A Virtually Synchronous Group Multicast Algorithm for WANs: Formal Approach[*]

Idit Keidar Roger Khazan
Massachusetts Institute of Technology Lab for Computer Science
545 Technology Square, Cambridge, MA 02139, USA
Email: {idish, roger}@theory.lcs.mit.edu
URL: http://theory.lcs.mit.edu/{∼idish, ∼roger}

September 28, 2000

Abstract

This paper presents a formal design for a novel group communication service targeted for WANs. The service provides Virtual Synchrony semantics. Such semantics facilitate the design of fault tolerant distributed applications. The presented design is more suitable for WANs than previously suggested ones. In particular, it features the first algorithm to achieve Virtual Synchrony semantics in a single communication round. The design also employs a scalable WAN-oriented architecture: it effectively decouples the main two components of Virtually Synchronous group communication — group membership and reliable group multicast.

The design is carried out formally and rigorously. This paper includes formal specifications of both safety and liveness properties. The algorithm is formally modeled and assertionally verified.

Subjects: 68M14 Distributed systems, 68M15 Reliability, testing and fault tolerance, 68W15 Distributed algorithms, 68Q85 Models and methods for concurrent and distributed computing.
Keywords: Group Communication, Virtual Synchrony, Reliable Multicast, Formal Modeling.

[*] This paper is an extended version of a paper, entitled "A Client-Server Approach to Virtually Synchronous Group Multicast: Specifications and Algorithms" by the same authors, that appeared in the 20th International Conference on Distributed Computing Systems (ICDCS 2000), April 2000, pages 344–355. This work was supported by Air Force Aerospace Research (OSR) grants F49620-00-1-0097 and F49620-00-1-0327, Nippon Telegraph and Telephone (NTT) grant MIT9904-12, and NSF grants CCR-9909114 and EIA-9901592.

1 Introduction

Group communication services (GCSs) [ACM96, Bir96] are powerful middleware systems that facilitate the development of fault-tolerant distributed applications. These services provide a notion of *group abstraction*, which allows application processes to easily organize themselves into multicast groups. Application processes can communicate with the members of a group by addressing messages to the group. Most GCSs strive to present different members of the same group with mutually consistent perceptions of the communication done in the group. This perception is known as *Virtual Synchrony* semantics [BJ87].

Traditionally, GCSs were designed for deployment in local area networks (LANs). Efficient GCSs that operate in wide area networks (WANs) is still an open area of research. Designing such GCSs is challenging because in WANs communication is more expensive and connectivity is less stable than in LANs.

In this paper we present a novel algorithm for a GCS targeted for WANs. The service provided by our GCS satisfies a variant of the Virtual Synchrony (VS) semantics that has been shown useful for facilitating the design of distributed applications [VKCD99, Bir96]. Our algorithm for implementing this semantics is more appropriate for WANs than the existing solutions: it requires less rounds of communication and is designed for the scalable WAN-oriented architecture of [ACDK98, KSMD00]. Our design is carried out at a very high level of formality and rigor, much higher than that of most previous designs of Virtually Synchronous GCSs. It includes formal and precise specifications, algorithms, and proofs.

The rest of this section is organized as follows: In Section 1.1 we present some basic background on GCSs and Virtual Synchrony. Section 1.2 summarizes the contributions made by our work, and Section 1.3 gives a brief overview of our design. Section 1.4 gives a roadmap to the rest of the paper.

1.1 Background

Modern distributed applications often involve large groups of geographically distributed processes that interact by sending messages over an asynchronous fault-prone network. Many of these applications maintain a replicated state of some sort. In order for these applications to be correct, the replicas must remain mutually consistent throughout the execution of the application. For example, in an online game, the states of the game maintained by different players must be mutually consistent in order for the game to be meaningful to the players. Designing algorithms that maintain state consistency is difficult however: different application processes may perceive the execution of the application inconsistently because of asynchrony and failures. For example if Alice, Bob, and Carol are playing an online game, the following asymmetric scenario is possible: Alice and Bob perceive each other as alive and well, but they differ in the way they perceive Carol; one sees Carol as crashed or disconnected, while the other sees her as alive and well. Middleware systems that hide from the application some of the underlying inconsistencies and instead present them with a more consistent picture of the distributed execution facilitate development of distributed applications.

Group communication services, such as [ADKM92, AMMS^{+}95, vRBM96, BJ87], are examples of such middleware systems. They are particularly useful for building applications that require reliable multi-point to multi-point communication among a group (or groups) of processes. Examples of such applications are data replication (for example, [KD96, ADMSM94, FLS97, KFL98, FV97], and [DP99] Ch. 7), highly-available servers (for example, [ADK99]), and online games. GCSs allow application processes to easily organize themselves into groups and to communicate with all the members of a group by addressing messages to the group. The semantics of this abstraction are such that different members of the group have consistent perceptions of the communication done in the group. The abstraction is typically implemented through the integration of two types of services: membership and reliable multicast.

Membership services maintain information about membership of groups. The membership of a group can change dynamically due to new processes joining and current members departing, failing, or disconnecting. The membership service tracks these changes and reports them to group members. The report given by the membership service to a member is called a *view*. It includes a unique identifier and a list of currently active and mutually connected members. Failures can partition a group into disconnected components of mutually connected members. Membership services strive to form and deliver the same views to all mutually connected members of the group. While this is not always possible, they typically succeed once network connectivity more or less stabilizes (see, for example, [KSMD00, VKCD99]).

In addition, GCSs provide reliable multicast services that allow application processes to send messages to the entire membership of a group. GCSs guarantee that message delivery satisfies certain properties. For example, one property can be that messages sent by the same sender are delivered in the order in which they were sent. Different GCSs differ in the specific message delivery properties they provide, but most of them provide some variant of *Virtual Synchrony* semantics. We refer to a GCS providing such semantics as a *Virtually Synchronous GCS*, and to an algorithm implementing this semantics as a *Virtual Synchrony algorithm.*

Virtual Synchrony semantics specifies how message deliveries are synchronized with view deliveries. This synchronization is done in a way that simulates a "benign" world in which message delivery is reliable within each view. Many variants of Virtual Synchrony have been suggested (for example, [MAMSA94, FvR95, VKCD99, BJ87, SR93, BDM98]). Nearly all of them include a key property, called *Virtually-Synchronous Delivery*, which guarantees that *processes that receive the same pair of views from the GCS receive the same sets of messages in between receiving the views.* Henceforth, when we refer to Virtual Synchrony, we assume the semantics includes Virtually-Synchronous Delivery.

Example 1.1 *Assume Alice, Bob, and Carol are playing an online game. Each of them is initially given a view* $\langle\{Alice,\ Bob,\ Carol\}, 1\rangle$*, where* $\{Alice,\ Bob,\ Carol\}$ *is a set of members and* 1 *is a view id. Then Carol disconnects, and Alice and Bob are given a new view* $\langle\{Alice,\ Bob\}, 2\rangle$*. The Virtually-Synchronous Delivery property guarantees that both Alice and Bob receive the same messages before receiving the new view. In particular, if Bob receives a message from Carol before it receives the new view, then Alice also receives this message before the new view. Therefore, if Alice and Bob modify their game states only when they receive messages, they remain in consistent states and can safely continue playing the game after they receive the new view.*

In general, Virtually Synchronous GCSs are especially useful for building applications that maintain a replicated state of some sort using a variant of the well-known *state-machine/active replication* approach [Lam78, Sch90]. With such approach, processes that maintain state replicas are organized into multicast groups. Actions that update the state are sent using a multicast primitive that delivers messages to different processes in the same order. When processes receive these actions, they apply them to their local replicas. Virtual Synchrony guarantees that processes that remain connected receive the same messages. This implies that processes that remain connected apply the same sequences of actions to their replicas. Hence, their replicas remain mutually consistent. Examples of GCS applications that use this technique are [ABCD96, ADMSM94, KD96, SM98, FV97, ADK99].

Let us consider what is involved in implementing the Virtually-Synchronous Delivery property. Imagine that GCS processes are forming a new view because someone has disconnected from their current view. The GCS processes must make sure that they deliver the same messages to their application clients before delivering to them the new view. However, it may be the case that some of these GCS processes received messages that others did not. In the scenario illustrated in Example 1.1, the last messages from Carol may have reached the GCS process of only Bob, and not of Alice; Bob and Alice need to agree on whether or not to deliver these messages. To ensure such agreement, GCS processes invoke a *synchronization* protocol whenever a new view is forming.

Designing correct and efficient algorithms that implement Virtual Synchrony is not trivial. Different GCS processes may perceive connectivity changes inconsistently. Since the desired synchronization depends on who the members of the new view are, the algorithm has to tolerate transient inconsistent views and cascading connectivity changes.

In particular, a Virtual Synchrony algorithm needs to know which synchronization messages sent by different processes pertain to the same view formation attempt. Existing algorithms, such as [FvR95, ADKM92, SR93, BDM98, GVvR96, AMMS$^+$95], identify such synchronization messages by tagging them with a common identifier. Some initial communication is performed first, before synchronization messages are communicated, in order to agree upon a common identifier and to distribute it to the members of the forming view.

While a view is forming and a synchronization protocol is executing, there may be changes in connectivity that call for views with altogether different memberships. When such situations happen, existing Virtual Synchrony algorithms, for example [FvR95, GVvR96, SR93, BDM98, AMMS$^+$95], continue executing their current synchronization protocol to termination and then deliver to the application a view that does not reflect the already detected changes in connectivity. Afterwards, the algorithm is invoked anew to incorporate the new changes.

We refer to a view as *obsolete* [KSMD00] when it is delivered by a GCS even though the GCS already has information that the view's membership has changed. Obsolete views cause an overhead not just for the GCS, but also for applications. Since application processes do not know when the views delivered to them are obsolete, they handle such views just as they do any other view, for example by running state-synchronization protocols [KD96, FLS97, KFL98].

In a WAN, connectivity changes tend to occur frequently, message latency tends to be high and unpredictable, and message loss is not uncommon. Therefore, WANs call for algorithms that execute fewer communication rounds and respond to connectivity changes promptly, without wasting resources on handling obsolete memberships.

1.2 Our Contributions

In this paper, we present a novel design for a Virtually Synchronous GCS targeted for WANs. We make the following contributions:

- We present a new algorithm for implementing Virtual Synchrony. Our algorithm is more efficient than existing ones. It neither processes nor delivers views with obsolete memberships. Moreover, the synchronization protocol run by our algorithm involves just a single message exchange round among members of the new view. We are not aware of any other algorithm for implementing Virtual Synchrony that has these two features.

- Our design demonstrates how to more effectively decouple the algorithm for achieving Virtual Synchrony from the algorithm for maintaining membership. As suggested in [ACDK98, KSMD00], such efficient decoupling is important for providing scalable GCSs in WANs. Existing designs typically have a single algorithm handling both Virtual Synchrony and membership. The few designs that do employ two separate algorithms [SR93, BFHR98] still have the two algorithms tightly coupled. In particular, the Virtual Synchrony algorithms control the membership algorithms: the membership algorithms are not allowed to incorporate newly joining members while the synchronization protocols are running. Moreover, the communication between the two algorithms is in two directions. In contrast, our design allows the membership algorithm to freely change memberships of forming views at any time. The interaction between the membership and Virtual Synchrony algorithms is only in one direction, from the former to the latter, and it has low overhead. The decoupling is such that the synchronization protocol can execute in parallel with the view formation protocol.

- Our design is carried out much more rigorously and formally than most previous designs of Virtually Synchronous GCSs. The presented specifications of our GCS and its environment, description of the algorithm, and proof of correctness are all precise and formal. Our project is the first to use formal methods for modeling a Virtually Synchronous GCS and to provide an assertional proof of its correctness.

  In order to manage the complexity of the design, we have developed a novel, inheritance-based methodology [KKLS00]. This methodology allows for incremental construction of formal specifications, algorithms, and, very importantly, proofs. In addition to making the design tractable, the use of this methodology makes it evident which part of the algorithm implements which property.

We now discuss each of the different aspects of our design in more detail.

1.3 Design Overview

The novelty of our algorithm for achieving Virtual Synchrony is concentrated in its synchronization protocol. Recall that this protocol is run among GCS processes in order for those that remain connected to agree upon a common set of messages each of them must deliver before moving into the new view. The protocol depends on a simple, yet powerful idea. Instead of using common identifiers to designate which synchronization messages pertain to the same view formation attempt, we use locally generated identifiers. These identifiers are then included as part of the formed views[1]. Once a view formation completes at a GCS process, the process knows which synchronization messages of other members to consider for the view – the messages tagged with the identifiers that are included in the view.

Example 1.2 *View* $\langle\{Alice, Bob, Carol\}, [4,3,7], 8\}\rangle$ *has membership* $\{Alice, Bob, Carol\}$*, vector of local identifiers* $[4,3,7]$*, and view identifier* 8*. When a GCS process forms this view, it uses the synchronization messages from Alice, Bob, and Carol tagged respectively with* 4*,* 3*, and* 7 *to decide on the set of messages it must deliver before delivering this view to its application. Thus, if Alice, Bob, and Carol form the same view, they use the same synchronization messages, and thus agree on which application messages each of them needs to deliver.*

The use of local identifiers eliminates the need to pre-agree on common identifiers and allows the synchronization protocol to complete in a single message exchange round. It also allows the

[1] A similar view structure is suggested in [SR93], for the purpose of not having concurrent views intersect.

algorithm to promptly and efficiently react to connectivity changes, without wasting resources on obsolete views. The protocol works correctly even if, because of network instability, GCS processes send multiple synchronization messages during the same synchronization protocol.

Our design decouples the algorithm for implementing Virtually Synchronous multicast from the algorithm for maintaining membership. The membership algorithm handles generation of local identifiers and formation of views. The algorithm for implementing Virtually Synchronous multicast synchronizes views and application messages to implement the Virtual Synchrony semantics. In particular, it handles multicast requests submitted by the application, delivers application messages and views back to the application, and runs the synchronization protocol to synchronize processes that transition together into new views. The decoupling involves low-cost, one-directional communication from the membership to the Virtually Synchronous multicast algorithm. It also allows the synchronization protocol to execute in parallel with the view formation protocol.

Efficient decoupling of membership and Virtually Synchronous multicast algorithms allows for an architecture in which the membership service is implemented by a small set of dedicated membership servers maintaining the membership information on behalf of a large set of clients. This architecture was proposed in [ACDK98, KSMD00] for supporting scalable membership services in WANs. Our work extends this architecture by specifying how it can be used as a base for a Virtually Synchronous GCS. In particular, we present precise specifications of the interface and semantics that a membership service has to provide in order to be decoupled from the Virtually Synchronous multicast algorithm.

Our interface and membership service specifications allow for straightforward and efficient membership and Virtually Synchronous multicast algorithms. The Virtually Synchronous multicast algorithm presented in this paper is an example of such an algorithm: the synchronization protocol requires a single message exchange round, which can occur in parallel with the formation of the view. The algorithm has been implemented (in C++) [Tar00] using the scalable one-round membership algorithm of [KSMD00]. This membership algorithm was specifically tailored for our design, but other existing membership algorithms (for example, [DMS94, AMMS$^+$95]) can be also easily extended to provide the required interface and semantics.

Our design has been carried out and is presented at a level more formal and rigorous than that of most previous designs of Virtually Synchronous GCSs. We precisely specify the properties satisfied by our Virtually Synchronous multicast algorithm, the external membership service, and the underlying communication substrate. We then give a formal description of the Virtually Synchronous multicast algorithm. The algorithm is accompanied by a careful formal correctness proof. The safety properties are proved by using invariant assertions and simulation mappings; the liveness properties are proved by using invariant assertions and careful operational arguments. We found this level of rigor to be important: in the process of specifying and verifying the algorithm, we uncovered several ambiguities and errors.

