Practical consistency management for geographically distributed MMOG servers

June 6, 2011

Abstract

1 Introduction

2 System model and definitions

We assume a system composed of nodes, divided into *players* and *servers*, distributed over a geographical area. Nodes may fail by crashing and subsequently recover, but do not experience arbitrary behavior (i.e., no Byzantine failures). Communication is done by message passing, through the primitives $send(p, m)$ and $receive(m)$, where p is the addressee of message m. Messages can be lost but not corrupted. If a message is repeatedly resubmitted to a *correct node* (defined below), it is eventually received.

Our protocols ensure safety under both asynchronous and synchronous execution periods. The FLP impossibility result [2] states that under asynchronous assumptions consensus cannot be both safe and live. We thus assume that the system is initially asynchronous and eventually becomes synchronous. The time when the system becomes synchronous is called the *Global Stabilization Time (GST)* [1], and it is unknown to the nodes. Before GST, there are no bounds on the time it takes for messages to be transmitted and actions to be executed. After GST, such bounds exist but are unknown. After GST nodes are either *correct* or *faulty*. A correct node is operational "forever" and can reliably exchange messages with other correct nodes. This assumption is only needed to prove liveness properties about the system. In practice, "forever" means long enough for one instance of consensus to terminate.

A game is composed of a set of objects. The game state is defined by the individual states of each one of its objects. We assume that the game objects are partitioned among different servers. Since objects have a location in the game, one way to perform this partitioning is by zoning the virtual world of the game. Each partition consists of a set of objects of the game and the server responsible for them is their *coordinator*. As partitions represent physical regions in the game world, we define *neighbor* regions as those which share a border with each other. We consider that each server is replicated, thus forming several groups which consist of the coordinator and its replicas. Therefore, for each region of the game world, there is a group assigned to it. This way, a group is said to coordinate an object when the group's coordinator is the coordinator for that object. Finally, groups are called neighbors when they are assigned to neighbor regions. From now on, the word 'server' will be used for any kind of server, be it a coordinator or a replica.

Each player may send his command to one of the servers – which might not be the group's coordinator, if that provides a lower delay between the issuing of a command and its delivery. In the case of avatar based virtual environments, and as an avatar is usually also an object, the server to which a player is connected belongs to the group of his avatar's coordinator. A command $C = \{c_1, c_2, ...\}$ is composed of one or more subcommands, one subcommand per object it affects. We refer to the set of objects affected by command C as $obj(C)$. Also, we refer to the objects coordinated by a server S as $obj(S)$. Finally, we define $obj(C, S) = \{o : o \in obj(C) \text{ and } o \in obj(S)\}$.

Our consistency criterion is "eventual linearizability". (I'm not sure this is indeed what we want and how to define it, but we do need some consistency criterion...)

3 Protocol

To ensure reliability despite server crashes, each server is replicated using state-machine replication, implemented with consensus (e.g. Paxos [3]). Each player sends his commands to the server to which he is connected, which

then proposes that command in a consensus instance. Each command is assigned a timestamp and executed against objects in timestamp order. We implement this by using a logical clock in each server group. Guaranteeing that the same set of commands is executed by the respective affected objects in the same order provides the level of consistency we are seeking. Therefore, the challenge is how to assign timestamps to commands such that consistency is not violated and commands are not discarded due to stale timestamp values.

However, providing such level of consistency may prove to be costly in an MMOG context, since there may be several communication steps between the sending of a command and its atomic delivery. For this reason, we use a primitive we call quasi-genuine global fifo total order multicast, which is described in section 3.1.

3.1 Quasi-genuine Global FIFO Multicast

To reduce the time needed to deliver a message, we use a multicast primitive which delivers messages optmistically, based on the time when they are created in a sender. This *optimistic* delivery, although not guaranteeing that all replicas will receive all messages, is designed to have a fairly low latency, counting from when each message is sent until it is delivered.

The final – conservative, fault tolerant, but costly in terms of communication steps for each message – delivery order should be as close as possible to the optimistic one, so that no rollbacks would be deemed necessary. To explain how this works, we first should understand the idea behind the optimistic delivery. Also, we must define what "quasi-genuine" and "global FIFO" means.

