$_h$)
17+ broadcast(<initACK.id$_h$,seq$_C_i$.Ses$_{S_i}$>)
18* case <join.id$_h$.Ses$_i$>/<init.id$_h$.seq$_C_i$.Ses$_{LC_i}$>:
19+ if id$_h$ registered then
20+ !lock$_j$ then
21+ if Ses$_{S_i}$ $\neq$ then
22+ clear(m$_{nd_i}$,seq$_C_i$
23+ seq$_C_i$=seq$_C$
24+ Ses$_i$=Ses$_i$
25+ broadcast(<initACK.id$_h$,seq$_C_i$.Ses$_{S_i}$,Ses>)
26+ case <initACK.id$_h$,seq$_C_i$.Ses$_{S_i}$>:
27+ if m$_{nd_i}$ $\neq$ then
28+ seq$_C_i$=calcSeq(id$_h$,seq$_C_i$)
29+ if m$_{nd_i}$ $\neq$ then
30+ broadcast(<initACK.id$_h$,seq$_C_i$.Ses$_{S_i}$>)
31* else
32+ h={id$_h$.0,minSeq(),Ses.$\emptyset$,false,0,$\emptyset$}
33* if type $==$ join then
34* broadcast(<initACK.id$_h$.minSeq(),Ses>)
35* broadcast(<Delete.id$_h$.Ses>)
36* else
37+ Hlock$_j$=true
38+ broadcast(<Req1.id$_h$.seq$_C_i$.Ses$_{LC_i}$>)
39* case <leave.id$_h$.Ses>: 
40+ if h[id$_h$] then
41+ delete[h[id$_h$]]
42+ broadcast(<Delete.id$_h$.Ses>)
43+ broadcast(<leaveACK.id$_h$>)

Function: Disseminate(data,id$_h$,seq$_h$)

44* broadcast(<App$_S$.h.data,seq$_C_i$.id$_h$.0>)
45* SBuffer$_i$.insert(<App$_S$.h.data,seq$_C_i$.id$_h$.0,0.id$_h$.seq$_h$>)
46* broadcast(<App$_S$.data,id$_h$,seq$_h$>)
47* seq$_C_i$++
disseminate (resp. deliver) them in ascending order (Algo2.6) (resp. Algo1.23) in ascending order of seqh (resp. seqC).

We denote pending messages that have not been acknowledged by all receiver(s). Regarding SBuffer, a message broadcasted by a h node is considered pending by this h node till it receives an acknowledge from its respective cell’s s node (Algo1.34-36), while a message re-broadcasted by a s node remains pending till it receives acknowledgments (ack messages) from all connected h nodes of the cell (Algo2.9-12). In both cases, as soon as a message is not pending anymore, it is removed from the SBuffer. On the other hand, every pending message is periodically retransmitted within a time interval whose duration, recalculated at each retransmission, depends on the number of pending messages in the SBuffer (Algo1.49-53 & Algo3).

A s (resp., h) node regularly sends an ack message (timeout mechanism), confirming the reception (Algo3.48-51) (resp., delivery (Algo1.49-50)) of those application messages whose sequence number value is smaller or equal to the one included in the ack message in question. Note that the ack messages of s nodes contain a vector with an entry for each connected h node of the s node’s cell. A final remark is that a h node delivers a message it has broadcasted only after receiving this same message from its cell’s s node (Algo1.17).

Figure 4 shows the broadcast of two messages: h1 broadcasts m1 and h2 broadcasts m2 after delivering m1 (broadcast(m1) → broadcast(m2)). The notation of the messages also includes their sequence number. Pending messages in SBuffer (bold) and non FIFO ordered ones in Rbuffer (italic) are also shown. h1 is connected to s1 and h2 to s2 while s1 and s2 are neighbors.

Upon reception of m1, s1 sends it to s2 and also broadcasts it within its cell, which contains h1. When receiving m1, s2 broadcasts it in its cells. Note that s2 does not send m1 back to s1. Node h2 receives and delivers m1. Then, h2 broadcasts m2, which is disseminated like m1. Remark that h2 delivers m2 only after receiving m2 from s2, confirming the reception of m2. At expiration of a timeout, s2 (resp., h2) sends an ack message to h2 (resp., s2) with seq=1 (resp., seq=2) to acknowledge m2 (resp., m1 and m2). h2 and s2 then stop sending m2 and clear their respective SBuffer.

Node s1 re-broadcasts m1 within its cell after receiving it, but it is lost. Thus, at the next timeout expiration, h1 re-broadcasts m1, because s1 did not acknowledged it. However, s1 has received m1 and, therefore, ignores m1’s second reception. Upon receiving m2 from s2, s1 broadcasts it within

---

**Algorithm 3:** upon expiration of M’s timeout at s nodes

48* if M==ack, then
49* vSeq={seq, ∀ connected h1}
50* broadcast<(ack, vSeq)>
51* setTimer(ackTimeout, calcAckTimeout())
52* else
53* broadcast(<M>)

**Algorithm 4:** upon reception of m from s nodes at sj

52* switch (m)
53* case <Apps_s.m, idh, seqh, Ses>:
54* Disseminate(m, idh, seqh)
55* case <Delete, idh, Ses>:
56* if h[idh] then
57* if Ses>Sesj then
58+ if CoRequestj then
59+ broadcast<CoRequestj>
60+ delete(h[idh])
61+ broadcast(<Delete, idh, Ses>)
62* default:
63+ Handoff(m)