Previously, formal approaches were used to specify the semantics of Virtually Synchronous GCSs and to model and verify their applications, for example, in [Cho97, FLS97, DPFLS98, KFL98, DPFLS99, HLvR99]. Existing algorithms implementing Virtual Synchrony are modeled in pseudocode and proven correct operationally. However, due to their size and complexity, such algorithms were not previously modeled using formal methods nor were they assertionally verified.

To manage the complexity of this project we have developed a formal inheritance-based methodology [KKLS00] for incrementally constructing specifications, algorithms, and proofs. In addition to making the project tractable, the use of this construct makes clear which parts of the algorithm implement which property. The modularity of this approach facilitates further modifications and alterations of the design. Our project and the inheritance-based construct are both developed in the framework of the I/O automaton formalism (see [LT89] and [Lyn96], Ch. 8).

1.4 Roadmap

The rest of this paper is organized as follows: Section 2 reviews the formal model and notation. In Section 3 we present the client-server architecture of our GCS and formally specify the assumptions we make on the membership service and the underlying communication substrate. Section 4 contains precise specifications of the safety and liveness properties satisfied by our GCS. The algorithm is then given in Section 5 and is accompanied by informal correctness arguments. Section 6 concludes the paper. A formal correctness proof that the algorithm of Section 5 satisfies the specifications of Section 4 is given in Appendix: safety properties – in B, and liveness properties – in C. Appendix A reviews the proof techniques used in Appendices B and C.

2 Formal Model and Notation

In the I/O automaton model (cf. [LT89] and [Lyn96], Ch. 8), a system component is described as a state-machine, called an *I/O automaton.* The transitions of this state-machine are associated with named actions, which are classified as either *input*, *output*, or *internal.* Input and output actions model the component's interaction with other components, while internal actions are externally-unobservable.

Formally, an I/O automaton is defined as the following five-tuple: a signature (input, output and internal actions), a set of states, a set of start states, a state-transition relation (a cross-product between states, actions, and states), and a partition of output and internal actions into *tasks.* Tasks are used for defining fairness conditions.

An action π is said to be *enabled* in a state $\mathtt{s}$ if the automaton has a transition of the form ($\mathtt{s}$, π, $\mathtt{s}'$); input actions are enabled in every state. An *execution* of an automaton is an alternating sequence of states and actions that begins with its start state and in which every action is enabled in the preceding state. An infinite execution is *fair* if, for each task, it either contains infinitely many actions from this task or infinitely many occurrences of states in which no action from this task is enabled; a finite execution is *fair* if no action is enabled in its final state. A *trace* is a subsequence of an execution consisting solely of the automaton's external actions. A *fair trace* is a trace of a fair execution.

When reasoning about an automaton, we are interested in only its externally-observable behavior as reflected in its traces. There are two types of trace properties: *safety* and *liveness.* Safety properties usually specify that some particular bad thing never happens. In this paper we specify safety properties using centralized, global, I/O automata that generate the legal sets of traces; for such automata we do not specify task partitions. Each external action in such a centralized automaton is tagged with a subscript which denotes the process at which this action occurs. An algorithm automaton *satisfies* a specification if all of its traces are also traces of the specification automaton. Refinement mappings are a commonly used technique for proving trace inclusion, in which one automaton (the algorithm) *simulates* the behavior of another automaton (the specification). Refinement mappings and other related proof techniques are reviewed in Appendix A. Liveness properties usually specify that some good thing eventually happens. An algorithm automaton satisfies a liveness property if the property holds in all of its *fair* traces.

The *composition operation* defines how automata interact via their input and output actions: It matches output and input actions with the same name in different component automata; when a component automaton performs a step involving an output action, so do all components that have this action as an input one. When reasoning about a certain system component, we compose it with abstract specification automata that specify the behavior of its environment.

I/O automata are conveniently presented using the *precondition-effect* style: In this style, typed state variables with initial values specify the set of states and the start states. A variable type is a set; if S is a set, the notation $S_\perp$ refers to the set $S \cup \{\perp\}$. Transitions are grouped by action name, and are specified as a list of triples consisting of an action name, possibly with parameters, a `pre` : block with preconditions on the states in which the action is enabled, and an `eff` : block which specifies how the pre-state is modified *atomically* to yield the post-state.

We have developed a novel formal notion of inheritance for automata [KKLS00]. A *child* automaton is specified as a modification of the parent automaton's code. When presenting a child we first specify a *signature extension* which consists of new actions, labeled new, and modified actions. A modified action is labeled with the name of the action which it modifies as follows: modifies `parent.action(parameters)`). We next specify the *state extension* consisting of new state variables added by the child. Finally, we describe the *transition restriction* which consists of new preconditions and effects added by the child to both new and modified actions. For modified actions, the preconditions and effects of the parent are appended to those added by the child. New effects added by the child are performed before the effects of the parent, all of them in a single atomic step. The child's effects are not allowed to modify state variables of the parent. This ensures that the set of traces of the child, when projected onto the parent's signature, is a subset of the parent's set of traces [KKLS00].

Inheritance allows us to reuse code and avoid redundancies. It also allows us to reuse proofs: Assume that an algorithm automaton $\mathtt{A}$ can simulate a specification automaton $\mathtt{S}$, and let $\mathtt{A}'$ and $\mathtt{S}'$ be child automata of $\mathtt{A}$ and $\mathtt{S}$, respectively. Then the Proof Extension theorem of [KKLS00] asserts that in order to prove that $\mathtt{A}'$ can simulate $\mathtt{S}'$ it is sufficient to show that the restrictions added by $\mathtt{A}'$ are consistent with the restrictions $\mathtt{S}'$ places on $\mathtt{S}$, and that the new functionality of $\mathtt{A}'$ can simulate new functionality of $\mathtt{S}'$. Appendix A contains more details.

3 Client-Server Architecture and Environment Specification

Our service is designed to operate in an asynchronous message-passing environment. Processes and communication links may fail and may later recover, possibly causing network partitions and merges. For simplicity, we assume that processes recover with their running state intact; this is a plausible assumption as processes can keep their running state on stable storage. We do not explicitly model process crashes and recoveries because under this assumption a crashed process is indistinguishable from a slow one. In Section 5.4, we argue that our algorithm also provides meaningful semantics when group communication processes lose their entire state upon a crash and recover with their state reset to an initial value.

Our Group Communication service is implemented by a collection of GCS *end-points*, which are the GCS processes that run at the application clients' locations. GCS end-points handle clients' multicast requests and inform their clients of view changes.

The GCS architecture is depicted in Figure 1. All GCS end-points run the same algorithm. The algorithm relies on the underlying membership and multicast services to handle respectively formation of views and transmission of messages. The algorithm's task is to synchronize output of the two underlying services to implement the Virtual Synchrony semantics.

Sections 3.1 and 3.2 below give precise specifications of the interface and semantics that the underlying membership and multicast services have to provide in order to be suitable for our algorithm. Services that satisfy these (or very similar) requirements have been previously used for

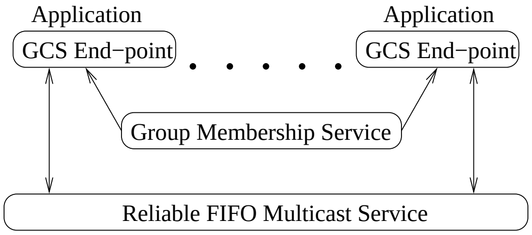


Figure 1: The client-server architecture: GCS end-points using an external membership service. Arrows represent interaction between GCS end-points and underlying services.

GCSs, and efficient implementations of these services for WANs exist.

3.1 The membership service specification

This section presents a formal specification of the membership services that are appropriate for our GCS design. For simplicity, here and in the rest of the paper, we assume that there is a single process group; multiple groups can be supported by treating each independently. We also omit part of the interface that handles processes' requests to join and leave groups.

Figure 2 contains an I/O automaton, called MBRSHP, that defines the interface and the safety properties of the membership service. The service interface is given by the automaton's signature; Informally, it consists of the following two output actions:

$\texttt{start\_change}_{\texttt{p}}(\texttt{cid}, \texttt{set})$ notifies process p that the membership service is attempting to form a view with the members of set; cid is a local start-change identifier.

$\texttt{view}_{\texttt{p}}(\texttt{v})$ notifies process p that the membership service has succeeded in forming view v. A view v is a triple consisting of an identifier v.id, a set of members v.set, and a function v.startId that maps members of v to start-change identifiers. Two views are the same if they consist of identical triples.

Automaton MBRSHP maintains two state variables, mbrshp_view[p] and start_change[p], for each client p. These variables contain respectively the last view and the last start_change message issued to client p; the variables are updated in the effects of the transitions. The safety properties satisfied by the MBRSHP automaton include two basic properties, which are provided by virtually all group membership services (for example, [BvR94, DMS94, AMMS$^+$95, FvR95, BDM98, KSMD00, SR93, ADKM92]), as well as some new properties concerning the start_change notifications.

The two basic properties are *Self Inclusion* and *Local Monotonicity.* Self Inclusion requires every view issued to a client p to include p as a member; this property is enforced with a precondition $\texttt{p} \in \texttt{v.set}$ on the $\texttt{view}_{\texttt{p}}(\texttt{v})$ action. Local Monotonicity requires that view identifiers delivered to p be monotonically increasing; this property is enforced with a precondition $\texttt{v.id} > \texttt{mbrshp\_view}[\texttt{p}]$ on the $\texttt{view}_{\texttt{p}}(\texttt{v})$ action. Local Monotonicity has two important consequences: the same view is not delivered more than once to the same client, and clients that receive the same two views receive them in the same order [VKCD99].

In addition, the MBRSHP automaton specifies that the membership service must issue at least one start_change notification to client p before issuing a new view v to p. Also, the start-change

```
AUTOMATON MBRSHP

Type:
    Proc: Set of end-points.
    StartChangeId: Total-order; cid_0 is smallest.
    ViewId: Partial-order; vid_0 is smallest.
    View: ViewId × SetOf(Proc) × (Proc → StartChangeId).
Def:  v_p = ⟨vid_0, {p}, {(p →cid_0)}⟩.

Signature:
 Output: start_change_p(cid, set), Proc p, StartChangeId cid, SetOf(Proc) set
         view_p(v), Proc p, View v

State:
 For all Proc p: View  mbrshp_view[p], initially v_p
 For all Proc p: (StartChangeId × SetOf(Proc)) start_change[p], initially ⟨cid_0, {}⟩

Transitions:
 OUTPUT  start_change_p(cid, set)
 pre: cid > mbrshp_view[p].startId(p)
      cid ≥ start_change[p].id
      p ∈ set
 eff: start_change[p] ← ⟨cid, set⟩

 OUTPUT  view_p(v)
 pre: p ∈ v.set ∧ v.id > mbrshp_view[p].id
      v.set ⊆ start_change[p].set
      v.startId(p) = start_change[p].id
      v.startId(p) > mbrshp_view[p].startId(p)
 eff: mbrshp_view[p] ← v
```

Figure 2: Membership service interface and safety specification.

identifier $\mathtt{v.startId(p)}$ contained in the new view $\mathtt{v}$ must be the same as the identifier of the latest preceding $\mathtt{start\_change}$ issued to $\mathtt{p}$. These two requirements are enforced by the last two preconditions on $\mathtt{view}_\mathtt{p}(\mathtt{v})$. In particular, the former one is achieved by requiring that a bigger start-change identifier than the one associated with $\mathtt{p}$ in the last view has been issued to $\mathtt{p}$.

The MBRSHP specification allows the membership service to react to connectivity changes happening during view formation. Whenever the service wants to add new members to the membership, it has to issue a new $\mathtt{start\_change}$ notification to the clients: the second precondition on $\mathtt{view}_\mathtt{p}(\mathtt{v})$ actions requires the membership $\mathtt{v.set}$ to be a subset of the tentative membership $\mathtt{set}$ included in the last $\mathtt{start\_change}$ notification. In order to remove members from a forming view, the service does not need to issue a new $\mathtt{start\_change}$ notification.

The first $\mathtt{start\_change}$ notification issued to $\mathtt{p}$ after a view marks the beginning of a new view formation period. It includes a new local identifier $\mathtt{cid}$, different from the ones that were previously sent to $\mathtt{p}$: the first precondition on $\mathtt{start\_change}_\mathtt{p}(\mathtt{cid}, \mathtt{set})$ requires $\mathtt{cid}$ to be strictly greater than $\mathtt{mbrshp\_view[p].startId(p)}$. Subsequent $\mathtt{start\_change}$ notifications sent during an on-going view formation may either reuse the last start-change identifier or issue a new one, as specified by the second precondition on $\mathtt{start\_change}$ actions. We ensure uniqueness of local start-change identifiers by generating them in increasing order.

Notice that the MBRSHP automaton does not specify any relationship between views issued to different clients.

Example 3.1 *Figure 3 presents a sample execution that shows the* MBRSHP *service delivering different sequences of views to two different clients,* $\mathtt{a}$ *and* $\mathtt{b}$*. Arrows represent time passage at each client; gray dots represent events. First, both clients receive the same view* $\mathtt{v} = \langle 2, \{\mathtt{a}, \mathtt{b}\}, [\mathtt{a} : 1, \mathtt{b} : 1]\rangle$*; we illustrate this with a circle around the view events at both clients. Then, client* $\mathtt{b}$ *receives a view* $\mathtt{v}_{\mathtt{mid}} = \langle 3, \{\mathtt{b}\}, [\mathtt{b} : 2]\rangle$ *by itself. Then, both clients receive another common view* $\mathtt{v}' =$

$\langle 4, \{\mathtt{a},\mathtt{b}\}, [\mathtt{a}:2, \mathtt{b}:3]\rangle$. *Notice how the start-change identifiers included in the views correspond to the last start-change identifiers issued to the clients.*

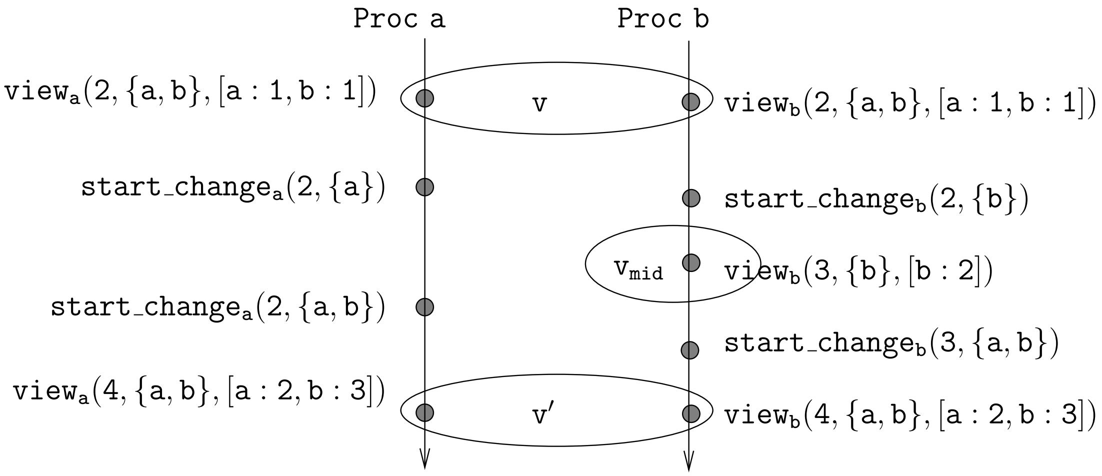

Proc a
Proc b
view$_a$(2, {a, b}, [a : 1, b : 1])
v
view$_b$(2, {a, b}, [a : 1, b : 1])
start_change$_a$(2, {a})
start_change$_b$(2, {b})
v$_{mid}$
view$_b$(3, {b}, [b : 2])
start_change$_a$(2, {a, b})
start_change$_b$(3, {a, b})
view$_a$(4, {a, b}, [a : 2, b : 3])
v′
view$_b$(4, {a, b}, [a : 2, b : 3])

Figure 3: A sample execution of MBRSHP.