Genuine multicast protocols are those where two multicast groups only communicate with each other when one has some message to send to the other. A **quasi-genuine** multicast protocol assumes that:

A1: every group knows from which other groups a multicast message can possibly arrive, and to which other groups it can send a multicast message.

In a quasi-genuine multicast protocol, different groups communicate with each other even if there is no message to be sent from one to the other, but, from A1, there is no need for a group g_i to communicate with some other group g_j which cannot send messages to g_i, or receive messages from g_i[1]. This information may be given by the application which is making use of such primitive, so, although not genuine, a quasi-genuine protocol does not imply that each group has to keep sending messages to every other group.

We define $sendersTo(G)$ as the set of groups which are able to send a message to G. Also, we define $receiversFrom(G)$ as the set of groups who are able to receive a message from G.

As messages are delivered according to their generation time, the delivery order is FIFO. Besides, as every two messages which are delivered in different groups should be delivered in the same order in these groups, we need total order. However, even FIFO total order is not enough to guarantee one property that we need: let the send time $st(m)$ of a message m be the value of the wallclock of p when m was sent, where p is the process which sent it. We need that, if $st(m_1) < st(m_2)$, then m_1 should be delivered before m_2, if m_1 and m_2 are delivered by the same process. Although this property may be hard to observe, it is used by the optimistic delivery protocol, so that messages can be delivered in order in one communication step, in the best case. If, when each process sends a message m, it stamps the message with the current value $st(m)$ of its wallclock, we define:

Global FIFO order: if processes p_i and p_j send respectively m_i and m_j to the same destination p_d, then if $st(m_i) < st(m_j)$, m_i is not delivered after m_j in p_d.

This property implies the FIFO order, although it does not guarantee that, if a message m was sent, it will ever be delivered. As for the other properties of the algorithm, assuming a crash-stop model, we have:

Uniform Agreement: if a process GF-CONS-Delivers m, then every correct process that is a destination of m also GF-CONS-Delivers m.

Uniform Integrity: for any message m, every correct process p GF-CONS-Delivers m at most once, and only if some process executed GF-Mcast(m) and p is one of m's destinations.

Uniform Total Order: if processes p and p' both GF-CONS-Deliver m and m', then p GF-CONS-Delivers m before m' if and only if p' GF-CONS-Delivers m before m'.

Regarding the **termination** property, as presented in [4], for being guaranteed by the GF-Mcast primitive, it requires that an optimistic assumption is true:

A2: every process p knows a value $w(p)$, which is at least the maximum sum of the message delay bound plus the clock deviation between p and any possible sender process p' which could send a message to p.

[1]This relation may even be asymmetric: it could be possible to send a message from group g_i to g_j, but not in the opposite direction. In this case, in a protocol where a group is blocked waiting for possible messages from other groups, g_j may block its delivery of messages waiting for some kind of "clearance" from g_i, but g_i will never block waiting for messages from g_j.

We then relaxed the termination property to another one, which considers the optimistic assumption we just stated:

Optimistic termination: For any message m, if a correct process GF-Mcasts m **and the assumption A2 is true** then all correct processes that are destinations of m GF-CONS-Deliver m.

Here, we consider that there are several processes. Each process p belongs to a group $G = group(p)$, that is, $p \in G = group(p)$.

3.1.1 Optimistic delivery

The basic idea of the optimistic ordering is the following: assuming that the processes have a synchronized clock[2], whenever a process p receives a message m from a client, it immediately applies a timestamp ts to it, which consists simply of the current value of p's wallclock, now. Therefore, $m.ts = now$. A wait window of length $w(p)$ is considered, where $w(p)$ is defined as the highest value of the estimated communication delay plus the wallclock deviation between the process p and any of the other processes in its group G or in any of the groups in $sendersTo(G)$.

More formally, let $\delta(p', p)$ be the maximum time for a message from p' to arrive at p. Also, let $\epsilon(p', p)$ be the deviation between the wallclocks of p and p'. The value of $w(p)$ is estimated as the maximum value of $(\delta(p', p) + \epsilon(p', p))$ for every p' in $G = group(p)$ or in some group of $sendersTo(G)$.