**Function:** Handoff(m)

64* if m.idh not registered then
65* broadcast(m)
66* if m==<Req1, idh, seqC, SesC, Ses> then
67* if Sesj>Ses ∧ Hlockj then
68* if CoRequestj then Sesj then
69* CoRequestj=m
70* else if Sesj>Ses then
71* broadcast(<Delete, idh, Ses>)
72* return
73* if SesC == Sesj then
74* mnd={idm, ∀m∈SBuffer ∪ mndj ∩ idm ∈ Mm ∧ seqm = seqC}
75* else
76* mnd={idm, ∀m∈SBuffer ∪ mndj ∧ idm /∈ Mm}
77* broadcast(<Rsp1, idh, seqhj, mnd, Ses>)
78* seqCojj = seqC
79* else if m>Sesj then
80* broadcast(m)
81* else
82* if m==<Req2, idh, msgseq, Ses> then
83* seqhj = seqh
84* seqCoj = seqC
85* mndj = mnd
86* msgseq = {m∈mnd /∈ SBufferj, m′∈msgseq}
87* mrev = {m′∈SBufferj, m′>seqCoj}
88* delete(h[idh])
89* clear(SBufferj)
90* broadcast(<Rsp2, idh, msg, mrev, Ses>)
91* else if m==<Rsp2, idh, msg, msgseq, Ses> then
92* ∀m′∈SBufferj \{msgseq ∪ mndj}, m′>seqCoj
93* BroadcastCo(msg)
94* Hlockj = false
95* if CoRequestj then
96* Receive(CoRequestj)
97* else if msg==[] then
98* seqCoj = calcSeq(idh)
99* broadcast(<initACK, idh, seqCj, Sesj>)

its cell. On the other hand, for \( h_1 \), \( m_2 \) does not respect FIFO order, because it awaits a message with seq_{C1} = 1 and \( m_2 \) has seq_{C1} = 2. Hence, \( h_1 \) stores \( m_2 \) in its RBuffer. At expiration of the timeout related to \( m_1 \), \( s_1 \) broadcasts it again, and, upon reception, \( h_1 \) delivers the two messages in FIFO order and remove them from its RBuffer. Finally, \( h_1 \) and \( s_1 \) send ack messages at their next timeout expiration, which are both received. Hence, they remove both messages from their respecting SBuffer. Note that all copies of \( m_1 \) and \( m_2 \) have been deleted from all node buffers.

C. Join/leave the system

An extra control is necessary to identify the connection in which \( h \) nodes are, because each \( s \) node has its own local sequence number (seq) (a same message may receive a different seq from one \( s \) node than from another \( s \) node), messages may be lost on the wireless network (nodes cannot determine if the other node has received the connection messages), and \( h \) nodes can move and change cells (several connections of the same \( h \) node might be processed simultaneously).

To solve those problems, we have introduced the concept of session sequence numbers that uniquely identify a wireless network connection between a \( h \) node and a cell’s \( s \) node. For this purpose, every \( h \) node keeps the following variables whose values are also included in some messages of the protocol, whenever necessary:

- \( \text{Ses} \) identifies \( h_i \)’s current connection. It is incremented when \( h_i \) changes its current cell.
- \( \text{Ses}_{LC} \) identifies the last established connection, i.e. the last session in which the \( h \) node received a message from its cell’s \( s \) node acknowledging the connection request.

The Join() primitive (Algo1.1-5), called by \( h_i \), chooses a cell \( c_j \) and sends to it a join message which includes the id and Ses of \( h_i \). In its turn, if \( h_i \) is not already connected to it (Algo2.32-36), node \( s_j \) associates a structure to \( h_i \) to control the connection. Otherwise (Algo2.30), \( h_i \)’s session number is updated (Algo2.25) in order to take into account that \( h_i \) might have tried to connect to another \( s \) node without succeeding.

In both cases, \( s_j \) sends an initACK message to \( h_i \) in order to finish the handoff by giving to \( h_i \) the sequence number of the oldest stored message in its SBuffer (determined by minSeq()). Moreover, a Delete message, including the session number piggybacked by the join message, is disseminated to the other \( s \) nodes (Algo4.55-61), so that they delete the structure they have associated with \( h_i \), if they store one whose associated session number is lower than the one of the Delete message, i.e., if Delete concerns a more recent session/connection. We should emphasize that a \( h \) node can join the system at any moment, but it will not deliver those messages which were discarded by the \( s \) node to which it connects before its connection.

\( h_i \) can leave the system at any time by calling the Leave() primitive (Algo1.14) which will re-broadcasts a leave message until it is received by some \( s \) node. The latter forwards the leave message to the other \( s \) nodes (Algo2.39-43). \( h_i \) leaves the system once it received an acknowledge to its Leave request (Algo1.46-48). This message exchange is necessary since \( s \) nodes store messages until all \( h \) nodes of their cell have acknowledged them. Hence, the \( s \) node(s) where \( h_i \) is registered would never discard messages if \( h_i \) would leave without sending a leave message.