We do *not* specify liveness properties for membership services. Instead, when we specify the liveness properties of our GCS in Section 4.2, we *condition* them on the behavior of the membership service. For example, we state that if the same view is delivered to all the members and the members do not receive any subsequent membership events, then they eventually deliver this view to their application clients. Existing membership services do satisfy meaningful liveness properties. For example, [KSMD00] guarantees that, when network stabilizes, all members receive "correct" view and no other views thereafter. By combining our GCS liveness properties with such membership liveness properties, we can restate the liveness properties of our GCS conditionally on the network behavior.

The MBRSHP specification allows for simple and efficient distributed implementations that also satisfy meaningful liveness properties. In our implementation [Tar00], we use the service of [KSMD00], where a small number of servers support a large number of clients, communicating with them asynchronously via FIFO ordered channels (TCP sockets). In case a server fails, clients can migrate to another server. Other existing membership algorithms (for example, [DMS94, AMMS$^+$95]) could also be extended easily to provide the specified here interface and semantics.

3.2 The reliable FIFO multicast service specification

The group communication end-points communicate with each other using an underlying multicast service that provides reliable FIFO communication between every pair of connected processes. Many existing group communication systems (for example, [GVvR96, BDM98, DMS94, ADKM92]) implement Virtual Synchrony over similar communication substrates. In our implementation [Tar00], we use the service of [ACSD00].

Figure 4 presents an I/O automaton, CO_RFIFO, that specifies a multicast service appropriate for our GCS design. Portions of the code that define liveness properties are colored gray.

Automaton CO_RFIFO maintains a FIFO queue channel[p][q] for every pair of end-points. An input action send$_p$(set, m) models a multicast of message m from end-point p to the end-points

```
AUTOMATON CO_RFIFO

Signature:
 Input:                                                  Output:   deliver_p,q(m), Proc p, Proc q, Msg m
   send_p(set,m), Proc p, SetOf(Proc) set, Msg m         Internal: lose(p,q), Proc p, Proc q
   reliable_p(set), Proc p, SetOf(Proc) set                        skip_task(p,q), Proc p, Proc q
   live_p(set), Proc p, SetOf(Proc) set

State:
  For all Proc p, Proc q: SequenceOf(Msg) channel[p][q], initially empty
  For all Proc p: SetOf(Proc) reliable_set[p], initially {p}
  For all Proc p: SetOf(Proc) live_set[p], initially {p}

Transitions:
  INPUT  send_p(set, m)                                  INTERNAL  lose(p, q)
  eff: (∀ q ∈ set) append m to channel[p][q]             pre: q ∉ reliable_set[p]
                                                         eff: dequeue last message from channel[p][q]
  OUTPUT  deliver_p,q(m)
  pre: m = first(channel[p][q])                          INPUT live_p(set)
  eff: dequeue m from channel[p][q]                      eff: live_set[p] ← set

  INPUT  reliable_p(set)                                 INTERNAL skip_task(p, q)
  eff: reliable_set[p] ←  set                            pre: q ∉ live_set[p]

Tasks:
  For each Proc p, Proq q:  C_p,q = ({deliver_p,q(m) | m ∈ Msg} ∪ {skip_task(p,q)} ∪ {lose(p,q)})
```

Figure 4: Reliable FIFO multicast service specification. Liveness-related code is colored gray.

listed in the set by appending m to the channel[p][q] queues for every end-point q in set. The $\mathtt{deliver}_{\mathtt{p,q}}(\mathtt{m})$ action removes the first message from channel[p][q] and delivers it to q.

In addition, the CO_RFIFO specification allows an end-point p to use the $\mathtt{reliable}_{\mathtt{p}}(\mathtt{set})$ action to require that the multicast service maintain a reliable (gap-free) FIFO connection to the end-points listed in set. Whenever this action occurs, set is stored in a special variable reliable_set[p]. For every process q not in reliable_set[p], the multicast service may lose an arbitrary suffix of the messages sent from p to q, as modeled by an internal action lose(p, q).

In order for the multicast service to be considered live, messages sent to live and connected processes must eventually reach their destinations. The CO_RFIFO specification enforces this property in the gray-colored portion of its code.

Recall from Section 2 that an infinite fair execution of an automaton must contain either infinitely many events from each task C or infinitely many occurrences of states in which no action in C is enabled. Automaton CO_RFIFO defines the set $\mathtt{C}_{\mathtt{p,q}} = (\{\mathtt{deliver}_{\mathtt{p,q}} \mid \mathtt{m} \in \mathtt{Msg}\} \cup \mathtt{skip\_task}(\mathtt{p,q}) \cup \mathtt{lose}(\mathtt{p,q}))$ to be a task for each pair of end-points p and q. This definition implies that $\mathtt{deliver}_{\mathtt{p,q}}$ actions must occur in an infinite fair execution of CO_RFIFO, provided the following three conditions hold: there are messages sent from p to q – hence, $\mathtt{deliver}_{\mathtt{p,q}}$ is enabled; the client at p is interested in maintaining reliable connection to q – hence, lose(p, q) is disabled; and q is believed to be lively connected to p – hence, a special action skip_task(p, q) is disabled, as explained below.

Action skip_task(p, q) is defined only to provide an alternative to $\mathtt{deliver}_{\mathtt{p,q}}$ actions so that $\mathtt{deliver}_{\mathtt{p,q}}$ actions are not required to happen when q is believed to be disconnected from p. skip_task(p, q) is an internal action that has no effect on the state of CO_RFIFO and is enabled when q is believed to be disconnected from p. Such belief is modeled using special $\mathtt{live}_{\mathtt{p}}(\mathtt{set})$ input actions. The `set` argument is assumed to represent a set of processes that are alive and connected to `p`; when such an input happens, `set` is stored in a state variable `live_set`[`p`]. The precondition on the `skip_task`(`p`, `q`) action is $q \notin$ `live_set`[`p`].

An important implication of how tasks are defined in CO_RFIFO is that, if `q` remains in both `live_set`[`p`] and `reliable_set`[`p`] from some point on in a fair execution of CO_RFIFO, then all the messages that `p` sends to `q` from that point on are eventually delivered to `q`.

4 Specifications of the Group Communication Service

The next two subsections contain specifications of the safety and liveness properties satisfied by our group communication service. These properties have been shown useful for many distributed applications (see [VKCD99]).

4.1 Safety properties

We present the safety specification of our group communication service incrementally, as four automata: In Section 4.1.1 we specify a simple group communication service that synchronizes delivery of views and application messages to require *Within-View Delivery* of messages. In Section 4.1.2 we extend the specification of Section 4.1.1 to also require *Virtually-Synchronous Delivery*, the key property of Virtual Synchrony (see Section 1.1). In Section 4.1.3 we specify the *Transitional Set* property, which complements Virtually-Synchronous Delivery. Finally, in Section 4.1.4, we specify the *Self Delivery* property, which requires the GCS to deliver to each client the client's own messages.

The incremental development of the safety specification is matched later when we develop the algorithm and its correctness proof in Section 5 and Appendix B.

4.1.1 Within-View reliable FIFO multicast

In this section we specify a GCS that captures the following properties:

- Views delivered to the application satisfy the Self Inclusion and Local Monotonicity properties of the MBRSHP service, see Section 3.1.
- Messages are delivered in the same view in which they were sent. This property is useful for many applications (see [FvR95, VKCD99, SM98]) and appears in several systems and specifications (for example, [BvR94, vRBM96, AMMS$^+$95, MAMSA94, FLS97, HS95, DPFLS98]). A weaker property that requires each message to be delivered in the same view at every process that delivers it, but not necessarily the view in which it was sent, is typically implemented on top of an implementation of Within-View Delivery (see [VKCD99]).
- Messages are delivered in gap-free FIFO order (within views). This is a basic property upon which one can build services with stronger ordering guarantees, such as causal order or total order. The totally ordered multicast algorithm of [CHD98] is implemented atop a service with a similar specification.

Figure 5 presents automaton WV_RFIFO : SPEC that models this specification. The automaton uses centralized queues `msgs`[`p`][`v`] of application messages for each sender `p` and view `v`. It also maintains a variable `current_view`[`p`] that contains the last view delivered to each process `p`, and a variable `last_dlvrd`[`q`][`p`], for every pair of processes `q` and `p`, containing the index in the `msgs`[`q`][`current_view`[`p`]] queue of the last `q`'s message delivered to `p` in `p`'s current view.

```
AUTOMATON WV_RFIFO : SPEC

Signature:
 Input:  send_p(m), Proc p, AppMsg m
 Output: deliver_p(q, m), Proc p, Proc q, AppMsg m
         view_p(v), Proc p, View v

State:
 For all Proc p, View v: SequenceOf(AppMsg) msgs[p][v], initially empty
 For all Proc p, Proc q: Int last_dlvrd[p][q], initially 0
 For all Proc p: View current_view[p], initially v_p

Transitions:
 INPUT  send_p(m)
 eff: append m to msgs[p][current_view[p]]

 OUTPUT  deliver_p(q, m)
 pre: m = msgs[q][current_view[p]][last_dlvrd[q][p]+1]
 eff: last_dlvrd[q][p] ← last_dlvrd[q][p]+1

 OUTPUT  view_p(v)
 pre: p ∈ v.set ∧ v.id > current_view[p].id
 eff: (∀ q) last_dlvrd[q][p] ← 0
      current_view[p] ← v
```

Figure 5: WV_RFIFO service specification.

Action $\mathtt{view}_\mathtt{p}(\mathtt{v})$ models the delivery of view v to process p; the precondition on this action enforces Self Inclusion and Local Monotonicity. Action $\mathtt{send}_\mathtt{p}(\mathtt{m})$ models the multicast of message m from process p to the members of p's current view by appending m to msgs[p][current_view[p]]. Action $\mathtt{deliver}_\mathtt{p}(\mathtt{q}, \mathtt{m})$ models the delivery to process p of message m sent by process q. The gap-free FIFO ordered delivery of messages within-views is enforced by its precondition, which allows delivery of only the message indexed by last_dlvrd[q][p] + 1 in the msgs[q][current_view[p]] queue.

4.1.2 Virtually-Synchronous delivery

In this section we use the inheritance-based methodology to modify the WV_RFIFO : SPEC automaton to also enforce the *Virtually-Synchronous Delivery* property. The modified automaton, VSRFIFO : SPEC is defined by the code contained in both Figures 5 and 6.

```
AUTOMATON VS_RFIFO : SPEC    MODIFIES  WV_RFIFO : SPEC

Signature Extension:
 Output:   view_p(v) modifies wv_rfifo.view_p(v)
 Internal: set_cut(v, v′, c), View v, View v′, (Proc → Int)_⊥ c new

State Extension:
 For all View v, v′: (Proc→Int)_⊥ cut[v][v′], initially ⊥

Transition Restriction:
 OUTPUT  view_p(v)
 pre: cut[current_view[p]][v] ≠ ⊥
      (∀ q) last_dlvrd[q][p]=cut[current_view[p]][v](q)

 INTERNAL  set_cut(v, v′, c)
 pre: cut[v][v′] = ⊥
 eff: cut[v][v′] ← c
```

Figure 6: VS_RFIFO service specification.

Figure 6 contains the code that enforces the *Virtually-Synchronous Delivery* property. Recall from Section 1.1 that this property requires processes moving together from view v to view v′ to deliver same set of messages while in view $\mathtt{v}$. Since the parent specification, WV_RFIFO : SPEC, imposes gap-free FIFO delivery of messages, a message set can be represented by a set of indices, each pointing to the last message from each member of $\mathtt{v}$; such representation of a set is called a *cut*.

The WV_RFIFO : SPEC automaton fixes a cut for processes that wish to move from some view $\mathtt{v}$ to some view $\mathtt{v}'$: A new internal action $\mathtt{set\_cut}(\mathtt{v}, \mathtt{v}', \mathtt{c})$ sets a new variable $\mathtt{cut}[\mathtt{v}][\mathtt{v}']$ to a cut mapping $\mathtt{c}$. For a given pair of views, $\mathtt{v}$ and $\mathtt{v}'$, the cut is chosen only once, *nondeterministically.* Delivery of a view $\mathtt{v}$ to process $\mathtt{p}$ is allowed only if a cut for moving from $\mathtt{p}$'s current view into $\mathtt{v}$ has been set and if $\mathtt{p}$ has delivered all the messages identified in this cut. These conditions are enforced by the two new preconditions of the $\mathtt{view}_{\mathtt{p}}(\mathtt{v})$ action (see Figure 6). Since VSRFIFO : SPEC is a modification of WV_RFIFO : SPEC the new preconditions work in conjunction with the preconditions in $\mathtt{view}_{\mathtt{p}}(\mathtt{v})$ of WV_RFIFO : SPEC.

The VSRFIFO : SPEC automaton, being a safety specification, does not require liveness properties to hold, such as, that processes actually deliver messages specified by the cuts, and hence, are able to satisfy conditions for delivering new views. Such liveness specifications are stated in Section 4.2.

4.1.3 Transitional Set

While Virtually-Synchronous Delivery is a useful property, a process that moves from view $\mathtt{v}$ to view $\mathtt{v}'$ cannot tell locally which of the processes in $\mathtt{v.set} \cap \mathtt{v}'\mathtt{.set}$ move to view $\mathtt{v}'$ directly from view $\mathtt{v}$, and which move to $\mathtt{v}'$ from some other view. In order for the application to be able to exploit the Virtually-Synchronous Delivery property, application processes need to be informed which other processes move together with them from their current view into their new view. The set of processes that transition together from one view into the next is called a *transitional set* [VKCD99]:

Definition 4.1 *A transitional set from view* $\mathtt{v}$ *to view* $\mathtt{v}'$*, is a subset of* $\mathtt{v.set} \cap \mathtt{v}'\mathtt{.set}$ *that includes: (a) all processes that receive view* $\mathtt{v}'$ *while in view* $\mathtt{v}$*; and (b) no process that receive view* $\mathtt{v}'$ *while in a view other than* $\mathtt{v}$*.*

The notion of a transitional set was first introduced as part of a special transitional view in the EVS [MAMSA94] model. In our formulation (as in [VKCD99]), transitional sets are delivered to the application along with views, as an additional parameter $\mathtt{T}$.