After applying the timestamp to m, the process p immediately forwards it to all the other processes involved, including those in other groups. A process is involved with a message m when it is one of its destinations, which can be inferred from $m.dst$. Then, p puts m in an *optPending* list, where it stays until $now > m.ts + w(p)$, which means that m has been in that list for a time longer than the defined wait window. In the meantime, other messages, sent from other processes, may have been received and also inserted in that list, always sorting by their timestamps.

If $w(p)$ has been correctly estimated, and no message was lost, than all the messages that were supposed to be delivered before m have necessarily been received already. If the same has occurred for all the processes, then all of them have received all the messages, and can deliver them in the same order of ts.

To avoid out-of-order deliveries, if when m arrives at p, $m.ts < now - w(p)$, wich means that m arrived too late, the message m is simply discarded by p, since another message $m' : m'.ts > m.ts$ may have already been delivered[3].

However, even if the optimistic order is the same for all the processes, it won't be valid if the conservative order is different from it. For that reason, we devised a way to make the conservative delivery order as close as possible to the optimistic one.

3.1.2 Conservative delivery

Instead of having the conservative delivery done completely independently from the optimistic one, we can actually use the latter as a hint for the final delivery order. Since the optimistic delivery should be fast when compared to the conservative one, waiting for it should not decrease the system performance significantly. Also, if we wait a short period longer, we can avoid a rollback later caused by mismatches between the two delivery orders, which is not desirable. The basic idea is, then, to pick the optimistic delivery order seen at the processes of each group.

The complicating factor is the possibility of a message having at least one destination group different from its source group. So, all involved groups must somehow agree regarding the delivery order of these messages. However, from assumption A1, each group knows which other groups it could send messages to – or receive messages from. We can use this by defining *barriers* for multicast, such that $barrier(G_{send}, G_{recv}) = t$ means that the group G_{send} promised that it would send no more messages with a timestamp lower than t to group G_{recv}. We have defined that $sendersTo(G) = \{G' : G' \text{ is able to send a message to } G\}$. When a process p, from group G, has received all the barrier values from all the groups in $sendersTo(G)$, and they are all greater than a value t, then p knows that no more messages with timestamp lower than t are coming from other groups and that, once the local ordering (the ordering of messages originated in G) is done, all the messages with timestamp up to t can be conservatively delivered. Besides, a barrier is sent along with the bundle of all messages with timestamp greater than the last previous barrier sent from G_{send} to G_{recv}, so that when a process has received a barrier from a group, it means that it knows all the messages sent by that group until the time value stored in that barrier.

We use consensus to conservatively deliver messages. Consider that each consensus instance I from each group $I.grp$ receives a monotonically increasing unique integer identifier, without gaps, that is, for any two instances I_i

[2]We don't require here perfectly synchronized clocks, as the optimistic protocol tolerates mistakes by its very definition. We only need clocks which are synchronized enough, so that our delivery order prediction succeeds and matches the conservative delivery order.

[3]We could make this in a way such that m is only discarded by p if, in fact, there was a delivered message $m' : m'.ts > m.ts$. If there was no such message, m could still be delivered without violating the order we defined.

and I_k, such that $I_i.grp = I_k.grp$, if $I_i.id + 1 < I_k.id$, there is necessarily an instance $I_j : I_i.grp = I_j.grp = I_k.grp \wedge I_i.id < I_j.id < I_k.id$. No group runs two consensus instances in parallel: before initiating an instance of id $k+1$ each process checks whether the instance k has already been decided, so some messages may wait to be proposed. When a process is allowed to initiate a new consensus instance, the pending messages may be proposed as a batch.

There are three possibilities for each message m, in the perspective of a process p of G:

- The message m was originated in G, which is the only destination of m:

  In this case, when m is optimistically delivered by p, p checks whether the latest consensus instance I_{prev} in which it participated, or is trying to start, has already been decided – if not, p enqueues m in a *propPending* queue as the next message being proposed by it, so other tasks can keep being executed. Then, once I_{prev} has been decided, p may start a new instance. Before that, all messages in *propPending* that have either been already decided, or that have a timestamp smaller than that of some already decided message, are discarded from *propPending*. The rest is proposed as a batch in the consensus instance. Once m is decided, it is not immediately delivered to the application. Instead, it is inserted in a *consPending* list for later being conservatively delivered, which will happen once every group G' in $sendersTo(G)$ has already sent a message $barrier(G', G) = t$, such that $t > m.ts$. This is done because there could be a message m' yet to come from another group G', such that $m'.ts < m.ts$.