D. Handoff procedure

The handoff procedure ensures the causal order delivery of application messages when \( h \) nodes move between cells after having joined the system. We denote \( s_p \) and \( s_q \) the previous and the new cell \( s \) node of \( h_i \) respectively. Basically, the handoff procedure consists of a set of messages exchanged between the moving node \( h_i \) and \( s_m \), as well as between the latter and \( s_p \), as shown in Figure 5. The handoff procedure must cope with the two following constraints.

a) Single delivery: \( s \) nodes do not necessarily assign sequence numbers in the same order to those application messages which are not causally related, i.e., which have been concurrently broadcasted. For example, let’s suppose that \( s_p \) receives \( m \) then \( m' \) while \( s_q \) receives \( m' \) then \( m \) and that \( h_i \) delivers \( m \) when connected to \( s_p \), moves, connects to \( s_q \), and then delivers \( m' \). Without any extra control, \( h_i \) would deliver \( m \) again, since \( s_q \) has ordered \( m \) after \( m' \). In order to avoid these multiple deliveries, a \( s \) node assigns a small set, denoted \( M_d \), to every application message it broadcasts. The \( M_d \) of a message contains the ids of all \( h \) nodes of the cell that have already delivered the message at another cell. By exchanging messages with \( s_p \), \( s_q \) acquires knowledge about which are the pending messages of its SBuffer which \( h_i \) have already delivered, including then \( h_i \)’s id to the \( M_d \) of each of those messages (Algo4.88). \( h_i \) will not deliver those messages whose \( M_d \) set contains its id (Algo1.22). For every one of those messages, it just updates the related sequence number (seq_{C1}). Note that the size of \( M_d \) is quite small as, at a given moment, few \( h \) nodes are changing cells.

On the other hand, a \( s \) node discards a message \( m \), i.e., removes it from its SBuffer, once all connected \( h \) nodes have acknowledged \( m \) (Algo2.13-14). This message deletion procedure renders more difficult the comparison of the SBuffer of \( s_q \) and \( s_p \) to find which messages \( h_i \) has not delivered, since
the SBuffer of one s node may contain messages removed from the other one, and messages received by one may not have been received yet by the other. Such a conflict is handled by message exchanges over the wired network connecting the s nodes (Handoff function of Algorithm 4).

b) Session consistency: Messages may be lost on the wireless network. In this case, nodes cannot determine if their handoff messages are received nor to which connection received messages belong. This uncertainty is handled by including the h node’s session number Ses in every handoff message, identifying, therefore, the connection in which it is sent. Moreover, a h node is unable to know which s node holds its latest connection information, because its connection requests may be lost. In order to tackle this problem, handoff messages are propagated over the wired network to all s nodes. A third remark is that due to its movement, a h node may try to connect to different s nodes in a short time interval, starting, therefore, several handoff procedures simultaneously. It happens, for instance, if h1 tries to connect to a second s node just before connecting to sn. Hence, during the handoff procedure with h1, sn may receive old connection requests from h1. The s nodes manage concurrent requests sequentially in increasing order of Ses (Algo4.56).

Handoff principle: h1 starts the handoff procedure (Algo1.6-10) when it moves to sn’s cell, by sending to sn an init message containing the sequence number (seqC) of the latest message it delivered, as well as the connection’s session number Ses and the session number of the last established connection SesLC. Note that h1 stops the sending of ack messages until its new s node gave him a new sequence number seqC, since its actual seqC is associated to h1’s previous s node sp’s cell.

sn processes h1’s init message (Algo2.18-38) differently following if it has already registered h1.

If h1 is already registered (Algo4.19), then the request is processed if sn is not currently in a handoff procedure for h1 (Algo4.20), to ensure that the handoffs are done sequentially. If the request’s session number Ses is greater than the stored session number Sesij, then it is the first time sn receives a connection message for this session, and the structure is thus updated. The session number associated to h1 is updated, so that messages related to h1’s previous connection attempts to other s nodes, i.e., those with lower Ses values, will be discarded. Moreover, if the last established connection of h1 was with sn itself (Algo4.22), it considers the seqC value of the init message as an acknowledge (ack message) because, in this case, h1 may have received and delivered messages during that SesLC connection without acknowledging them. mndj is updated in order to remove the messages h1 might have acknowledged, as well as to update the remaining messages in mndj, so that their associated sequence number order begins with 0, because h1 will await ascending ordered application messages with a sequence number beginning with 0. The sequence number of h1 is also updated if mndj is empty, to take into account the messages init might acknowledged. Finally, if mndj is empty, then the handoff is finished and sn sends to h1 an initACK message.

On the other hand, if h1 is not already connected to sn (Algo4.31-38), then the latter associates a structure to h1, locks the structure, and disseminates a handoff request Req1 message including seqC and SesLC to its s node neighbors in the tree overlay that will forward it to their neighbors and so on.

As previously explained, only one handoff procedure per node is executed at a given time, following the Ses value of Req1, even if h1 has tried to connect to several s nodes in a short time interval. Req1 messages related to more recent connections of h1, and received during the execution of another handoff procedure, are handled only after the latter ends (Algo4.67-69&Algo4.99-100). sn discards every Req1 message concerning an older connection than the current one and propagates a Delete message containing the request’s Ses value (Algo4.72), so that the s node which sent the Req1 in question deletes the structure it associated with h1. Before deleting the structure associated to h1, a s node checks if it has a pending Req1 request (Algo4.59), and broadcasts it, so that no Req1 message is lost.