Example 4.1 *Assume that Alice and Bob are using a Virtually Synchronous GCS that eventually reports the views produced by the* MBRSHP *service to Alice and Bob. Consider the scenario described in Example 3.1: both Alice and Bob receive views* $\mathtt{v}$ *and* $\mathtt{v}'$ *with the membership* $\{$*Alice, Bob*$\}$*. Just from these views, Alice does not know whether Bob receives view* $\mathtt{v}'$ *while in view* $\mathtt{v}$*, or while in some other view,* $\mathtt{v}_{\mathtt{mid}}$ *with the membership* $\{$*Bob*$\}$*. If the former holds, then Alice does not need to synchronize with Bob because Virtually-Synchronous Delivery guarantees that they have received the same messages while in view* $\mathtt{v}$*; otherwise, she does. The transitional set given to Alice together with view* $\mathtt{v}'$ *provides this information.*

Figure 7 presents an automaton TS : SPEC that specifies delivery of transitional sets satisfying Definition 4.1. The automaton has two types of actions: output actions $\mathtt{view}_{\mathtt{p}}(\mathtt{v}, \mathtt{T})$, which deliver view $\mathtt{v}$ with transitional set $\mathtt{T}$ to process $\mathtt{p}$; and internal actions $\mathtt{set\_prev\_view}_{\mathtt{p}}(\mathtt{v})$, which declare that $\mathtt{q}$ intends to deliver view $\mathtt{v}$ while in its current view. The intentions are recorded in the variable $\mathtt{prev\_view}[\mathtt{p}][\mathtt{v}]$, and the current views are recorded in the variable $\mathtt{current\_view}[\mathtt{p}]$.

```
AUTOMATON TRANS_SET : SPEC

Signature:
Output: $view_p$(v,T), Proc p, View v, SetOf(Proc) T
Internal: $set\_prev\_view_p$(v), Proc p, View v

State:
For all Proc p:  View current_view[p], initially $v_p$
For all Proc p, View v: $View_\bot$ prev_view[p][v], initially $\bot$

Transitions:
 OUTPUT  $view_p$(v, T)
 pre: prev_view[p][v] = current_view[p]
      ($\forall$ q $\in$ v.set $\cap$ current_view[p].set)
            prev_view[q][v] $\neq \bot$
      T = {q $\in$ v.set $\cap$ current_view[p].set |
            prev_view[q][v] = current_view[p]}
 eff: current_view[p] $\leftarrow$ v

 INTERNAL  $set\_prev\_view_p$(v)
 pre: p $\in$ v.set
      prev_view[p][v] = $\bot$
 eff: prev_view[p][v] $\leftarrow$ current_view[p]
```

Figure 7: Transitional set specification.

Before process p can deliver a view v, each member q in the intersection of these views must execute $\mathtt{set\_prev\_view_q(v)}$, as enforced by the second precondition. The transitional set T delivered by p with v is then computed to consist of those processes q in the intersection current_view[p].set $\cap$ v.set for which prev_view[q][v] is the same as current_view[p]; this is specified by the third precondition on $\mathtt{view_p(v, T)}$.

4.1.4 Self Delivery

We now specify the *Self Delivery* property, which requires that each client receives all the messages it sent in a given view before receiving a new view. We specify this property as a simple modification of the WV_RFIFO : SPEC automaton presented in Section 4.1.1; the modified automaton is defined by the code contained in both Figures 5 and 8.

```
AUTOMATON WV_RFIFO+SELF : SPEC    MODIFIES  WV_RFIFO : SPEC

 Signature Extension:
  Output: $view_p$(v) modifies wv_rfifo.$view_p$(v)

 Transition Restriction:
  OUTPUT  $view_p$(v)
  pre: last_dlvrd[p][p] = LastIndexOf(msgs[p][current_view[p]])
```

Figure 8: WV_RFIFO+SELF service specification.

To enforce Self Delivery, a new precondition on the $\mathtt{view_p(v)}$ action requires the last_dlvrd[p][p] index to point to the last message sent by client p in its current view. Since the parent automaton, WV_RFIFO : SPEC, guarantees within-view gap-free FIFO delivery, this precondition implies that all of p's messages have in fact been delivered back to p.

In order for a GCS to be live and satisfy Within-View Delivery, Self Delivery, and Virtually-Synchronous Delivery, the GCS must *block* its application from sending new messages during view formation periods; this is proved in [FvR95]. Therefore, we introduce a block/block_ok synchronization when we extend our algorithm to support the Self Delivery property in Section 5.3.

Our formulation of Self Delivery as a safety property, when combined with the liveness property of Section 4.2, implies the formulations in [VKCD99] and [MAMSA94] of Self Delivery as a liveness property. These formulations require a GCS to *eventually* deliver to each process its own messages.

4.2 Liveness property

In a fault-prone asynchronous model, it is not feasible to require that a group communication service be live in every execution. The only way to specify useful liveness properties without strengthening the communication model is to make these properties *conditional* on the underlying network behavior (as specified, for example, in [FLS97, CS95, VKCD99]). Since our GCS uses an external membership service, we condition the GCS liveness on the behavior of the membership service. Provided the membership service eventually delivers the same last view to all the end-points comprising the view and does not deliver to them any subsequent start_change events, the end-points are required to eventually deliver to their applications this last view and all the messages sent in this view. Formally:

Property 4.1 *Let* GCS *be a group communication service whose interface with its clients consists of* `send`*,* `deliver`*, and* `view` *events as defined in the automaton signature in Figure 5. Furthermore, assume that the* GCS *uses a membership service* MBRSHP *described in Section 3.*

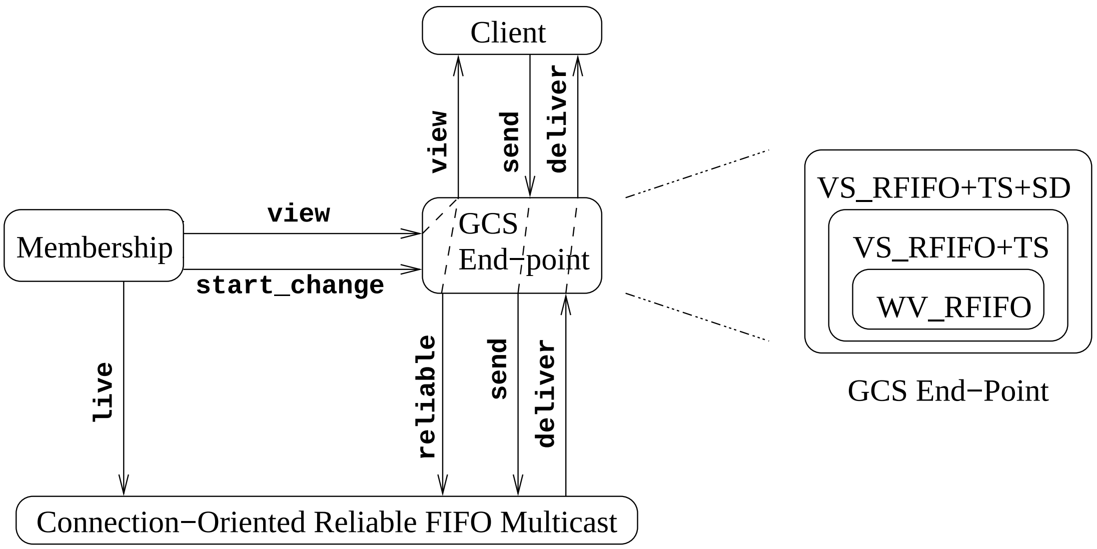
Let $\mathtt{v}$ *be a view. Let* α *be a fair execution of* GCS *in which, for every* $\mathtt{p} \in \mathtt{v.set}$*, the* MBRSHP.$\mathtt{view_p(v)}$ *action occurs and is followed by neither* MBRSHP.$\mathtt{view_p}$ *nor* MBRSHP.$\mathtt{start\_change_p}$*. Then at each* $\mathtt{p} \in \mathtt{v.set}$*,* GCS.$\mathtt{view_p(v)}$ *eventually occurs. Furthermore, for every* GCS.$\mathtt{send_p(m)}$ *that occurs after* GCS.$\mathtt{view_p(v)}$*, and for every* $\mathtt{q} \in \mathtt{v.set}$*,* GCS.$\mathtt{deliver_q(p,m)}$ *also occurs.*

It is important to note that although our liveness property requires the GCS to be live only in *certain* executions, any implementation that satisfies this property has to attempt to be live in *every* execution because it cannot test the external condition of the membership becoming stable. Also note that, even though membership stability is formally required to last forever, in practice it only has to hold "long enough" for the GCS to reconfigure, as explained in [DLS88, GS97]. However, we cannot explicitly introduce the bound on this time period in a fully asynchronous model, since it depends on external conditions such as message latency, process scheduling, and processing time.

5 The Virtually Synchronous Group Multicast Algorithm

In this section we present an algorithm for a group communication service, GCS, that satisfies the specifications in Section 4. The group communication service is implemented by a collection of GCS end-points, each running the same algorithm. Figure 9 (a) shows the interaction of a GCS end-point with its environment: a membership service MBRSHP and a reliable FIFO multicast service CO_RFIFO; these services are assumed to satisfy specifications of Section 3. The end-point interacts with its application client by accepting the client's send-requests and by delivering application messages and views to the client. The end-point uses the CO_RFIFO service to send messages to other GCS end-points and to receive messages sent by other GCS end-points. When necessary, the end-point uses the `reliable` action to inform CO_RFIFO of the set of end-points to which CO_RFIFO must maintain reliable (gap-free) FIFO connections. The GCS end-point also receives `start_change` and `view` notifications from the membership service.

The algorithm running at each GCS end-point is constructed incrementally using the inheritance-based methodology of [KKLS00]. We proceed in three steps, at each step adding support for a new property (see Figure 9 (b)):



(a) Interaction with Membership and Multicast Services

(b) Inheritance Hierarchy

Figure 9: A GCS end-point and its environment.

- First, in Section 5.1, we present an algorithm WV_RFIFO$_p$ for an end-point of the within-view reliable FIFO multicast service specified in Section 4.1.1, and argue that this service satisfies safety specification WV_RFIFO : SPEC and liveness Property 4.1.
- Then, in Section 5.2, we add support for the Virtually-Synchronous Delivery and Transitional Set properties specified in Sections 4.1.2 and 4.1.3. We present a child VS_RFIFO+TS$_p$ of WV_RFIFO$_p$, and argue that the service built from VS_RFIFO+TS$_p$ end-points satisfies safety specifications VSRFIFO : SPEC and TS : SPEC, and liveness Property 4.1.
- Finally, in Section 5.3, we add support for the Self Delivery property specified in Section 4.1.4. The resulting automaton VS_RFIFO+TS+SD$_p$ models a complete GCS end-point. Due to the use of inheritance, the service built from these end-points automatically satisfies safety specifications WV_RFIFO : SPEC, VSRFIFO : SPEC, and TS : SPEC. We argue that it also satisfies safety specification SELF : SPEC and liveness Property 4.1.

In the presented automata, each locally controlled action is defined to be a task by itself, which means that, if it becomes and stays enabled, it eventually gets executed.

When composing automata into a service, actions of the type $\texttt{mbrshp.start\_change}_p(\texttt{id}, \texttt{set})$ are linked with $\texttt{co\_rfifo.live}_p(\texttt{set})$, and $\texttt{mbrshp.view}_p(\texttt{v})$ are linked with $\texttt{co\_rfifo.live}_p(\texttt{v.set})$. This way, the live_set[p] variable of CO_RFIFO matches the MBRSHP's perception of which end-points are alive and connected to p. (We assume that every permanently disconnected end-point is eventually excluded by either a start_change or a view notification.) In the composed system, all output actions except the application interface are reclassified as internal.

For simplicity of the code, the presented automata do not include certain practical optimizations, such as for example garbage collection; we point out some of the important ones in Section 5.4.

5.1 Within-view reliable FIFO multicast algorithm

In this section we present the WV_RFIFO$_p$ algorithm running at an end-point p of a basic group communication service, WV_RFIFO. The end-point algorithm is quite simple: It relies on the MBRSHP service to form and deliver views involving end-point p; the end-point forwards these views to its client. The algorithm also relies on the CO_RFIFO service to provide reliable gap-free FIFO multicast communication. When the end-point receives a message-send request from its client, it uses CO_RFIFO to send the message to other end-points in the client's current view. The end-point delivers to its client the messages received from other end-points via CO_RFIFO, provided the client's current view matches the views in which the messages were sent. The algorithm keeps track in which views messages are sent using the following technique: each time the end-point delivers a view v to its client, it sends a special view_msg message to the end-points in v.set, informing them that the end-point's future messages will be sent in view v. Reliable delivery of messages is ensured by having CO_RFIFO maintain a reliable connection to every member of the end-point's view.

Figure 10 models the WV_RFIFO$_p$ algorithm as an automaton. The signature defines the interface through which end-point p interacts with its client and with the MBRSHP and CO_RFIFO services.

When a view v is received from MBRSHP via action mbrshp.view$_p$(v), end-point p saves it in a variable mbrshp_view and then delivers v to its client by executing action view$_p$(v). Variable current_view contains the last view delivered to the client. The precondition, $\mathtt{v} = \mathtt{mbrshp\_view} \neq \mathtt{current\_view}$, on the view$_p$(v) action ensures that v is indeed the last view received from MBRSHP and that it has not already been delivered to the client. After end-point p delivers view v to its client, it sends a view_msg containing v to the rest of the members of current_view.set by using action co_rfifo.send$_p$(set, tag = **view_msg**, v) with $\mathtt{set} = \mathtt{current\_view.set} - \{\mathtt{p}\}$ and $\mathtt{v} = \mathtt{current\_view}$. Variable view_msg[p] contains the last view sent as a view_msg. The first precondition, $\mathtt{view\_msg[p]} \neq \mathtt{current\_view}$, on co_rfifo.send$_p$(set, tag = **view_msg**, v) ensures that each view_msg is sent only once, and the second precondition, $\mathtt{current\_view.set} \subseteq \mathtt{reliable\_set}$, ensures that, prior to sending the view_msg, end-point p has requested CO_RFIFO to maintain reliable connection to every member of the client's view by executing action co_rfifo.reliable$_p$(set), which sets variable reliable_set to the value of set. When end-point p receives a view_msg from some end-point q via the co_rfifo.deliver$_{q,p}$(tag = **view_msg**, v) action, it stores v in a variable view_msg[q].

End-point p maintains a queue msgs[q][v] per each end-point q and view v; these queues are used for storing application messages received both from other end-points via co_rfifo.deliver$_{q,p}$ and from the end-point's own client via send$_p$. When action send$_p$(m) occurs, m is appended to msgs[p][current_view]. The end-point maintains indices that enforce message handling in the order of their appearances in the msgs queues: index last_sent points to the last application message m on msgs[p][current_view] that was sent using co_rfifo.send$_p$(set, tag = **app_msg**, m); index last_rcvd[q], for each end-point q, points to the last message m on msgs[q][view_msg[q]] that was delivered to p by co_rfifo.deliver$_{q,p}$(tag = **app_msg**, m); index last_dlvrd[q], for each end-point q, points to the last message m on msgs[q][current_view] that was delivered to p's client using deliver$_p$(q, m). The first precondition of co_rfifo.send$_p$(set, tag = **app_msg**, m) ensures that a view_msg containing current_view has been already sent to everybody in $\mathtt{set} = \mathtt{current\_view} - \{\mathtt{p}\}$. The preconditions on sending view_msgs, imply that CO_RFIFO already maintains a reliable connection to everyone in set, when co_rfifo.send$_p$(set, tag = **app_msg**, m) occurs.