- When m is originated in G, but it has at least one group other than G as a destination:

  In this case, when $now > m.ts + w(p)$, p tries to initiate a consensus instance within G to decide m. If p cannot start the proposal now, m is enqueued in *propPending* for being proposed later along with other pending messages. Then, once p may start a new instance, all messages in *propPending* that have either been already decided, or that have a timestamp smaller than that of some already decided message, are discarded from *propPending*. The rest is proposed as a batch in the new consensus instance. Once any message m is decided, if $G \in m.dst$, m is inserted in the *consPending* list. Besides, when m is decided, p sends a message $\{m, \text{`cons'}\}$ to every $p' : (\exists G' \in m.dst \setminus \{G\}) \wedge (p' \in G')$. When $\{m, \text{`cons'}\}$ is received by each $p' \in G'$, p' checks whether it has ever inserted m in its own *consPending* list. If not, p' inserts m into *consPending* and adjusts $barrier(G, G')$ to $m.ts$. To ensure that, once a message $\{m, \text{`cons'}\}$ is received, every message $\{m', \text{`cons'}\} : m'.ts < m.ts$ has already been received from $m.src$, every $\{m, \text{`cons'}\}$ message is sent through a lossless FIFO channel[4].

- When G is one of the destinations of m, but m was originated in some other group G':

  In this case, when m is optimistically delivered by a process p of group G, nothing else is done. The message m is inserted into the *consPending* list of p only when p receives some message $\{m, \text{`cons'}\}$.

The messages in the *consPending* list are always sorted in ascending order of their timestamps. When the first message m, in the *consPending* list of a process $p : group(p) = G$, is such that $m.ts < barrier(G, G')$ for all $G' \in sendersTo(G)$, then m is conservatively delivered by p to the application as the next message. We claim that this delivery respects the "global FIFO total order"[5].

A more formal description of the protocol is given in Algorithm 1. We consider that three primitives are given: $getTime()$, which returns the current value of the local wallclock; $Propose(k, val)$, which proposes a value val for the consensus instance of id k within its group; and also $Decide(k, val)$, which is called when the consensus instance of id k finishes. $Decide(k, val)$ is called for all the processes of the group that initiated it, when they learn that the value val has been agreed upon in instance of id k. For the sake of simplicity, we assume that, for consensus instances within the same group, the values are decided in the same order of the instances id's[6]. Finally, we also use a reliable multicast primitive R-MCast(m,*groupSet*), which R-Delivers m to all the processes in all the groups in *groupSet* in one communication step (e.g. the one described in [**?**]).

Moreover, each process p of group G keeps some lists of messages:

- *optPending*, which contains the messages waiting to be *GF-OPT-Delivered* (optimistically delivered);

- *propPending*, containing the messages waiting to be proposed by p;

[4] An ordinary TCP connection would be enough to provide such FIFO lossless channel.

[5] Proof needed.

[6] This can be easily done by delaying the callback of $Decide(k, val)$ while there is some unfinished consensus instance of id $k' : k' < k$ from the same group.

- *consPending*, with the messages ready to be *GF-CONS-Delivered* (conservatively delivered), but which may be waiting for barriers from the groups in $sendersTo(G)$;
- *decided*, which contains the messages that have already been proposed and decided within $group(p)$;
- *delivered*, which contains the messages that have been GF-CONS-Delivered already.

Also, each message m has at least three fields: dst, which is the set of m's destination groups; src, which is the group where m was generated; and ts, which is the value of the wallclock of the process when m was generated. Note that, by abuse of notation, we have 'process $p \in m.dst$' instead of '$\exists$ group $G \in m.dst$: process $p \in G$'.