If no handoff procedure is currently in progress and Req1’s Ses value is higher than the one sp stores, then sp replies to Req1 with Rsp1 which contains the list of id’s of the messages which h1 has not delivered. Moreover, if the last established connection (SesLC) of h1 was with sp, then sp considers Req1’s sequence number as an acknowledgment (Algo4.75). Finally, sp saves the state of its SBuffer by saving its seqC, which identifies the sequence number of the next broadcasted message, identifying therefore also the maximum sequence number of currently broadcasted messages.

Several s nodes may associate a structure to h1 simultaneously, since h1 might tried to connect to several s nodes in a short time interval. Hence, to ensure that the Rsp1,Req2 and Rsp2 handoff messages are only processed by sn and sp, they are only taken into account of by s nodes if their associated Ses session number is equal to the session number Sesij to which the stored structure is associated (Algo4.80). Otherwise, the messages are broadcasted to the neighbor s nodes to propagate them in the overlay network.

When receiving Rsp1, sn sends a Req2 message to sp, asking for messages it deleted among those of Req1’s list (Algo4.87). Moreover, sn stores the state of h1 as viewed by sp (seqh1,seqCj,1,mndj).

sp replies to the Req2 message with a Rsp2 message that includes two lists: (1) the list of requested messages that sn has discarded (Algo4.90); (2) the list of messages that sp has received since the sending of (Rsp1) (Algo4.91). Finally, sp deletes the structure associated to h1 and removes those messages from SBuffer whose only missing acknowledge was the one from h1.

When receiving the Rsp2 message, sn has the information needed to ensure that h1 delivers all messages exactly once (at least once and at most once), respecting the causal order of them.

At least once: sn keeps all messages received after the reception of the init message and sends them to h1. In fact sn associates a structure to h1 (Algo2.32), and sets the seqC of this structure to minSeq(), corresponding to the sequence
number of the oldest message it stores in its SBuffer. Therefore, no message of $s_n$’s SBuffer will be discarded until $h_i$ acknowledged it, since a message is only discarded once all connected $h$ nodes have acknowledged it. Moreover, due to the FIFO channels between $s$ nodes, the messages received by $s_n$ before $init$ are received by $s_p$ before it receives $Req_1$. Thus, $h_i$ always delivers these messages, since $s_n$ requests them to $s_p$ in $Req_2$ (Algo4.87), if it has discarded those $s_p$ identified as not delivered by $h_i$.

At most once: $h_i$ might deliver messages from $s_p$ until it tries to connect to $s_n$ by sending an $init$ message. Since the wired network is FIFO, and $s_n$ sends $Req_1$ after receiving $init$, the messages that $h_i$ might have already delivered are those $s_n$ receives before the reception of $Rsp_1$. Among these messages which are still in its SBuffer, $s_n$ determines which ones have not been delivered yet by $h_i$: (1) messages identified by $s_p$ in the $Rsp_1$ (Algo4.96) message ($m_1$ in Figure 5); (2) messages that $s_p$ received between the sending of $Rsp_1$ and reception of $Req_2$ (Algo4.96) message ($m_3$ in Figure 5). Due to FIFO channels, $s_p$ has received these messages before the reception of $Req_2$ and includes them in the $Rsp_2$ message. Messages $h_i$ has already delivered are identified by $s_n$ which adds the id of $h_i$ to the $M_d$ of all the other messages received before $Rsp_1$ not included in (1) or (2).

$s_n$ and $h_i$ exchange messages to conclude the handoff procedure. $h_i$ must first deliver the messages it has not delivered but which $s_n$ has discarded ($m_2$ in Figure 5), before delivering $s_n$’s pending messages ($m_3$ in Figure 5). Those messages are those contained in $m_{nd1}$ and are sent by the BroadcastCo function (Algo4.98). They are sent as AppCo messages, to distinguish them with regular broadcast application messages by the $s$ nodes (App$_{S,h}$ messages). $h_i$ changes its state to connecting when receiving the first AppCo message (Algo1.27). These messages contain $h_i$’s id to identify the handoff (Algo1.26). They have a sequence number and are thus delivered respecting causal order, as App$_{S,h}$ messages. Moreover, Hlock$_{s}$ is set to false, since the handoff between the $s$ nodes is finished. Furthermore, if a handoff request message $Req_1$ was received during the handoff, then $s_n$ processes it (Algo4.99-100).

$s_n$ concludes the handoff by sending an initACK message to $h_i$, if no AppCo messages need to be sent to $h_i$ and there is no pending $Req_1$ request (Algo4.101), or once $h_i$ confirmed the delivery of all pending AppCo messages. It assigns to $h_i$ the sequence number ($seq_c$) of the oldest pending message of its SBuffer which was not delivered by $h_i$. Finally, $h_i$ updates its sequence ($seq_C$) and session ($Ses_{LC}$) numbers, and removes from its SBuffer the messages that initACK has acknowledged. $h_i$ also restarts the sending of ack messages.

Handoff example: Figure 5 shows a simple example of a handoff procedure when $h_i$ moves from cell 1 to cell 2 but that covers the general case. Three messages $m_1$, $m_2$, and $m_3$, are broadcasted: broadcast($m_1$) is concurrent to broadcast($m_2$) and broadcast($m_2$) → broadcast($m_3$). $h_i$ has already delivered $m_1$. We consider that $s_1$ has discarded $m_1$ and stores $m_2$, while $s_2$ has discarded $m_2$ and stores $m_3$ ($SBuffer_1 = \{m_2\}$ and $SBuffer_2 = \{m_1\}$). Both $s$ nodes receive $m_3$ during the handoff.