Automaton WV_RFIFO$_p$ implements auxiliary functionality that allows end-point p to forward an application message received from some end-point to some other end-points. Specifically, using co_rfifo.send$_p$(set, tag = **fwd_msg**, r, v, m, i), end-point p can forward to some set of end-points

AUTOMATON WV_RFIFO$_p$

```
Type:
 ViewMsg = View
 FwdMsg  = Proc × View × AppMsg × Int

Signature:
 Input:  $send_p$(m), AppMsg m
         co_rfifo.$deliver_{q,p}$(m), Proc q,
               (AppMsg + ViewMsg + FwdMsg) m
         mbrshp.$view_p$(v), View v

 Output: $deliver_p$(q, m), Proc q, AppMsg m
         co_rfifo.$send_p$(set, m), SetOf(Proc) set,
               (AppMsg + ViewMsg + FwdMsg) m
         co_rfifo.$reliable_p$(set), SetOf(Proc) set
         $view_p$(v), View v
```

```
State:
 // Variables for handling application messages
 For all Proc q, View v: SequenceOf($AppMsg_\bot$)
     msgs[q][v], initially empty
 Int last_sent, initially 0
 For all Proc q: Int last_rcvd[q],  initially 0
 For all Proc q: Int last_dlvrd[q], initially 0

 // Variables for handling views and view messages
 View current_view, initially $v_p$
 View mbrshp_view,  initially $v_p$
 For all Proc q: View view_msg[q], initially $v_q$

 SetOf(Proc) reliable_set, initially $v_p$.set
```

```
Transitions:
 INPUT  mbrshp.$view_p$(v)
 eff: mbrshp_view ← v

 OUTPUT  $view_p$(v)
 pre: v = mbrshp_view ≠ current_view
 eff: current_view ← v
      last_sent  ← 0
      (∀ q) last_dlvrd[q] ← 0

 OUTPUT  co_rfifo.$reliable_p$(set)
 pre: current_view.set ⊆ set
 eff: reliable_set ← set

 OUTPUT  co_rfifo.$send_p$(set, tag=view_msg, v)
 pre: view_msg[p] ≠ current_view
      current_view.set ⊆ reliable_set
      set =  current_view.set - {p}
      v = current_view
 eff: view_msg[p] ← current_view

 INPUT  co_rfifo.$deliver_{q, p}$(tag=view_msg, v)
 eff: view_msg[q]  ← v
      last_rcvd[q] ← 0
```

```
 INPUT  $send_p$(m)
 eff: append m to msgs[p][current_view]

 OUTPUT  $deliver_p$(q, m)
 pre: m = msgs[q][current_view][last_dlvrd[q]+1]
 eff: last_dlvrd[q] ← last_dlvrd[q] + 1

 OUTPUT  co_rfifo.$send_p$(set, tag=app_msg, m)
 pre: view_msg[p] = current_view
      set =  current_view.set - {p}
      m = msgs[p][current_view][last_sent + 1]
 eff: last_sent ← last_sent + 1

 INPUT  co_rfifo.$deliver_{q,p}$(tag=app_msg, m)
 eff: msgs[q][view_msg[q]][last_rcvd[q]+1]←m
      last_rcvd[q] ← last_rcvd[q] + 1

 OUTPUT  co_rfifo.$send_p$(set,tag=fwd_msg,r,v,m,i)
 pre: (p ∉ set)  ∧  (m = msgs[r][v][i])

 INPUT  co_rfifo.$deliver_{q,p}$(tag=fwd_msg,r,v,m,i)
 eff:  msgs[r][v][i] ← m
```

Figure 10: Within-view reliable FIFO multicast end-point automaton.

the ith message, m, sent by the client at r in view v. In turn, when the end-point receives a forwarded message from end-point q via co_rfifo.$\text{deliver}_{q,p}$(tag = **fwd_msg**, r, v, m, i), it stores m in the ith location of the msgs[r][v] queue. The code of WV_RFIFO$_p$ does not specify a particular strategy for forwarding messages; the strategy can be chosen non-deterministically. Such a strategy can be specified by more refined versions of the algorithm and/or by modifications of WV_RFIFO$_p$, as we do in the VS_RFIFO+TS$_p$ modification of the WV_RFIFO$_p$ automaton in Section 5.2 below.

Leaving certain level of non-determinism at the parent automaton, with the intention of resolving it later at the child automaton, is a technique similar to the use of *abstract methods* or *pure virtual methods* in object-oriented methodology [KKLS00]. We use the same technique in the co_rfifo.reliable$_p$(set) action when we require set to be a nondeterministic superset of current_view.set. The VS_RFIFO+TS$_p$ modification of WV_RFIFO$_p$ places additional preconditions on this action, thereby specifying precise values for the set argument.

The WV_RFIFO automaton resulting from the composition of all the end-point automata and the MBRSHP and CO_RFIFO automata models the WV_RFIFO service. The automaton satisfies the safety properties specified by WV_RFIFO : SPEC: it preserves the Local Monotonicity and Self Inclusion properties of view deliveries guaranteed by the MBRSHP service; and it also extends the gap-free FIFO-ordered message delivery of CO_RFIFO with the Within-View Delivery property. The Within-View Delivery is achieved by delivering messages to the clients only if the views in which the messages were sent match the clients' current views.

Appendix B.1 contains a simulation from WV_RFIFO to WV_RFIFO : SPEC: Actions of automaton WV_RFIFO : SPEC involving view$_p$(v), send$_p$(m), and deliver$_p$(q, m) are simulated when WV_RFIFO takes the corresponding view$_p$(v), send$_p$(m), and deliver$_p$(q, m) actions. Steps of WV_RFIFO involving other actions correspond to empty steps of WV_RFIFO : SPEC. We define the following function R that maps every reachable state s of WV_RFIFO to a reachable state of WV_RFIFO : SPEC, where s[p].var denotes an instance of a variable var of end-point p in a state s:

```
R(s ∈ ReachableStates(WV_RFIFO)) = t ∈ ReachableStates(WV_RFIFO : SPEC), where
  For each Proc p, View v:       t.msgs[p][v]   =   s[p].msgs[p][v]
  For each Proc p, Proc q: t.last_dlvrd[p][q]   =   s[q].last_dlvrd[p]
  For each Proc p:          t.current_view[p]   =   s[p].current_view
```

Lemma B.1 states that R is a refinement mapping from WV_RFIFO to WV_RFIFO : SPEC; the proof relies on a number of invariant assertions, stated and proved in Appendix B.1 as well.

The WV_RFIFO automaton also satisfies liveness Property 4.1. Consider a fair execution in which each end-point p in v.set receives the same view v from the membership and no view events afterwards. Starting from the time the mbrshp.view$_p$(v) action occurs, the view$_p$(v) action stays enabled; therefore it eventually happens due to the fairness of the execution. After view v is delivered to the clients, all messages sent in view v are also eventually delivered to the clients. This is due to the liveness property of CO_RFIFO, which guarantees that messages sent between live and connected end-points (as perceived by the membership service) are eventually delivered to their destinations. We prove these claims formally for the complete GCS algorithm in Appendix C.

5.2 Adding support for Virtually Synchronous Delivery and Transitional Sets

The WV_RFIFO service of the previous section guarantees that each member p of a view v receives *some* prefix of the FIFO ordered stream of messages sent by every member q in v. In this section, we modify the WV_RFIFO$_p$ algorithm to yield an end-point VS_RFIFO+TS$_p$ of a service, VS_RFIFO+TS, that, in addition to the semantics provided by WV_RFIFO, guarantees that those members that transition from v in to *the same* view v′, receive not just *some* but *the same* prefix of the message stream sent by each member q in v. This is the Virtually-Synchronous Delivery property, the key property of Virtual Synchrony semantics (see Section 4.1.2). Overall, the VS_RFIFO+TS service satisfies the VSRFIFO : SPEC and TS : SPEC safety specifications, as well as liveness Property 4.1; we prove these claims respectively in Appendixes B.2, B.3, and C.

In a nutshell, here is how the VS_RFIFO+TS$_p$ algorithm computes transitional sets and enforces Virtually-Synchronous Delivery: When end-point p is notified via mbrshp.start_change$_p$(cid, set) of the MBRSHP's attempt to form a new view, p sends via CO_RFIFO a synchronization message tagged with cid to every end-point in set; if subsequent start_change notifications with the same cid but a different set occur, p forwards its last synchronization message to the joining end-points. The synchronization message includes p's current view v and a mapping cut, such that cut(q) is the index of the last message from each q in v.set that p commits to deliver in view v. Notice that a synchronization message is sent right after a start_change notification is received, without waiting for a new view to be formed. Once p receives a new view v′ from MBRSHP and also a synchronization message tagged with v′.startId(q) from each end-point q in v.set ∩ v′.set, p computes a transitional set from v to v′ and decides on which messages it needs to deliver to its client in view v before delivering view v′. A transitional set T from v to v′ is computed to include every client q in v.set ∩ v′.set whose synchronization message tagged with v′.startId(q) contains the same view as p's current view v. For each client r in v.set, end-point p decides to deliver as much messages sent by r as committed to by any member q of T in its synchronization message tagged with v′.startId(q). Section 5.2.1 describes two message-forwarding strategies that ensure p's ability to actually deliver all the messages it decides to deliver. After p delivers all these messages to its client, it then delivers to its client the new view v′ along with the transitional set T.

Virtually-Synchronous Delivery follows from the fact that all end-points transitioning from view v to v′ consider the same synchronization messages, compute the same set T, and hence use the same data to decide which messages to deliver in view v before delivering view v′. Set T satisfies Definition 4.1 of a transitional set from v to v′ because (a) every end-point that computes T is itself included in T, and (b) no end-point q in T is allowed to deliver v′ while in some view other than v because v′.startId(q) is linked through q's synchronization message to v.

Figures 10, 11 and 12, together, contain the code of the VS_RFIFO+TS$_p$ automaton that models end-point p of the VS_RFIFO+TS service. Figures 11 and 12 specify how the WV_RFIFO$_p$ automaton of Figure 10 is modified to support Virtually-Synchronous Delivery and Transitional Sets. Figure 11 contains Signature Extension that defines the signatures of new and modified actions; Figure 12 contains State Extension and Transition Restriction defining respectively new state variables and new precondition/effect code. We now describe automaton VS_RFIFO+TS$_p$ in detail.

AUTOMATON VS_RFIFO+TS$_p$ MODIFIES WV_RFIFO$_p$

Type: SyncMsg = StartChangeId × View × (Proc→Int)

Signature Extension:

Input: mbrshp.start_change$_p$(id, set), StartChangeId id, SetOf(Proc) set new
co_rfifo.deliver$_{q,p}$(m), Proc q, SyncMsg m new

Output: deliver$_p$(q, m) modifies wv_rfifo.deliver$_p$(q, m)
view$_p$(v, T), SetOf(Proc) T modifies wv_rfifo.view$_p$(v)
co_rfifo.reliable$_p$(set), SetOf(Proc) set modifies wv_rfifo.co_rfifo.reliable$_p$(set)
co_rfifo.send$_p$(set, m), SetOf(Proc) set, SyncMsg m new
co_rfifo.send$_p$(set, m) modifies wv_rfifo.co_rfifo.send$_p$(set, m), FwdMsg m

Internal: set_cut$_p$() new

Figure 11: Virtually Synchronous reliable FIFO multicast: Signature Extension.

Upon receiving mbrshp.start_change$_p$(cid, set), VS_RFIFO+TS$_p$ stores the cid and set parameters in the id and set fields of a variable start_change. When start_change has a value different from ⊥, it indicates that VS_RFIFO+TS$_p$ is engaged in a synchronization protocol, during which it exchanges synchronization messages tagged with start_change.id with the end-points in start_change.set; after VS_RFIFO+TS$_p$ delivers a view to its client it resets start_change to ⊥.

```
AUTOMATON VS_RFIFO+TS_p   MODIFIES   WV_RFIFO_p

State Extension:
 (StartChangeId × SetOf(Proc))_⊥ start_change, initially ⊥
 For all Proc q, ViewId id:  (View v, (Proc→Int) cut)_⊥ sync_msg[q][id], initially ⊥
 SetOf(Proc) sync_set, initially empty
 SetOf((Proc × Proc × View × Int)) forwarded_set, initially empty

Transition Restriction:
 INPUT  mbrshp.start_change_p(cid, set)
 eff: if start_change ≠ ⊥ ∧ start_change.id = cid
         then  sync_set ← sync_set ∩ set
         else  sync_set ← { }
      start_change ← ⟨cid, set⟩

 OUTPUT  co_rfifo.reliable_p(set)
 pre: start_change = ⊥ ⇒  set = current_view.set
      start_change ≠ ⊥ ⇒  set = current_view.set ∪ start_change.set

 INTERNAL  set_cut_p()
 pre: start_change ≠ ⊥ ∧ sync_msg[p][start_change.id] = ⊥
 eff: Let cut = {⟨q, LongestPrefixOf(msgs[q][current_view])⟩ | q ∈ current_view.set}
      sync_msg[p][start_change.id] ← ⟨current_view, cut⟩
      sync_set ← {p}

 OUTPUT  co_rfifo.send_p(set, tag=sync_msg, cid, v, cut)
 pre: start_change ≠ ⊥  ∧  sync_msg[p][start_change.id] ≠ ⊥
      set = (start_change.set - sync_set) ≠ { }
      set ⊆ reliable_set
      cid = start_change.id ∧ ⟨v, cut⟩ = sync_msg[p][cid]
 eff: sync_set ← start_change.set

 INPUT  co_rfifo.deliver_q,p(tag=sync_msg, cid, v, cut)
 eff: sync_msg[q][cid] ← ⟨v, cut⟩

 OUTPUT  deliver_p(q, m)
 pre: if (start_change ≠ ⊥ ∧ sync_msg[p][start_change.id] ≠ ⊥) then
         if start_change.id ≠ mbrshp_view.startId(p) then
            last_dlvrd[q]+1 ≤ sync_msg[p][start_change.id].cut(q)
         else  let S = {r ∈ mbrshp_view.set ∩ current_view.set |
                           sync_msg[r][mbrshp_view.startId(r)].view = current_view}
              last_dlvrd[q]+1 ≤ max_{r ∈ S} sync_msg[r][mbrshp_view.startId(r)].cut(q)

 OUTPUT  view_p(v, T)
 pre: v.startId(p) = start_change.id    // to prevent delivery of obsolete views
      v.set - sync_set = { }     // all sync msgs are sent
      last_sent ≥ sync_msg[p][v.startId(p)].cut(p)    // sent out your own msgs
      (∀ q ∈ v.set ∩ current_view.set) sync_msg[q][v.startId(q)] ≠ ⊥
      T = {q ∈ v.set ∩ current_view.set | sync_msg[q][v.startId(q)].view = current_view}
      (∀ q ∈ current_view.set) last_dlvrd[q] = max_{r ∈ T} sync_msg[r][v.startId(r)].cut(q)
 eff: start_change ← ⊥
      sync_set ← { }

 OUTPUT  co_rfifo.send_p(set,tag=fwd_msg,r,v,m,i)
 pre: (∀ q ∈ set) (⟨q, r, v, i⟩ ∉ forwarded_set)   ∧   ForwardStrategyPredicate(set, r, v, i)
 eff: (∀ q ∈ set) add ⟨q, r, v, i⟩ to forwarded_set
```

Figure 12: Virtually Synchronous reliable FIFO multicast: State Extension & Transition Restriction.

Variable sync_set indicates a set of end-points to which a synchronization message tagged with the latest start_change.id has already been sent. When MBRSHP modifies the membership of a forming view by issuing a new $\texttt{mbrshp.start\_change}_{\texttt{p}}(\texttt{cid}, \texttt{set})$ with the previous cid and an updated set, the disconnected end-points are removed from sync_set by setting sync_set to $\texttt{sync\_set} \cap \texttt{set}$. This way, if any of the disconnected end-points later re-join, the synchronization message will be re-sent to them. If MBRSHP issues a new cid, then sync_set is reset to $\{\}$ to indicate that a new synchronization message needs to be sent to every end-point in set.

After $\text{VS\_RFIFO+TS}_{\texttt{p}}$ receives a $\texttt{mbrshp.start\_change}_{\texttt{p}}(\texttt{cid}, \texttt{set})$ input from MBRSHP, it executes an internal action, $\texttt{set\_cut}_{\texttt{p}}()$, that commits p to deliver to its client all the messages it has so far received from the members of its current view. For each member q of current_view.set, cut(q) is set to the length of the longest continuous prefix of messages in msgs[q][current_view].[2] Action $\texttt{set\_cut}_{\texttt{p}}()$ results in p's current view being stored in sync_msg[p][start_change.id].view, the committed cut – in sync_msg[p][start_change.id].cut, and sync_set being set to $\{\texttt{p}\}$.