Something that must be noticed is that some messages might not have their source groups as a destination. Anyway, each message m of this type still has to be agreed upon in it group of origin G, so that its order among other messages from G may be decided and m will be retrievable, even in the presence of failures.

Assuming that $w(p)$ is the same for every process p, the time needed to decide a message m after it has been GF-Mcast by some process is equal, in the worst case, to $w(p) + 2T_{cons}$, where T_{cons} is the time needed to execute a consensus instance. This value is counted twice because m may have been inserted into *propPending* right after a consensus instance has been initiated. In that case, m would have to wait such consensus to finish, to be then proposed and finally decided. However, it may also happen that m has been inserted into *propPending* right before some proposal has been made, so it would take only $w(p) + T_{cons}$ to decide m. As m may go into *propPending* anytime between the worst and the best case with the same probability, the average time needed for deciding m would be equal to $w(p) + 1.5T_{cons}$. Nevertheless, the time needed to finally GF-CONS-Deliver m will depend on when barriers are received from other groups.

3.1.3 Addressing liveness

The problem with Algorithm 1 is that it does not guarantee liveness when a group has no message to receive from some other group and then keeps waiting for a new message to increase the barrier value and proceed with the conservative delivery. However, it does not disrupt the optimistic delivery. Also, liveness for the conservative delivery can be easily provided by sending periodic empty messages from G to each $G' \in receiversFrom(G)$ to which no message has been sent for a specified time threshold *barrierThreshold*. Algorithm 2 describes this. When a message m from group G to some other group G' has been inserted into the *optPending* list of $p \in G$, it knows that the other processes of G did the same and that m will be eventually decided and sent to G', which will serve as barrier from G to G' (l. 37 of Algorithm 1). However, when there is a long period after the last time when such kind of message has been created, p decides to create some empty message to send to the processes of G' with the sole purpose of increasing their barrier values and allow for the delivery of possibly blocked messages in G'.

The problem with addressing liveness this way is that, in the worst case, G' has decided a message m and has just received the last barrier b from G, such that $b < m.ts$. This would mean that, if G has no messages to send to G', G' will have to wait, at least, for $barrierThreshold$ – maybe just to receive some $b' : b' < m.ts$, having to wait again and so on. How long exactly it will take to GF-CONS-Deliver m depends on many variables, such as how far in the past m was created and how long it takes for some barrier $b : b > m.ts$ to arrive.

It is necessary to guarantee that a *null* message will eventually be proposed, decided, and a barrier will be sent to some group which might be needing it, so that progression is guaranteed. Therefore, this kind of messages are created by every process in the group, since if any one of them does not have it, such message might never be decided. Besides, they are excluded from the *propPending* list only when decided in the group (l. 20 of Algorithm 1). Since they are obviously never delivered to the application, they can remain in such list and be decided even after some message with higher timestamp is decided. Not removing them unless they are decided ensures that they will be sent and that other groups will be able to increase their barrier values. To prevent the multiple *null* messages – created by different processes within the group – of being decided, they could be created in a way such that the different processes can somehow figure out that two different *null* messages are equivalent[7].

However, there is a way to provide liveness withouth such a possibly long delay for the conservative delivery of messages, although that would imply creating more messages and making deeper changes in the delivery algorithm. Let $blockers(m)$ be defined as the set of groups whose barrier is needed in order for some group to deliver m. More formally, $blockers(m) = \{G_B : \exists G_{dst} \in m.dst \wedge G_B \in sendersTo(G_{dst})\}$. The idea is that, once each group G_B in $blockers(m)$ have sent a barrier $b > m.ts$ to all the groups belonging to $m.dst \cap receiversFrom(G_B)$, all possible destinations of m can deliver it. This way, instead of relying on periodic messages, whenever a process p in a group

[7]This could be done by assuming no timestamp collisions and by using them to uniquely identify messages. Then, only one of the messages created with the same timestamp (l. 9 of Algorithm 2) would be decided.