The handoff procedure starts when $h_i$ sends an init message to $s_2$. The init message contains seq=1, since $h_i$ delivered $m_1$. We assume that no other handoff takes place for $h_i$ simultaneously. Upon reception of init, $s_2$ sends Req1 message to $s_1$ with seq=1 and Ses$_{LC}$ which contains the session number of the connection with $s_1$.

When receiving $Req_1$, $s_1$ learns that $h_i$ has not delivered $m_2$ since seq=1. It thus replies with $Rsp_1 = \{\{id(m_2)\}, seq_h = 0\}$, where seq$_h$=0 because $h_i$ did not broadcast any message.

Upon reception of $Rsp_1$, $s_2$ requests $m_3$ (Req$_2$) since it discarded $m_2$. $s_1$ replies with $Rsp_2$ message which contains the list of requested messages (\{m$_3$\}) and the list of received messages since $s_1$ sent $Rsp_1$ (\{id(m$_3$)\}). Moreover, $s_1$ deletes the structure associated to $h_i$.

Based on $Rsp_2$, $s_2$ can determine which messages of its SBuffer=\{m$_1$, m$_3$\} $h_i$ has delivered: $s_2$ received $m_1$ before $Rsp_2$ and $m_1$ is not identified by $s_1$ as not delivered by $h_i$. Therefore, $h_i$ already delivered $m_1$ and $s_2$ adds $h_i$’s id to $m_1$’s $M_d$. $s_1$ received $m_3$ between the send of $Rsp_1$ and the send of $Rsp_2$. Hence, $h_i$ did not deliver $m_2$ and must deliver it before delivering $s_2$’s pending messages. Thus, $s_2$ sends $m_2$ with seq=0 to $h_i$ and discards $m_2$ once $h_i$ acknowledged it.

Finally, $s_2$ assigns to $h_i$ seq=3, since $h_i$ delivered $m_1$ and $m_2$, and sends an initACK message to $h_i$, concluding the connection process. When receiving initACK, $h_i$ sets seq=3. Hence, $m_3$ is the next message $h_i$ will deliver.

E. Fault resilience

$h$ nodes are subject to transient faults. For recovery sake, $h$ nodes save the following variables on persistent local storage: the sequence number of broadcasted and received messages, both session numbers, and the SBuffer. Note that our experiments show that $h$ nodes’ SBuffer are of small size. The SBuffer and its associated sequence number are saved whenever the $h$ node broadcasts a message, while the sequence number of received messages is saved when it sends ack messages. Finally, Ses is saved each time it is incremented and Ses$_{LC}$ whenever a new session connection is confirmed (reception of initACK message). Upon recovering, a $h$ node restores these variables and sends an init message, similarly to when it changes cells. Therefore, $h$ node transient faults are tolerated with few persistent information.

On the other hand, permanent failures are not tolerated. A $s$ node keeps a pending message in its SBuffer until all $h$ nodes connected to its cell acknowledged it, move to another
cell, or leave the system properly. Hence, $s$ nodes would never discard unacknowledged messages in the presence of permanent failures. The memory footprint of $s$ nodes would then grow infinitely, and the wireless network would rapidly be overloaded. In fact, all these unacknowledged messages would be periodically broadcasted on the cell’s wireless network, increasing message loss rate until all messages on the wireless network would be lost due to interferences. Based on this observation, we point out that transient failures cannot last too long. However, this limitation is inherent to wireless networks’ nature. No algorithm can safely discard messages and cope with interferences in such conditions. Nevertheless, transient failures of long duration, as well as permanent failures, could be handled with the assumption that a $h$ node recovers and reconnects to the $s$ node to which it was connected before failing within at least $T$ seconds. In this case, a $s$ node would delete the information it stores about a failed $h$ node if no message is received from it during $T$ seconds, considering this $h$ node as a new one when it recovers.

### V. Performance Evaluation

We have conducted experiments on OMNeT++, with the INET extension [18]. INET renders simulations more realistic by implementing communication layers (e.g., TCP/UDP/Ethernet/IPv4/MAC), node mobility, propagation delays, and wireless networks with interferences.

The wireless network’s antennas have a communication range of 120m. The wired network has a bandwidth of 10 Mb/s and a delay of 10 ms between each link. The network contains 7 $s$ nodes, configured as in Figure 2. Initially, 70 $h$ nodes are placed randomly, connected to the closest $s$ node.

Application messages have a fixed size of 100 bytes and are encapsulated into UDP/TCP/IPv4/MAC packets. These protocols have headers of 8 bytes for UDP, and 20 bytes for TCP, IPv4, and MAC. Thus, application messages sent with UDP have a length of $100+8+20+20=148$ bytes, and those sent with TCP (TCP Reno) a length of $100+20+20+20=160$ bytes. Our algorithm uses TCP only on the wired network while UDP on the wireless network. A separately implemented control module verifies that messages are causally delivered.

Experiments were executed several times, and the initial position of the $h$ nodes is set randomly at each run. Moreover, experiments are run with different message emission frequencies. In the first experiment, we compare our algorithm with a TCP flooding one, considering that $h$ nodes are static. We then evaluate our algorithm in a dynamic context where $h$ nodes move but do not fail. We then extend this experiment where $h$ nodes can fail and recover (transient failures). Finally, we analyze the memory footprint.