$\text{VS\_RFIFO+TS}_{\texttt{p}}$ specifies precise preconditions on the $\texttt{co\_rfifo.reliable}_{\texttt{p}}(\texttt{set})$ actions. When $\text{VS\_RFIFO+TS}_{\texttt{p}}$ is not engaged in a synchronization protocol (i.e., when $\texttt{start\_change} = \bot$), CO_RFIFO is asked to maintain reliable connection just to the end-points in p's current view, current_view.set. When $\text{VS\_RFIFO+TS}_{\texttt{p}}$ is engaged in a synchronization protocol, it requires CO_RFIFO to maintain reliable connection to the members of a forming view, start_change.set, as well as to those in current_view.set. Thus, CO_RFIFO avoids loss of messages sent to the disconnected end-points in case these end-points are later added to the forming view.

After setting the cut and telling CO_RFIFO to maintain reliable connection to everyone in $\texttt{current\_view.set} \cup \texttt{start\_change.set}$, $\text{VS\_RFIFO+TS}_{\texttt{p}}$ uses $\texttt{co\_rfifo.send}_{\texttt{p}}$ to send the synchronization message sync_msg[p][start_change.id] tagged with start_change.id to the end-points in $\texttt{start\_change.set} - \texttt{sync\_set}$. Synchronization messages received from other end-points via $\texttt{co\_rfifo.deliver}_{\texttt{q,p}}(\texttt{tag} = \textbf{sync\_msg}, \texttt{cid}, \texttt{v}, \texttt{cut})$ result in $\langle \texttt{v}, \texttt{cut} \rangle$ being saved in sync_msg[q][cid].

$\text{VS\_RFIFO+TS}_{\texttt{p}}$ restricts delivery of application messages while it is engaged in a synchronization protocol (i.e., when $\texttt{start\_change} \neq \bot$ and $\texttt{sync\_msg[p][start\_change.id]} \neq \bot$): Prior to receiving a new view from MBRSHP, only the messages identified in the cut of its own latest synchronization message, sync_msg[p][start_change.id].cut, can be delivered to the client. After $\texttt{mbrshp.view}_{\texttt{p}}(\texttt{v})$ occurs, $\text{VS\_RFIFO+TS}_{\texttt{p}}$ is allowed to deliver messages identified in the cut sync_msg[q][v.startId(q)].cut received from q, provided q is a member of a transitional set from current_view to v. An end-point $\texttt{q} \in \texttt{current\_view.set} \cap \texttt{v.set}$ is considered to be in a transitional set from current_view to v if sync_msg[q][v.startId(q)].view is the same as p's current_view.

$\text{VS\_RFIFO+TS}_{\texttt{p}}$ delivers a view v received from MBRSHP and a transitional set T to its client when p receives a synchronization message sync_msg[q][v.startId(q)] from every q in current_view.set $\cap$ v.set, computes T, and delivers all the application messages identified in the cuts of the members of T (as specified by the last three preconditions on $\texttt{view}_{\texttt{p}}(\texttt{v}, \texttt{T})$). The first two preconditions ensure respectively that no new $\texttt{mbrshp.start\_change}_{\texttt{p}}$ notification was issued after $\texttt{mbrshp.view}_{\texttt{p}}(\texttt{v})$ and that p sent its synchronization message to everybody in v.set. The third precondition specifies that, before delivering view v, p must send to others all of its own messages indicated in its own cut. All these preconditions work in conjunction with those in $\texttt{wv\_rfifo.view}_{\texttt{p}}(\texttt{v})$.

Recall from Section 5.1 that $\text{WV\_RFIFO}_{\texttt{p}}$ allows for nondeterministic forwarding of other end-points' application mesages. $\text{VS\_RFIFO+TS}_{\texttt{p}}$ resolves this nondeterminism by placing two additional preconditions on $\texttt{co\_rfifo.send}_{\texttt{p}}(\texttt{set}, \texttt{tag} = \textbf{fwd\_msg}, \texttt{r}, \texttt{v}, \texttt{m}, \texttt{i})$: The first checks a variable

[2] The longest continuous prefix can be different from the length of msgs[q][current_view] because forwarded messages may arrive out of order and introduce gaps in the msgs queues.

forwarded_set to make sure that message m was not previously forwarded to anyone in set. The second tests that a certain ForwardingStrategyPredicate(set, r, v, i) holds. This predicate is designed to ensure that all end-points in the transitional set T are able to deliver all the messages that each has committed to deliver in its synchronization message, in particular those sent by disconnected clients. End-points test ForwardingStrategyPredicate to decide whether they need to forward any messages to others.

5.2.1 Forwarding Strategy Predicate

We now provide two examples of ForwardingStrategyPredicates. With the first, multiple copies of the same message may be forwarded by different end-points. The second strategy minimizes the number of forwarded copies of a message. Many other possible strategies exist. For example, a strategy can employ randomization to decide whether an end-point should forward a message in a certain time slice, and suppress forwarding of messages that have already been forwarded by others.

A simple strategy: With our first strategy, a process p forwards a message m only if p has committed to deliver m. In addition, if m was originally sent in view v, p forwards m to a process q only if p does not know of any view of q later than v, and if the latest sync_msg from q sent in view v indicates that q has not received message m. The strategy is defined as follows:

```
ForwardingStrategyPredicate(set, r, v, i) $\equiv$
 ($\exists$ cid) (sync_msg[p][cid].view $=$ v  $\wedge$  i $\leq$ sync_msg[p][cid].cut(r))
 $\wedge$ set $=$ { q |  view_msg[q] $\leq$ v  $\wedge$  ($\exists$ cid') (sync_msg[q][cid'].view $=$ v
        $\wedge$ ($\nexists$ cid'' $>$ cid') sync_msg[q][cid''].view $=$ v $\wedge$ sync_msg[q][cid'].cut(r) $<$ i) }
```

If some process q is missing a certain message m, m will be forwarded to q by some end-point p that has committed to deliver m, when p learns from q's synchronization message that q misses m.

Minimizing the number of forwarded copies of a message: The second strategy relies on the computed transitional set T from view v to v′ to decide which message should be forwarded by which member of the transitional set. Assume that a member u of T misses a message m that was originally sent in v by a non-member r of T, but that was committed to delivery by some other members of T. Among these memebers, ForwardingStrategyPredicate selects the one with the minimal process-identifier to forward m to u; variations of this predicate may use a different deterministic rule for selecting a member, for example, accounting for network topology or communication costs. The selected end-point, p, forwards the message to u only if view v′ is the latest view known to p, as specified by the first conjunct below. Otherwise, v′ is an obsolete view, so there is no need to help u transition in to v′. The described strategy does not forward to u $\in$ T messages from the members of T because u is guaranteed to receive these messages directly from their original senders (unless v′ becomes obsolete because of further view changes occur).

```
ForwardingStrategyPredicate(set, r, v, i) $\equiv$
  Let v' $=$ mbrshp_view $\wedge$                            // latest view known to {p}
  sync_msg[p][v'.startId(p)] $\neq \bot$ $\wedge$                   // already sent own sync_msg
  Let v $=$ sync_msg[p][v'.startId(p)].view $\wedge$
  ($\forall$ q $\in$  v.set $\cap$ v'.set) sync_msg[q][v'.startId(q)] $\neq \bot$  $\wedge$   // received right sync_msgs
  Let T $=$ {q $\in$ v.set $\cap$ v'.set | sync_msg[q][v'.startId(q)].view $=$ v} $\wedge$
  r $\notin$ T   $\wedge$                                    // only forward messages from end-point not in T
  set $=$ {u $\in$ T  | sync_msg[u][v'.startId(u)].cut(r) $<$ i }  $\wedge$
  p $=$ min{u $\in$ T  | sync_msg[u][v'.startId(u)].cut(r) $\geq$ i }
```

If all end-points receive the same view from MBRSHP, only one copy of m will be forwarded to each u. In rare cases, however, when MBRSHP delivers different views to different end-points, more than one end-point may forward the same message m to the same end-point u.

Each end-point waits to receive a new view from MBRSHP and all the right synchronization messages before it forwards messages to others. Thus, compared to the first strategy, this strategy reduces the communication traffic at the cost of slower recovery of lost messages.

5.2.2 Correctness

The VS_RFIFO+TS automaton, resulting from the composition of all end-point automata and the MBRSHP and CO_RFIFO automata, satisfies the VSRFIFO : SPEC and TS : SPEC safety specifications, as well as Liveness Property 4.1, as we formally prove in Appendixes B.2, B.3, and C, respectively. Below we give high-lights of these proofs.

VSRFIFO : SPEC is a modification of WV_RFIFO : SPEC. The proof that VS_RFIFO+TS satisfies VSRFIFO : SPEC reuses the proof that WV_RFIFO satisfies WV_RFIFO : SPEC and involves reasoning about only how VSRFIFO : SPEC modifies WV_RFIFO : SPEC. The proof extends refinement mapping R between WV_RFIFO and WV_RFIFO : SPEC with a mapping R_n that maps the cuts used by the end-points of VS_RFIFO+TS to move from a view v to a view v' to the $\mathtt{cut}[v][v']$ variable of VSRFIFO : SPEC. The proof depends on Invariant B.9 and Corollary B.1, which state that all end-points that move from a view v to a view v' use the same synchronization messages, compute the same transitional set T, and therefore, use the same cuts.

The proof in Appendix B.3 shows that VS_RFIFO+TS satisfies TS : SPEC. The proof augments VS_RFIFO+TS_p with a *prophecy variable* that guesses, at the time p receives a $\mathtt{start\_change}_p(\mathtt{cid}, \mathtt{set})$ notification from MBRSHP, possible future views that may contain cid in their $\mathtt{startId}(p)$ mappings. For each of these views v', VS_RFIFO+TS simulates a $\mathtt{set\_prev\_view}_p(v')$ action of TS : SPEC, thereby fixing the previous view of v' to be p's current view v.

In a fair execution of VS_RFIFO+TS in which the same last view v' is delivered to all its members and no start_change events subsequently occur, the three preconditions on the $\mathtt{view}_p(v', T_p)$ delivery are eventually satisfied for every $p \in v'.\mathtt{set}$:

1. Condition $v'.\mathtt{startId}(p) = \mathtt{start\_change.id}$ remains true since by the assumption there are no subsequent start_change events at p.
2. End-point p eventually receives synchronization messages tagged with the "right" cid from every member of $v.\mathtt{set} \cap v'.\mathtt{set}$ because they keep taking steps towards reliably sending these synchronization messages to p (by low-level fairness of the code) and because CO_RFIFO eventually delivers these messages to p (by the liveness assumption on CO_RFIFO).
3. End-point p eventually receives and delivers all the messages committed to in the cuts of the members of the transitional set T_p because for each such message there is at least one end-point in T_p that has the message in its msgs buffer and that would reliably forward it to p (according to the ForwardingStrategyPredicate) if so necessary. Also, p never delivers any messages beyond those committed to in the cuts of the members of T_p because of the precondition on application message delivery.

5.3 Adding support for Self Delivery

As a final step in constructing the automaton that models an end-point of our group communication service, GCS_p, we add support for Self Delivery to the VS_RFIFO+TS_p automaton presented above.

Self Delivery requires each end-point to deliver to its client all the messages the client sends in a view, before moving on to the next view.

```
AUTOMATON GCS_p = VS_RFIFO+TS+SD_p   MODIFIES  VS_RFIFO+TS_p

Signature Extension:
Input:  block_ok_p() new                        Output: block_p() new
Internal: set_cut_p() modifies set_cut_p()              view_p(v,T) modifies vs_rfifo+ts.view_p(v,T)

State Extension:
block_status ∈ {unblocked, requested, blocked}, initially unblocked

Transition Restriction:
  INTERNAL  set_cut_p()                                   OUTPUT  block_p()
  pre: block_status = blocked                             pre: start_change ≠ ⊥
                                                               block_status = unblocked
                                                          eff: block_status ← requested

  OUTPUT  view_p(v,T)                                     INPUT  block_ok_p()
  eff: block_status ← unblocked                           eff: block_status ← blocked
```

Figure 13: GCS$_p$ end-point automaton.

In order to implement Self Delivery, Virtually-Synchronous Delivery, and Within-View Delivery together in a live manner, each end-point must *block* its client from sending new messages while a view change is taking place (as proven in [FvR95]). Therefore, we add to VS_RFIFO+TS$_p$ an output action block and an input action block_ok. We assume that the client at end-point p has the matching actions and that it eventually responds to every block request with a block_ok response and subsequently refrains from sending messages until a view is delivered to it. In Section B.4, we formalize this requirement as an abstract client automaton.

The GCS$_p$ automaton appears in Figure 13. After receiving the first start_change notification in a given view, end-point p issues a block request to its client and awaits receiving a block_ok response before executing set_cut$_p$(). As a result of set_cut$_p$(), p commits to deliver all the messages its client has sent in the current view. Therefore, p has to deliver all these messages before moving on to a new view, and Self Delivery is satisfied. Due to the use of inheritance, the GCS automaton preserves all the safety properties satisfied by its parent. Since end-point p has its own messages on the msgs[p][p] queue and can deliver them to its client, liveness is also preserved. Thus, GCS satisfies all the properties we have specified in Section 4.

5.4 Optimizations and Extensions

Having formally presented the basic algorithm for an end-point of our Virtually-Synchronous GCS, we now discuss several optimizations and extensions that can be added to the algorithm to make its implementation more practical. Specifically, we discuss ways to reduce the size and number of synchronization messages, as well as to avoid the use of non-volatile storage. We also discuss garbage collection.

The first optimization that reduces the size of synchronization messages relies on the following observation: An end-point p does not need to send its current view and its cut to end-points which are not in current_view.set because p cannot be included in their transitional sets. However, these end-points still need to hear from p if p is in their current views. Therefore, end-point p could send a smaller synchronization message to the end-points in start_change.set − current_view.set, containing its start_change.id only (but neither a view nor a cut). This message would be interpreted as saying "I am not in your transitional set", and the recipients of this message would know not to include p in their transitional sets for views v' with v'.startId(p) = p's start_change.id. When using this optimization, p also does not need to include its current view in the synchronization messages sent to current_view.set − start_change.set, since the view information can be deduced from p's view_msg.

An additional optimization can be used if we strengthen the membership specification to require a mbrshp.start_change with a new identifier to be sent every time MBRSHP changes its mind about the membership of a forming view. In this case, the latest mbrshp.start_change has the same membership as the delivered mbrshp.view. Therefore, the synchronization messages can be shortened to not include information about application messages delivered from end-points in start_change.set ∩ current_view.set: for an end-point p, end-points that have p in their transitional sets will deliver all the application messages that p sent before its synchronization message.

Other optimizations can reduce the total number of messages sent during synchronization protocol by all end-points. A simple way to do this is to transform the algorithm into a leader-based one, as [vRBM96, SR93]. A more scalable approach was suggested by Guo et al. [GVvR96]. Their algorithm uses a two-level hierarchy for message dissemination in order to implement Virtual Synchrony: end-points send synchronization messages to their designated leaders, which in turn exchange only the cumulative information among themselves. The number of messages exchanged to synchronize multiple groups can be also reduced, as suggested in [BFHR98, RGS+96], by aggregating information pertaining to multiple groups into a single message.