Algorithm 1 GF-Mcast(m) – executed by every process p from group G

1: Initialization
2: $k \leftarrow 0$, $nextProp \leftarrow 0$, $decided \leftarrow \emptyset$, $delivered \leftarrow \emptyset$, $propPending \leftarrow \emptyset$, $optPending \leftarrow \emptyset$, $consPending \leftarrow \emptyset$
3: **for all** $G' \in sendersTo(G)$ **do**
4: $barrier(G', G) \leftarrow -\infty$

5: *To send a message* m – GF-Mcast(m)
6: $m.ts \leftarrow getTime()$ {*current wallclock value as the timestamp of* m}
7: R-MCast(m,$m.dst \cup \{G\}$)

8: *When* R-Deliver(m')
9: **if** $m'.ts < getTime() - w(p)$ **then**
10: discard m' {*late commands probably lead to out-of-order delivery*}
11: **else**
12: $optPending \leftarrow optPending \cup \{m'\}$

13: *When* $\exists m \in optPending : getTime() > m.ts + w(p) \wedge \nexists m' \in optPending : m'.ts < m.ts$
14: $optPending \leftarrow optPending \setminus \{m\}$
15: **if** $G \in m.dst \wedge m \neq null$ **then**
16: GF-OPT-Deliver(m)
17: **if** $G = m.src$ **then**
18: $propPending \leftarrow propPending \cup \{m\}$

19: *When* $\exists m \in propPending \wedge nextProp = k$
20: $propPending \leftarrow propPending \setminus (decided \cup \{m' : \exists m'' \in decided \wedge m'.ts < m''.ts \wedge m' \neq null\})$
21: **if** $propPending \neq \emptyset$ **then**
22: $nextProp \leftarrow k + 1$
23: Propose(k,$propPending$)

24: *When* Decide(k,$msgSet$)
25: $decided \leftarrow decided \cup msgSet$
26: **for all** $m \in msgSet : G \in m.dst \wedge m \neq null$ **do**
27: $consPending \leftarrow consPending \cup \{m\}$
28: **while** $\exists m \in msgSet : (\forall m' \in msgSet : m \neq m' \Rightarrow m.ts < m'.ts)$ **do**
29: **for all** $p' \in m.dst \setminus \{G\}$ **do**
30: send(p', $\{m, \text{`cons'}\}$) {*this message is sent through a FIFO lossless channel*}
31: $msgSet \leftarrow msgSet \setminus \{m\}$ {*the messages are sent in ascending order of timestamp*}
32: $nextProp \leftarrow k + 1$
33: $k \leftarrow k + 1$

34: *When* receive($\{m', \{\text{`cons'}\}\}$)
35: **if** $m' \notin consPending \wedge m' \notin delivered \wedge m' \neq null$ **then**
36: $consPending \leftarrow consPending \cup \{m\}$
37: $barrier(m'.src, G) \leftarrow max(m'.ts, barrier(m'.src, G))$ {*channels are FIFO, but there are different senders*}

38: *When* $\exists m \in consPending : \forall G' \in sendersTo(G) : m.ts < barrier(G', G)$
$\wedge \nexists m' \in consPending : m'.ts < m.ts$
39: $consPending \leftarrow consPending \setminus \{m\}$
40: GF-CONS-Deliver(m)
41: $delivered \leftarrow delivered \cup \{m\}$

Algorithm 2 Achieving liveness by sending periodic messages; executed by every process p of group G

1: Initialization
2: **for all** $G' \in receiversFrom(G)$ **do**
3: $lastBarrierCreated(G') = -\infty$

4: *When* inserting into *optPending* a message $m : (m.src = G \wedge \exists G' \in m.dst : G' \neq G)$
5: **for all** $G' \in m.dst : G' \neq G$ **do**
6: $lastBarrierCreated(G') \leftarrow m.ts$

7: *When* $\exists G' \in receiversFrom(G) : getTime() - lastBarrierCreated(G') > barrierThreshold$
8: $null \leftarrow$ empty message
9: $null.ts \leftarrow lastBarrierCreated(G') + barrierThreshold$
10: $null.src \leftarrow G$
11: $null.dst \leftarrow \{G'\}$
12: $optPending \leftarrow optPending \cup \{null\}$ {*saving that nothing was sent until null.ts*}

G has a message m to send, p sends m to every $p' \in m.dst \cup G \cup blockers(m)$. It is sent to the groups in $blockers(m)$, so that they know that there is a message which will be blocked until they send a proper barrier to unblock it.