#### A. Static system

Figure 6 compares our algorithm with [13] (denoted TCP-Flooding), which is based on reliable FIFO channels implemented with TCP connections.

Each $h$ node broadcasts, on average, an application message every 12.5 seconds. Since there are 70 $h$ nodes, there are $70/12.5 \approx 5.6$ messages broadcasts per second. In TCP-Flooding, a $s$ node disseminates an application message by sending it point-to-point to each of its connected $h$ nodes. However, some application messages might be acknowledged right after reception, and the respective $	extit{ack}$ messages will then collide on the wireless network with the application messages sent to the other $h$ nodes. Hence, in order to reduce these collisions, we have included a 5 ms delay between every point-to-point sending of a given application message by a $s$ node to each of its connected $h$ nodes.

Figure 6a gives the average number of messages stored in $s$ nodes’ $SBuffer$. We use a logarithmic scale in the case of Figure 6a, due to the different order of magnitudes of $SBuffer$ sizes. Figure 6c shows that our algorithm’s $h$ nodes’ $RBuffer$ are much smaller (0.5$<$) than TCP-Flooding’s (<5). The size of $SBuffer$ of $h$ nodes and the $RBuffer$ of $s$ nodes are small, because $h$ nodes only disseminate one message every 12.5 seconds. Therefore, we do not discuss them. On the other hand, the number of messages in TCP-Flooding’s buffers is much higher than in our algorithm. Using TCP-flooding, $s$ nodes’ $SBuffer$ stores many more messages because of the choice of the communication protocol itself and the congestion avoidance strategy.

For broadcasting an application message inside its cell, a $s$ node needs to send the message to each $h$ node of its cell using TCP (point-to-point communication) whereas our algorithm sends only one UDP message (broadcast function). Hence, until reception of the corresponding acknowledgment, a $s$ node keeps in its $SBuffer$, on average, $\approx 70/7=10$ messages per application message for TCP-Flooding, and only one for our algorithm.

TCP’s congestion avoidance strategy increases the time interval during which messages are stored in $SBuffer$. It consists of multiplying by 2 the retransmission delay (beginning at 0.2s) at each attempt. On the other hand, in our algorithm, retransmission delays start at 1s, and decrease with the number of sent messages, down to 200ms, aiming at delivering long outstanding messages faster. Moreover, TCP bounds the number of messages simultaneously sent to 46 (congestion window), while our algorithm bounds it to 150. The congestion window’s standard size is 7504 bytes, corresponding to 7504/160≈46 application messages sent simultaneously, while we limit it by 150 messages. Hence, TCP-Flooding takes longer to deliver messages, even though congestion avoidance reduces collisions (5% vs 10%). Nevertheless, messages are mainly stored on $s$ nodes’ $SBuffer$ since the congestion window born the number of simultaneously sent messages at 46.

A final remark is that we observe a low variation of buffer sizes in our algorithm, while TCP-Flooding’s buffer sizes vary a lot due to congestion avoidance (peaks) and fast retransmission (rapid decrease). The latter happens when outstanding messages are re-sent without waiting for the trigger of the corresponding timeouts upon detection of low network load. Our algorithm does not implement such mechanisms, and the buffer sizes are quite stable.

Figure 6b shows that TCP-Flooding sends more information than our algorithm. The former sends bigger messages
Our algorithm
TCP-Flooding

Fig. 6: Experimental results in static configuration

than the latter, due to TCP’s aggregation of messages. Hence, more data is lost and must be re-sent, even though fewer messages are lost. Moreover, 6d shows that TCP-Flooding sends more application messages, as well as more acknowledge messages (at least one/200ms vs one/500ms for our algorithm). Hence, the average ratio of messages sent per delivery, which stabilizes quickly for both algorithms, is much lower for our algorithm (0.4 msg/delivery) than TCP-Flooding (1.58 msg/delivery).

We evaluate the average message delivery delay, defined as the average time between the broadcast of an application message \( m \) \((\text{coBroadcast}(m))\) by a \( h \) node and the delivery of \( m \) \((\text{coDeliver}(m))\) by the \( h \) nodes. Our algorithm delivers messages with an average of 0.20s, much faster than TCP-Flooding, whose average delivery delay is 2s.

We also observe that TCP-Flooding hardly handles heavier loaded networks, because some \( h \) nodes will have great difficulty in receiving messages due to repeated collisions and the ensuring congestion avoidance strategy. Moreover, the establishment of TCP connections is sometimes long and even fails (exceeds 75 seconds). These problems are much worse in highly loaded networks, leading to buffers, and delivery delays, which grow indefinitely. Consequently, we could not collect meaningful statistics for TCP-Flooding in heavier loaded scenarios. On the other hand, our algorithm tolerates it (up to 35 messages/second), with an average of \( s \) nodes’ SBuffer size of 35 messages, an average delivery time of 0.31 seconds, RBuffer size of \( h \) nodes of 6 messages, an average of 700 sent messages/second, and, on average, 4300000 bytes of data sent/second.

B. Dynamic system

In this section, we do not compare our algorithm with TCP-Flooding because the mobility model described in the article [13] is not comparable to ours: mobile nodes must always be connected to at least one base station preventing temporal disconnection. Furthermore, as we have shown in the previous section, the TCP channels approach used by the algorithm strongly degrades the performance of message flooding over mobile networks.