Another optimization addresses the use of stable storage. Recall that in Section 3 we assumed that end-points keep their running states on stable storage, and therefore, recover with their state intact. However, our group multicast service does provide meaningful semantics even when GCS end-points maintain their running state on volatile storage. When an end-point p recovers after a crash, it can start executing with its state reset to an initial value with current_view being the singleton view v_p. It needs to contact the MBRSHP service to be re-added to its groups. The client would refrain from sending any messages in its recovered view until it receives a new view from its end-point. This view would satisfy Local Monotonicity and Self Inclusion because these are the properties guaranteed by the MBRSHP service. The specification of Virtually-Synchronous Delivery should be changed to make recovery be interpreted as delivering a singleton view. The remaining safety properties are also preserved because they involve message delivery within a single view.

In a practical implementation of our service, some sort of garbage collection mechanism is required in order to keep the buffer sizes finite. The implementation of [Tar00] discards messages from older views when moving to a new view, and also when learning that they were already delivered to every client in the view. This implementation also discards older synchronization messages: an end-point only holds on to the latest synchronization message it received from each end-point. This optimization does not violate liveness since discarded synchronization messages necessarily pertain to obsolete views.

6 Discussion

We have designed a novel, efficient group multicast service targeted for WANs. Our service implements a variant of the Virtual Synchrony semantics that includes a collection of properties that have been shown useful for many distributed applications (see [VKCD99]). Many GCSs, for example [vRBM96, SR93, BDM98, AMMS+95, DM96], support these and similar properties. Our design has been implemented [Tar00] (in C++) as part of a novel architecture for scalable group communication in WANs, using the datagram service of [ACSD00] and the membership algorithm of Keidar et al. [KSMD00].

The main contribution of this paper is a Virtual Synchrony algorithm run by GCS end-points, in particular, its synchronization protocol, which enforces Virtually-Synchronous Delivery. This protocol is invoked when the underlying membership service begins to form a new view, and is run while the view is forming. The protocol involves a single message-exchange round during which members of the forming view send synchronization messages to each other. In contrast to previously suggested Virtual Synchrony algorithms (e.g., [FvR95, AMMS$^+$95, GVvR96, ADKM92, BDM98]), our algorithm does not require end-points to pre-agree upon a globally unique identifier before sending synchronization messages, and thus involves less communication. Performing less communication is especially important in WANs, where message latency tends to be high.

Furthermore, unlike the algorithms in [AMMS$^+$95, GVvR96, BDM98, SR93], our algorithm allows the membership service to change the membership of a forming view while the synchronization protocol is running; the protocol responds immediately to such membership changes. The following example demonstrates the benefits of this approach:

Example 6.1 *Figure 14 presents a sample execution involving two clients,* $\mathtt{a}$ *and* $\mathtt{b}$*. Vertical arrows represent time passage at each client, empty circles represent client-level events, and gray circles –* MBRSHP*-level events. First, both clients receive the same view* $\mathtt{v} = \langle 2, \{\mathtt{a}, \mathtt{b}\}, [\mathtt{a}:1, \mathtt{b}:1]\rangle$ *from their* GCS *end-points,* $\text{GCS}_\mathtt{a}$ *and* $\text{GCS}_\mathtt{b}$*; the circle around these view events highlights that the delivered views are the same. At some point, the* MBRSHP *service notifies* $\text{GCS}_\mathtt{b}$ *that it is starting to form a view without* $\mathtt{a}$*. While doing so, it detects that* $\mathtt{a}$ *is connected to* $\mathtt{b}$ *afterall, so it changes the membership of the forming view to* $\{\mathtt{a}, \mathtt{b}\}$*.* $\text{GCS}_\mathtt{b}$ *forwards to* $\text{GCS}_\mathtt{a}$ *its latest synchronization message; synchronization messages are denoted by dashed lines.* $\text{GCS}_\mathtt{a}$ *is also notified by* MBRSHP *of its attempt to form a new view with* $\mathtt{b}$*; this causes* $\text{GCS}_\mathtt{a}$ *to send a synchronization message to* $\text{GCS}_\mathtt{b}$*. When* MBRSHP *completes its view formation protocol, it delivers the new view* $\mathtt{v}' = \langle 3, \{\mathtt{a}, \mathtt{b}\}, [\mathtt{a}:2, \mathtt{b}:2]\rangle$ *to both* GCS *end-points. After the* GCS *end-points receive each-others' synchronization messages, they compute their transitional sets to be* $\mathtt{T}' = \{\mathtt{a}, \mathtt{b}\}$*, decide on which application messages they need to deliver, deliver these messages, and then deliver* $\mathtt{v}'$ *and* $\mathtt{T}'$ *to their clients. From* $\mathtt{T}'$*,* $\mathtt{a}$ *and*

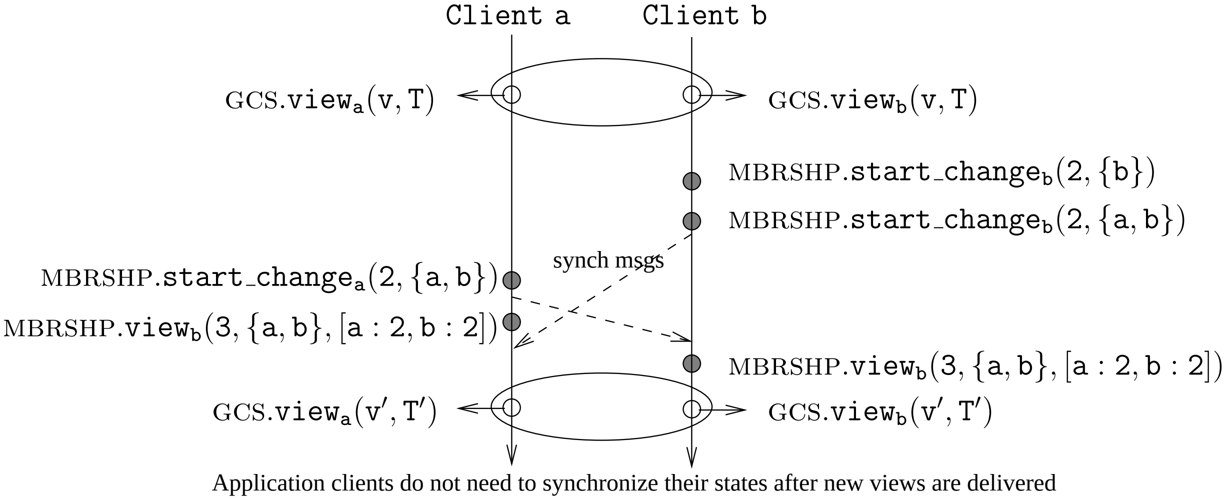

Client a
Client b
GCS.view$_\mathtt{a}$(v, T)
GCS.view$_\mathtt{b}$(v, T)
MBRSHP.start_change$_\mathtt{b}$(2, {b})
MBRSHP.start_change$_\mathtt{b}$(2, {a, b})
synch msgs
MBRSHP.start_change$_\mathtt{a}$(2, {a, b})
MBRSHP.view$_\mathtt{b}$(3, {a, b}, [a : 2, b : 2])
MBRSHP.view$_\mathtt{b}$(3, {a, b}, [a : 2, b : 2])
GCS.view$_\mathtt{a}$(v′, T′)
GCS.view$_\mathtt{b}$(v′, T′)
Application clients do not need to synchronize their states after new views are delivered

Figure 14: Handling membership changes while synchronization protocol is running.

b *can deduce that, due to Virtually-Synchronous Delivery, they received the same messages while in* v*, and therefore do not need to synchronize their states.*

Example 6.1 demonstrates two additional advantages of our algorithm: (a) the algorithm does not waste resources on synchronizing end-points in order to deliver views that are known to be obsolete; and (b) the application benefits from not seeing obsolete views, as it has to do fewer state synchronizations (or other view processing activity). Responding promptly to connectivity changes is therefore especially important in WANs, where transient connectivity changes may occur frequently due to variability of message latency and less reliable connectivity. In contrast to our algorithm, algorithms that do not allow new members to be added to the membership of an already forming view (such as, [AMMS$^+$95, GVvR96, BDM98, SR93]) lack these advantages, as illustrated by the following example:

Example 6.2 *When executed in the scenario of Example 6.1, algorithms that do not allow new members to be added to the membership of an already forming view would deliver an obsolete view* $\mathtt{v}_{\mathtt{mid}}$ *with membership* {b} *to client* b*, and then re-start the view formation and synchronization protocols anew in order to deliver to* a *and* b *a new view with membership* {a, b}*. As part of the synchronization protocol,* a *and* b *would first exchange messages to agree upon a common identifier before actually exchanging synchronization messages. The synchronization protocol would* not *synchronize end-points* a *and* b *because they would be transitioning into the new view from different views,* a *from* v *and* b *from* $\mathtt{v}_{\mathtt{mid}}$*. As a result, after the clients get the new view from* GCS*, they would have to run an additional state synchronization protocol.*

We are not aware of any other algorithm for Virtual Synchrony that does not pre-agree on common identifiers before sending synchronization messages and that always allows new members to join a forming view while the synchronization protocol is running. Our algorithm achieves these two features by virtue of a simple yet powerful idea: End-points tag their synchronization messages with start-change identifiers that are locally generated by the membership service; when the membership service forms a view and delivers it to the end-points, the view includes information about which start-change identifiers were given to which member. This information communicates to the end-points which synchronization messages they need to consider from each member. Since no pre-agreement upon a common identifier takes place, there is nothing that would inhibit the membership service and the Virtual Synchrony algorithm from allowing new members to join the forming view; end-points just have to forward their last synchronization messages to the joiners.

As a second contribution of this paper, our design has demonstrated how to effectively decouple the algorithm for achieving Virtual Synchrony from the algorithm for maintaining membership. As argued in [ACDK98, KSMD00], such decoupling is important for providing efficient and scalable group communication services in WANs. In previous designs that implement Virtual Synchrony atop an external membership service [BFHR98, SR93], the membership service is not allowed to add new members to an already forming view, and the membership service waits to synchronize with all end-points of the formed view before delivering the view to any of the clients.

A distinct and important characteristic of our design is the high level of formality and rigor at which it has been carried out. This paper has provided precise descriptions of the GCS algorithm and the semantics it provides, as well as a formal proof of the algorithm's correctness – both safety and liveness. Previously, formal approaches based on I/O automata were used to specify the semantics of Virtually Synchronous GCSs and to model and verify their applications, for example, in [Cho97, FLS97, DPFLS98, KFL98, DPFLS99, HLvR99]. However, due to their size and complexity, Virtual Synchrony algorithms were not previously modeled using formal methods, nor were they assertionally verified. Our experience has taught us the importance of careful modeling and verification: in the process of proving our algorithm's correctness we have often uncovered subtleties and ambiguities that had to be resolved.

In order to manage the complexity of our design, we developed a new formal inheritance-based methodology [KKLS00]. The incremental way in which we constructed our algorithms and specifications allowed us to also construct the simulation proof incrementally. For example, in order to prove that VS_RFIFO+TS simulates VS_RFIFO+TS : SPEC we extended the simulation relation from WV_RFIFO to WV_RFIFO : SPEC and reasoned solely about the extension, without repeating the reasoning about the parent components (see Appendix B.2). This reuse was justified by the Proof Extension theorem of [KKLS00] (see Appendix A.3). The use of incremental construction was the key to our success in formally modeling and reasoning about such a complex and sophisticated service. We hope that the methodology employed in this paper shall also be helpful to other researchers working on formal modeling of complex distributed systems.

Acknowledgments

We thank Alan Fekete, Nancy Lynch, Alex Shvartsman, and Jeremy Sussman for many interesting discussions and helpful suggestions.

References

[ABCD96] Y. Amir, D. Breitgand, G. Chockler, and D. Dolev. Group communication as an infrastructure for distributed system management. In *3rd International Workshop on Services in Distributed and Networked Environment (SDNE)*, pages 84–91, June 1996.

[ACDK98] T. Anker, G. Chockler, D. Dolev, and I. Keidar. Scalable group membership services for novel applications. In Marios Mavronicolas, Michael Merritt, and Nir Shavit, editors, *Networks in Distributed Computing (DIMACS workshop)*, volume 45 of *DIMACS*, pages 23–42. American Mathematical Society, 1998.

[ACM96] ACM. *Communications of the ACM 39(4), special issue on Group Communications Systems*, April 1996.

[ACSD00] T. Anker, G. Chockler, I. Shnaiderman, and D. Dolev. The Design of Xpand: A Group Communication System for Wide Area Networks. Technical Report 2000-31, Institute of Computer Science, Hebrew University, Jerusalem, Israel, July 2000.

[ADK99] T. Anker, D. Dolev, and I. Keidar. Fault tolerant video-on-demand services. In *19th International Conference on Distributed Computing Systems (ICDCS)*, pages 244–252, June 1999.

[ADKM92] Y. Amir, D. Dolev, S. Kramer, and D. Malki. Transis: A communication sub-system for high availability. In *22nd IEEE Fault-Tolerant Computing Symposium (FTCS)*, July 1992.

[ADMSM94] Y. Amir, D. Dolev, P. M. Melliar-Smith, and L. E. Moser. Robust and Efficient Replication using Group Communication. Technical Report CS94-20, Institute of Computer Science, Hebrew University, Jerusalem, Israel, 1994.

[AMMS^{+}95] Y. Amir, L. E. Moser, P. M. Melliar-Smith, D. A. Agarwal, and P. Ciarfella. The Totem single-ring ordering and membership protocol. *ACM Transactions on Computer Systems*, 13(4), November 1995.

[BDM98] Ö. Babaoğlu, R. Davoli, and A. Montresor. Group Communication in Partitionable Systems: Specification and Algorithms. TR UBLCS98-1, Department of Computer Science, University of Bologna, April 1998. To appear in IEEE Transactions on Software Engineering.

[BFHR98] K. Birman, R. Friedman, M. Hayden, and I. Rhee. Middleware support for distributed multimedia and collaborative computing. In *Multimedia Computing and Networking (MMCN98)*, 1998.

[Bir96] K. Birman. *Building Secure and Reliable Network Applications.* Manning, 1996.

[BJ87] K. Birman and T. Joseph. Exploiting virtual synchrony in distributed systems. In *11th ACM SIGOPS Symposium on Operating Systems Principles (SOSP)*, pages 123–138. ACM, Nov 1987.

[BvR94] K. Birman and R. van Renesse. *Reliable Distributed Computing with the Isis Toolkit.* IEEE Computer Society Press, 1994.

[CHD98] G. Chockler, N. Huleihel, and D. Dolev. An adaptive totally ordered multicast protocol that tolerates partitions. In *17th ACM Symposium on Principles of Distributed Computing (PODC)*, pages 237–246, June 1998.

[Cho97] G. V. Chockler. An Adaptive Totally Ordered Multicast Protocol that Tolerates Partitions. Master's thesis, Institute of Computer Science, Hebrew University, Jerusalem, Israel, 1997.

[CS95] F. Cristian and F. Schmuck. Agreeing on Process Group Membership in Asynchronous Distributed Systems. Technical Report CSE95-428, Department of Computer Science and Engineering, University of California, San Diego, 1995.

[DLS88] Cynthia Dwork, Nancy Lynch, and Larry Stockmeyer. Consensus in the presence of partial synchrony. *Journal of the ACM*, 35(2):288–323, April 1988.

[DM96] D. Dolev and D. Malkhi. The Transis approach to high availability cluster communication. *Communications of the ACM*, 39(4):64–70, April 1996.

[DMS94] D. Dolev, D. Malki, and H. R. Strong. An Asynchronous Membership Protocol that Tolerates Partitions. Technical Report CS94-6, Institute of Computer Science, Hebrew University, Jerusalem, Israel, 1994.