When the process p' of some group G' receives m, such that $m.src \neq G'$, p' knows that there might be other groups depeding on the barrier of G' to deliver m. For that reason, it will immediately create a *null* message with a timestamp equal to $m.ts$ and with $null.dst = m.dst \cap receiversFrom(G')$. Then, p' will insert such message in its *optPending* queue – to ensure that, once it goes to *propPending*, any message $m' : m'.ts < null.ts \wedge m'.src = G'$ is already there. As soon as $null.ts > now - w(p')$, the *null* message will be proposed in G' and, once *null* is decided, each process of G' will send a $\{null, \text{'cons'}\}$ message to each process in each group $G \in m.dst \cap receiversFrom(G')$. This way, any group which was waiting for a barrier from G' to deliver m will be able to do so as soon as it receives such *null* message. The new delivery algorithm would be as described in Algorithm 3.

Now, assuming that $w(p)$ is the same for every process p, we have a worst case delivery latency for m of $w(p) + 2T_{cons}$, as the *null* message is created and proposed in parallel with m. However, each sender process is sending m not only to every destination process, but also to every process in its group and to every process belonging to some group of $blockers(m)$. This can create a fairly high amount of messages. Unfortunately, this has to be done to guarantee that such *null* message will eventually be proposed, decided, and a barrier will be sent to some group which might be needing it. Although this kind of messages are excluded from the *propPending* list only when decided in the group (l. 26 of Algorithm 3), if at least one of the processes does not have it, it could never be decided. Again, like in Algorithm 2, there might be many processes in a group proposing a *null* message because of the same m. Although this is not wrong, it is inefficient, and a way to identify *null* messages created for the same purpose should be devised[8].

3.2 Recovering from mistakes

Unfortunately, even with a very good delay estimation (e.g. on an environment with a low jitter), there is absolutely no guarantee that the multicast protocol described in section 3.1 will deliver the game command messages optimistically and conservatively in the same order. When it doesn't, it is considered a *mistake*. Every mistake of the optimistic delivery – either a lost command message, or an out-of-order delivery – will cause a rollback of the optimistic state of the objects and re-execution of some of the optimistically delivered commands.

To perform that, we consider that each object has an optimistic delivery queue, Q_{opt}. Whenever a command is optimistically delivered, the optimistic state is updated and the command is pushed in the back of Q_{opt}. Whenever a command C_c is conservatively delivered, it updates the conservative state of each object in $obj(C_c)$ and, for each one of them, the algorithm checks whether it is the first command in Q_{opt}. If it is, C_c is simply removed from Q_{opt} and the execution continues. If it isn't, it means that C_c was either optimistically delivered out of order, or it was simply never optmistically delivered. It then checks whether Q_{opt} contains C_c. If it does, it means the command was optimistically deliverd out of order, and it is removed from the list – if Q_{opt} doesn't contain C_c, it was probably

[8]Again, we could use timestamps as unique identifiers and assume no timestamp collisions. But for that to work, instead of making $null.ts = m.ts$ (l. 12 of Algorithm 3), we could define $null.ts$ as a little bit greater than $m.ts$.

Algorithm 3 GF-Mcast(m) – executed by every process p from group G

1: Initialization
2: $k \leftarrow 0$, $nextProp \leftarrow 0$, $decided \leftarrow \emptyset$, $propPending \leftarrow \emptyset$, $optPending \leftarrow \emptyset$, $consPending \leftarrow \emptyset$
3: **for all** $G' \in sendersTo(G)$ **do**
4: $barrier(G', G) \leftarrow -\infty$

5: *To send a message m* – GF-Mcast(m)
6: $m.ts \leftarrow getTime()$ {*current wallclock value as the timestamp of m*}
7: R-MCast(m,$m.dst \cup \{G\} \cup blockers(m)$)