We keep the same network configuration, except that the \( h \) nodes move, with a velocity of 5km/h \( \approx 1.38 \text{m/s} \). Our algorithm handles \( h \) nodes moving outside the covered area, but the moving \( h \) node would then stop receiving and acknowledging messages, and the SBuffer of the \( s \) node at which it was previously registered would then grow until it reconnects to some \( s \) node. For the sake of clarity of results, \( h \) nodes move inside the area covered by the station cells.

Every \( h \) node sends, on average, an application message every 2.8 seconds, i.e., a total of 25 application messages are broadcasted per second. Note that the network load of this scenario is much higher (25 messages per second) than those of the static experiment (5.6 messages per second).

Figure 7a shows the size of \( s \) nodes’ SBuffers and \( h \) nodes’ RBuffers. We observe that network dynamics do not have a significant impact on performance. The variations of the buffer sizes are mostly due to the collision of messages (between
7.5-8.5%). The loss of a message delays the acknowledgment and the removal/delivery of all the messages whose sequence number is higher than the one of the lost message. On the other hand, when the latter is finally received, several messages are generally acknowledged. Hence, the buffer sizes first increase progressively, before decreasing abruptly. Moreover, the increasing peak of RBuffer’s size usually comprises the RBuffer of several h nodes of the same cell. In addition, h nodes in regions where cells overlap also have a bigger RBuffer, because of interferences caused by the collision of messages sent by the two s nodes whose cells are overlapping.

s nodes’ SBuffers usually contain no more than a few dozen messages. Moreover, h nodes mostly move between neighbor cells, which receive messages similarly, and which have a high probability to store the same messages. Hence, the list of discarded messages exchanged during the handoff is small or even empty and is quickly propagated among the neighbor h nodes. Some h nodes may not receive many messages when their cell is loaded, or when they are changing cell several times in a short time interval. Their cell’s SBuffer and the list of discarded messages exchanged during the handoff then become much bigger (up to 200 messages).

We point out that the number of messages kept by SBuffer of h nodes and the RBuffer of s nodes are not shown in any figure since their respective size keeps very small during the whole execution (≈0.1-0.3 messages/node).

Results of the experiment for the number of sent messages (Figure 7b) are quite similar to the ones for static networks, since the size of the buffers are almost the same and only a few discarded messages must be sent to moving h nodes.

The average delivery latency is ≈0.3 seconds, slightly higher than in the static network (+0.05s), but keeps stable, even though buffers’ size varies. Indeed, the majority of the h nodes receive and deliver messages at the first broadcast. Therefore, most of them are delivered quickly and with the same latency. The slight increase in delivery delay is due to the slight increase of collisions and the handoff procedures, where h nodes do not deliver messages, and messages they send are discarded.

C. Dynamic system with transient faults

We keep the dynamic configuration and inject transient faults on randomly chosen h nodes connected to the s node s3. Beginning at t=15s of the experiment, a h node fails at every 30 seconds. The first failure lasts 3s and the duration of the failure increases by 1s at each new fault, i.e., the second lasts 4s, the third 5s, etc. A faulty node stops sending and receiving messages and comes back at its previous location.

Figure 8a shows the average size of all s nodes’ SBuffer, and Figure 8b the size of s3’s SBuffer. The former shows that the main impacted SBuffer of s nodes is the one of the cell in which the fault occurs, which is s3. Nevertheless, it drops quickly down to its previous size a few seconds after the faulty h node recovers.

We observe that the longer the failure duration, the bigger the SBuffer size of the s node to which the faulty h node was connected, and the longer the delay required for this h node to deliver the outstanding messages of the s node’s SBuffer when it recovers. Thus, the maximum duration of faults that s nodes can cope with is bounded by the maximum number of simultaneously disseminated messages per second (supported network load), and the maximum number of messages that s nodes’ SBuffer can keep. In our simulations, s nodes can send up to 400 messages simultaneously. Beyond it, messages will not be delivered fast enough to counterbalance new incoming messages, because of too many interferences. The number of outstanding messages of the s node’s SBuffer to which the faulty h node was connected will then increase indefinitely, rendering the cell unstable. This instability will eventually propagate to the other cells.

Figure 8b also shows two special cases. First, the SBuffer of s3 does not entirely recover between t=200 and 250s. A h node fails at t=195s and recovers at t=204s. s3’s SBuffer then begin to reduce, but a new fault occurs at 225s, before the SBuffer comes back to its original size. Secondly, between 165 and 180s and 225 and 250s the size of SBuffer of another station (resp. s5 and s2) also grows due to interferences when s3 broadcasts many messages, or by the changing of cells of a recovering h node before it received all outstanding messages.
D. Scalability and memory analysis

The control information included in application messages concerns just some few integers while experiments show that the list of id’s ($M_0$), piggybacked on $s$ nodes’ application messages usually contains one or no id.

The two lists of messages exchanged during the handoff between the $s$ nodes contain at most the number of messages stored by the previous $s$ node’s SBuffer to which the moving $h$ node was connected. Experiments show that the SBuffer of $s$ nodes remains small, and practically very few messages are received by $s$ nodes during the handoff procedure, since $h$ nodes mostly move between adjacent $s$ nodes. Hence, messages piggyback a few control information.

Additionally to the SBuffer and RBuffer, $h$ nodes keep some integer variables, and $s$ nodes some integers and a structure for each connected $h$ node. Experiments show that all buffers, except $s$ nodes’ SBuffer, remain very small. $s$ nodes’ SBuffer contains only a few dozen messages except temporarily when some previously connected $h$ node fails or if the $s$ node’s cell is overloaded. The structure $s$ nodes keep for each $h$ node contains only a few integers. The number of those structures a $s$ node keeps grows linearly with the number of $h$ nodes connected to the $s$ node’s cell.

In summary, the amount of control information required by the algorithm is low for $h$ nodes and grows linearly in terms of locally connected $h$ nodes for $s$ nodes.

E. Experiments conclusion

We can draw the following conclusions from the results of our experiments:

Firstly, TCP-Flooding algorithm induces a high local control information overhead, increased transmission delays, and a high number and size of sent messages. Therefore, it is not suited for wireless networks, even in static systems.

Secondly, our algorithm shows good results in a dynamic network. Node mobility has a slight impact on the performance of causal broadcast, and performance is similar to those presented for static networks. On the other hand, transient faults degrade performance, particularly within the cell in which they occur. However, such a degradation disappears after a few seconds when the faulty $h$ nodes recover. Hence, our algorithm tolerates high load, mobility, and transient faults.

Finally, in terms of scalability, very few control information is piggybacked on messages. Moreover, $h$ nodes maintain few information, and $s$ nodes’ memory footprint grows linearly with the number of locally connected $h$ nodes. Overall our algorithm has a small memory footprint.

VI. RELATED WORK

Several approaches have been proposed aiming at reducing the control information necessary to implement causal broadcast algorithms in distributed systems. However, they are not always suitable for tolerating the dynamics, and wireless interferences of mobile networks neither present solutions for discarding no longer usable messages.

Prakash and al. observed in [14] that, for controlling causal order among messages, it is sufficient to piggyback in a message just the information about its direct dependencies. Even though this approach copes with node dynamics (churn) of mobile networks, since it is based on message identifiers and not on node identifiers, every node maintains a matrix of size $N^2$, where $N$ is the maximum number of nodes of the system. Additionally, in the worst case, the control information attached to messages presents a size of $O(N)$.

Vector clocks [6][10] of size $N$ are the smallest data structures which capture causality of events of a system with $N$ nodes [5]. Nevertheless, to implement a causal broadcast algorithm that exploits vector clocks [16], each node keeps a local vector variable of size $N$, which is included in every broadcasted message. Hence, since the vector size is fixed to $N$, such clocks are not adjustable for systems where the number of nodes varies dynamically. Their size still grows linearly with $N$, even with solutions that compress them, such as [3]. Moreover, a garbage collector and additional control information are needed to delete obsolete messages.

By applying Bloom filters on messages, Ramabaja aims in [15] at reducing messages’ piggybacked information. These filters have a much lower space complexity, but they can throw false positive (but not false negative) requiring, therefore, a mechanism to handle them. Furthermore, in order to limit the number of false positives, Bloom filter’s size should increase proportionally to the number of nodes.

Both plausible [17] and probabilistic clocks [2] are vector clocks whose size is much smaller than the number of nodes in
the system. The quality of the detected causality information is related to the vector size: the greater the size of the vector, the higher the accuracy of the captured causality. In plausible clocks, an entry of the vector clock is associated with several nodes. The probabilistic clocks extend the plausible clock by also associating several entries to a node. The use of these vector clocks by causal broadcast algorithms strongly reduces the size of control information but they do only capture causality among broadcast messages with a high probability. An extra procedure is necessary to handle causally related messages that were not ordered. Moreover, contrary to our solution, the authors do not propose any mechanism to discard delivered messages.

Some works address scalability issues by organizing nodes in logical structures. In Adly et al. [1], nodes are grouped into clusters, logically organized into a tree. Thereby, a node needs to send and keep track of messages only to/from few nodes. However, at the presence of node churn, clusters often need to be reorganized. Moreover, the mobility of a node is restricted to its cluster. Hence, it is not suitable for mobile networks.

By organizing the nodes into an application-level tree on top of which messages are propagated, Blessing et al.’s causal broadcast algorithm does not require that messages carry any causal information. However, the authors do not consider system dynamics.

Nédelec et al. [13][12] extend this approach to dynamic topologies. A node discards a message once it has received it by each of its links. Links are FIFO, and they can be dynamically added or removed between nodes using handoff procedures, provided that the nodes of the system are always connected through initialized links. Therefore, dynamics are tolerated under certain conditions. Particularly, to add a new link between two nodes, an already initialized path must exist between them. Hence, the system can never be partitioned.

VII. CONCLUSION

We have presented in this article a causal broadcast algorithm that is tailor to the characteristics of mobile networks, such as nodes mobility, dynamic membership and connections, mobile nodes memory constraints, scalability issues, and wireless interferences. The size of required information piggybacked on messages is small, as well as mobile nodes’ memory footprint, while mobile support stations’ memory footprint grows linearly with the number of local connected mobile nodes. Simulation results on OMNet++ show that TCP induces a heavy message overhead in mobile networks and that our causal broadcast has good performance in both static and dynamic systems.

For future work, we aim to extend our causal broadcast algorithm in order to handle mobility of support stations. A second research direction will be a hybrid approach with FIFO ordering and probabilistic clocks [2], proposed in [12]. The latter partially restores causality tracking and thus allows the delivery of some concurrent messages without waiting for their FIFO ordering.

REFERENCES