8: *When* R-Deliver(m')
9: **if** $G \neq m'.src \wedge \exists G' \in (m'.dst \cap receiversFrom(G))$ **then**
10: $null \leftarrow$ empty message
11: $null.src \leftarrow G$ {*some other group needs a barrier from this one*}
12: $null.ts \leftarrow m'.ts$
13: $null.dst \leftarrow m'.dst \cap receiversFrom(G)$
14: $optPending \leftarrow optPending \cup \{null\}$
15: **if** $m'.ts < getTime() - w(p) \vee G \notin m'.dst$ **then**
16: discard m' {*late commands probably lead to out-of-order delivery*}
17: **else**
18: $optPending \leftarrow optPending \cup \{m'\}$

19: *When* $\exists m \in optPending : getTime() > m.ts + w(p) \wedge \nexists m' \in optPending : m'.ts < m.ts$
20: $optPending \leftarrow optPending \setminus \{m\}$
21: **if** $G \in m.dst \wedge m \neq null$ **then**
22: GF-OPT-Deliver(m)
23: **if** $G = m.src$ **then**
24: $propPending \leftarrow propPending \cup \{m\}$

25: *When* $\exists m \in propPending \wedge nextProp = k$
26: $propPending \leftarrow propPending \setminus (decided \cup \{m' : \exists m'' \in decided \wedge m'.ts < m''.ts \wedge m' \neq null\})$
27: **if** $propPending \neq \emptyset$ **then**
28: $nextProp \leftarrow k + 1$
29: Propose(k,$propPending$)

30: *When* Decide(k,$msgSet$)
31: $decided \leftarrow decided \cup msgSet$
32: **for all** $m \in msgSet : G \in m.dst \wedge m \neq null$ **do**
33: $consPending \leftarrow consPending \cup \{m\}$
34: **while** $\exists m \in msgSet : (\forall m' \in msgSet : m \neq m' \Rightarrow m.ts < m'.ts)$ **do**
35: **for all** $p' \in m.dst \setminus \{G\}$ **do**
36: send(p', $\{m, \text{`cons'}\}$) {*this message is sent through a FIFO lossless channel*}
37: $msgSet \leftarrow msgSet \setminus \{m\}$ {*the messages are sent in ascending order of timestamp*}
38: $nextProp \leftarrow k + 1$
39: $k \leftarrow k + 1$

40: *When* receive($\{m', \{\text{`cons'}\}$)
41: **if** $m' \notin consPending \wedge m' \notin delivered \wedge m' \neq null$ **then**
42: $consPending \leftarrow consPending \cup \{m\}$
43: $barrier(m'.src, G) \leftarrow max(m'.ts, barrier(m'.src, G))$ {*channels are FIFO, but there are different senders*}

44: *When* $\exists m \in consPending : \forall G' \in sendersTo(G) : m.ts < barrier(G', G)$
$\wedge \nexists m' \in consPending : m'.ts < m.ts$
45: $consPending \leftarrow consPending \setminus \{m\}$
46: GF-CONS-Deliver(m)
47: $delivered \leftarrow delivered \cup \{m\}$

lost[9]. Then, the optimistic state is overwritten with the conservative one and, from that state, all the remaining comands in Q_{opt} are re-executed, leading to a new optimistic state for that object.

References

[1] Dwork, C., Lynch, N., and Stockmeyer, L. Consensus in the presence of partial synchrony. *Journal of the ACM (JACM) 35*, 2 (1988), 288–323.

[2] Fischer, M. J., Lynch, N. A., and Paterson, M. S. Impossibility of distributed consensus with one faulty process. *J. ACM 32* (April 1985), 374–382.

[3] Lamport, L. The part-time parliament. *ACM Transactions on Computer Systems (TOCS) 16*, 2 (1998), 133–169.

[4] Rodrigues, L., and Raynal, M. Atomic broadcast in asynchronous crash-recovery distributed systems. In *Distributed Computing Systems, 2000. Proceedings. 20th International Conference on* (2000), IEEE, pp. 288–295.

[9] Also, when C_c is delivered, but it is not in Q_{opt}, the remaining possibility is the very unlikely case where the conservative delivery happened before the optimistic one. To handle this case, C_c is stored in a list of possibly delayed optimistic delivery and, if it is ever optimistically delivered, the algorithm will know that it should only discard that command, instead of updating the optimistic state.