[DP99] R. De Prisco. *On building blocks for distributed systems.* PhD thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, USA, December 1999.

[DPFLS98] R. De Prisco, A. Fekete, N. Lynch, and A. Shvartsman. A dynamic view-oriented group communication service. In *17th ACM Symposium on Principles of Distributed Computing (PODC)*, pages 227–236, June 1998.

[DPFLS99] R. De Prisco, A. Fekete, N. Lynch, and A. Shvartsman. A dynamic primary configuration group communication service. In *13th International Symposium on DIStributed Computing (DISC)*, pages 64–78, Bratislava, Slovak Republic, 1999.

[FLS97] A. Fekete, N. Lynch, and A. Shvartsman. Specifying and using a partitionable group communication service. In *16th ACM Symposium on Principles of Distributed Computing (PODC)*, pages 53–62, August 1997. Full version to appear in ACM Transactions on Computer Systems (TOCS).

[FV97] R. Friedman and A. Vaysburg. High-performance replicated distributed objects in partitionable environments. Technical Report 97-1639, Dept. of Computer Science, Cornell University, Ithaca, NY 14850, USA, July 1997.

[FvR95] Roy Friedman and Robbert van Renesse. Strong and Weak Virtual Synchrony in Horus. TR 95-1537, dept. of Computer Science, Cornell University, August 1995.

[GS97] R. Guerraoui and A. Schiper. Consensus: the big misunderstanding. In *Proceedings of the 6th IEEE Computer Society Workshop on Future Trends in Distributed Computing Systems (FTDCS-6)*, pages 183–188, Tunis, Tunisia, October 1997. IEEE Computer Society Press.

[GVvR96] Katherine Guo, Werner Vogels, and Robbert van Renesse. Structured virtual synchrony: Exploring the bounds of virtual synchronous group communication. In *7th ACM SIGOPS European Workshop*, September 1996.

[HLvR99] Jason Hickey, Nancy Lynch, and Robbert van Renesse. Specifications and proofs for ensemble layers. In *5th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, LNCS. Springer-Verlag, March 1999.

[HS95] M. Hiltunen and R. Schlichting. Properties of membership services. In *2nd International Symposium on Autonomous Decentralized Systems*, pages 200–207, 1995.

[KD96] I. Keidar and D. Dolev. Efficient message ordering in dynamic networks. In *15th ACM Symposium on Principles of Distributed Computing (PODC)*, pages 68–76, May 1996.

[KFL98] Roger Khazan, Alan Fekete, and Nancy Lynch. Multicast group communication as a base for a load-balancing replicated data service. In *12th International Symposium on DIStributed Computing (DISC)*, pages 258–272, Andros, Greece, September 1998.

[KKLS00] Idit Keidar, Roger Khazan, Nancy Lynch, and Alex Shvartsman. An inheritance-based technique for building simulation proofs incrementally. In *22nd International Conference on Software Engineering (ICSE)*, pages 478–487, June 2000.

[KSMD00] I. Keidar, J. Sussman, K. Marzullo, and D. Dolev. A Client-Server Oriented Algorithm for Virtually Synchronous Group Membership in WANs. In *20th International Conference on Distributed Computing Systems (ICDCS)*, pages 356–365, April 2000. Full version: MIT Technical Memorandum MIT-LCS-TM-593a.

[Lam97] Butler Lampson. Generalizing Abstraction Functions. Massachusetts Institute of Technology, Laboratory for Computer Science, principles of computer systems class, handout 8, 1997. ftp://theory.lcs.mit.edu/pub/classes/6.826/www/6.826-top.html.

[Lam78] L. Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558–565, July 78.

[LT89] N.A. Lynch and M.R. Tuttle. An introduction to Input/Output Automata. *CWI Quarterly*, 2(3):219–246, 1989.

[Lyn96] N.A. Lynch. *Distributed Algorithms.* Morgan Kaufmann Publishers, 1996.

[MAMSA94] L. E. Moser, Y. Amir, P. M. Melliar-Smith, and D. A. Agarwal. Extended virtual synchrony. In *14th International Conference on Distributed Computing Systems (ICDCS)*, pages 56–65, June 1994. Full version: technical report ECE93-22, Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA.

[RGS+96] Luis Rodrigues, Katherine Guo, Antonio Sargento, Robbert van Renesse, Brad Glade, Paulo Verissimo, and Ken Birman. A dynamic light-weight group service. In *15th IEEE International Symposium on Reliable Distributed Systems (SRDS)*, pages 23–25, October 1996. also Cornell University Technical Report, TR96-1611, August, 1996.

[Sch90] F. B. Schneider. Implementing fault tolerant services using the state machine approach: A tutorial. *ACM Computing Surveys*, 22(4):299–319, December 1990.

[SM98] J. Sussman and K. Marzullo. The *Bancomat* problem: An example of resource allocation in a partitionable asynchronous system. In *12th International Symposium on DIStributed Computing (DISC)*, September 1998. Full version: Tech Report 98-570 University of California, San Diego Department of Computer Science and Engineering.

[SR93] A. Schiper and A.M. Ricciardi. Virtually synchronous communication based on a weak failure suspector. In *23rd IEEE Fault-Tolerant Computing Symposium (FTCS)*, pages 534–543, June 1993.

[Tar00] Igor Tarashchanskiy. Virtual Synchrony Semantics: Client-Server Implementation. Master's thesis, MIT Department of Electrical Engineering and Computer Science, August 2000. Master of Engineering.

[VKCD99] R. Vitenberg, I. Keidar, G. V. Chockler, and D. Dolev. Group Communication Specifications: A Comprehensive Study. Technical Report CS99-31, Institute of Computer Science, Hebrew University, Jerusalem, Israel, September 1999. Also Technical Report MIT-LCS-TR-790, Massachusetts Institute of Technology, Laboratory for Computer Science and Technical Report CS0964, Computer Science Department, the Technion, Haifa, Israel.

[vRBM96] R. van Renesse, K. P. Birman, and S. Maffeis. Horus: A flexible group communication system. *Communications of the ACM*, 39(4):76–83, April 1996.

A Review of Proof Techniques

In this section we describe the main techniques used to prove correctness of I/O automata: invariant assertions, hierarchical proofs, refinement mappings, and history and prophecy variables. The material in this section is closely based on [Lyn96, pages 216-228] and [Lam97, pages 3,4, and 13]. In Section A.3 we present a proof-extension theorem of [KKLS00] that provides a formal framework for the reuse of simulation proofs based on refinement mappings.

A.1 Invariants

The most fundamental type of property to be proved about an automaton is an *invariant assertion*, or just *invariant*, for short. An invariant assertion of an automaton A is defined as any property that is true in every single reachable state of A.

Invariants are typically proved by induction on the number of steps in an execution leading to the state in question. While proving an inductive step, we consider only *critical actions*, which affect the state variables appearing in the invariant.

A.2 Hierarchical Proofs

One of the important proof strategies is based on a hierarchy of automata. This hierarchy represents a series of descriptions of a system or algorithm, at different levels of abstraction. The process of moving through the series of abstractions, from the highest level to the lowest level, is known as *successive refinement*. The top level may be nothing more than a problem specification written in the form of an automaton. The next level is typically a very abstract representation of the system: it may be centralized rather than distributed, or have actions with large granularity, or have simple but inefficient data structures. Lower levels in the hierarchy look more and more like the actual system or algorithm that will be used in practice: they may be more distributed, have actions with small granularity, and contain optimizations. Because of all this extra detail, lower levels in the hierarchy are usually harder to understand than the higher levels. The best way to prove properties of the lower-level automata is by relating these automata to automata at higher levels in the hierarchy, rather than by carrying out direct proofs from scratch.

A.2.1 Refinement Mappings

The simplest way to relate two automata, say A and S, is to present a *refinement mapping* R from the reachable states of A to the reachable state of S such that it satisfies the following two conditions:

1. If t_0 is an initial state of A, then $R(s_0)$ is an initial state of S.

2. If t and $R(t)$ are reachable states of A and S respectively, and (t, π, t') is a step of A, then there exists an execution fragment of S beginning at state $R(t)$ and ending at state $R(t)'$, with its trace being the same as the trace of π and its final state $R(t)'$ being the same as $R(t')$.

The first condition asserts that any initial state of A has some corresponding initial state of S. The second condition asserts that any step of A has a corresponding sequence of steps of S. This corresponding sequence can consist of one step, many steps, or even no steps, as long as the correspondence between the states is preserved and the external behavior is the same.

The following theorem gives the key property of refinement mappings:

Theorem A.1 *If there is a refinement mapping from A to S, then* `traces`*(A)* $\subseteq$ `traces`*(S).*

If automata A and S have the same external signature and the traces of A are the traces of S, then we say that *A implements S in the sense of trace inclusion*, which means that A never does anything that S couldn't do. Theorem A.1 implies that, in order to prove that one automaton implements another in the sense of trace inclusion, it is enough to produce a refinement mapping from the former to the latter.

A.2.2 History and Prophecy Variables

Sometimes, however, even when the traces of A are the traces of S, it is not possible to give a refinement mapping from A to S. This may happen due to the following two generic reasons:

- The states of S may contain more information than the states of A.
- S may make some premature choices, which A makes later.

The situation when A has been optimized not to retain certain information that S maintains can be resolved by augmenting the state of A with additional components, called *history variables* (because they keep track of additional information about the history of execution), subject to the following constraints:

1. Every initial state has at least one value for the history variables.
2. No existing step is disabled by the addition of predicates involving history variables.
3. A value assigned to an existing state component must not depend on the value of a history variable.

These constraints guarantee that the history variables simply record additional state information and do not otherwise affect the behavior exhibited by the automaton. If the automaton A_{HV} augmented with history variables can be shown to implement S by presenting a refinement mapping, it follows that the original automaton A without the history variables also implements S, because they have the same traces.

The situation when S is making a premature choice, which A makes later, can be resolved by augmenting A with a different sort of auxiliary variable, *prophecy variable*, which can look into the future just as history variable looks into the past. A prophecy variable guesses in advance some non-deterministic choice that A is going to make later. The guess gives enough information to construct a refinement mapping to S (which is making the premature choice). For an added variable to be a prophecy variable, it must satisfy the following conditions:

1. Every state has at least one value for the prophecy variable.
2. No existing step is disabled *in the backward direction* by the new preconditions involving a prophecy variable. More precisely, for each step (t, π, t') there must be a state (t, p) and a p such that there is a step $((t, p), \pi, (t', p'))$.
3. A value assigned to an existing state component must not depend on the value of the prophecy variable.
4. If t is an initial state of A and (t, p) is a state of the A augmented with the prophecy variable, then it must be its initial state.

If these conditions are satisfied, the automaton augmented with the prophecy variable will have the same (finite) traces as the automaton without it. Therefore, if we can exhibit a refinement mapping from A_{PV} to S, we know that the A implements S.

A.3 Inheritance and Proof Extension Theorem

We now present a theorem from [KKLS00] which lays the foundation for incremental proof construction. Consider the example illustrated in Figure 15, where a refinement mapping $\mathtt{R}$ from an algorithm $\mathtt{A}$ to a specification $\mathtt{S}$ is given, and we want to construct a refinement mapping $\mathtt{R}'$ from a child $\mathtt{A}'$ of an automaton $\mathtt{A}$ to a child $\mathtt{S}'$ of a specification automaton $\mathtt{S}$.

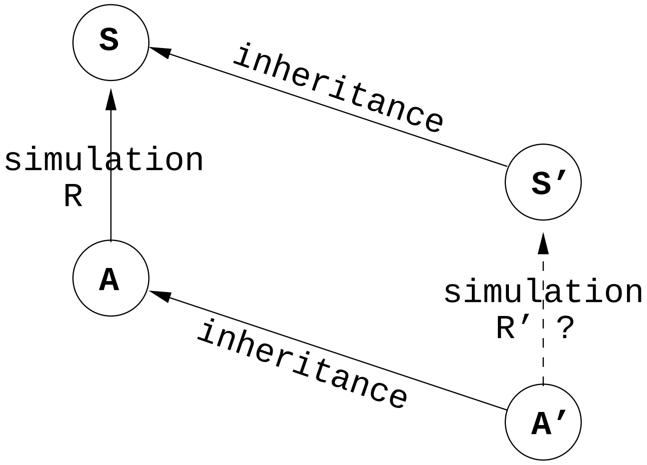


Figure 15: Algorithm $\mathtt{A}$ simulates specification $\mathtt{S}$ with $\mathtt{R}$. Can $\mathtt{R}$ be reused for building a refinement $\mathtt{R}'$ from a child $\mathtt{A}'$ of $\mathtt{A}$ to a child $\mathtt{S}'$ of $\mathtt{S}$?

Theorem A.2 below implies that such a refinement $\mathtt{R}'$ can be constructed by supplementing $\mathtt{R}$ with a mapping $\mathtt{R_n}$ from the states of $\mathtt{A}'$ to the state extension introduced by $\mathtt{S}'$. Mapping $\mathtt{R_n}$ has to map every initial state of $\mathtt{A}'$ to some initial state extension of $\mathtt{A}'$ and it has to satisfy a step condition similar to the one for refinement mapping (Section A.2.1), but only involving the transition restriction of $\mathtt{S}'$.

Theorem A.2 *Let automaton $\mathtt{A}'$ be a child of automaton $\mathtt{A}$. Let automaton $\mathtt{S}'$ be a child of automaton $\mathtt{S}$. Let mapping $\mathtt{R}$ be a refinement from $\mathtt{A}$ to $\mathtt{S}$.*

Let $\mathtt{R_n}$ be a mapping from the states of $\mathtt{A}'$ to the state extension introduced by $\mathtt{S}'$.

A mapping $\mathtt{R}'$ from the states of $\mathtt{A}'$ to the states of $\mathtt{S}'$, defined in terms of $\mathtt{R}$ and $\mathtt{R_n}$ as

$$\mathtt{R}'(\langle \mathtt{t}, \mathtt{t_n} \rangle) = \langle \mathtt{R}(\mathtt{t}), \mathtt{R_n}(\langle \mathtt{t}, \mathtt{t_n} \rangle) \rangle$$

is a refinement *from $\mathtt{A}'$ to $\mathtt{S}'$ if $\mathtt{R}'$ satisfies the following two conditions:*

1. *If $\mathtt{t}$ is an initial state of $\mathtt{A}'$, then $\mathtt{R_n}(\mathtt{t})$ is an initial state extension of $\mathtt{S}'$.*
2. *If $\langle \mathtt{t}, \mathtt{t_n} \rangle$ is a reachable state of $\mathtt{A}'$, $\mathtt{s} = \langle \mathtt{R}(\mathtt{t}), \mathtt{R_n}(\langle \mathtt{t}, \mathtt{t_n} \rangle) \rangle$ is a reachable state of $\mathtt{S}'$, and $(\langle \mathtt{t}, \mathtt{t_n} \rangle, \pi, \langle \mathtt{t}', \mathtt{t}'_\mathtt{n} \rangle)$ is a step of $\mathtt{A}'$, then there exists a finite sequence α of alternating states and actions of $\mathtt{S}'$, beginning from $\mathtt{s}$ and ending at some state $\mathtt{s}'$, and satisfying the following conditions:*
   (a) *α projected onto states of $\mathtt{S}$ is an execution sequence of $\mathtt{S}$.*
   (b) *Every step $(\mathtt{s_i}, \sigma, \mathtt{s_{i+1}}) \in \alpha$ is consistent with the transition restriction placed on $\mathtt{S}$ by $\mathtt{S}'$.*
   (c) *The parent component of the final state $\mathtt{s}'$ is $\mathtt{R}(\mathtt{t}')$.*
   (d) *The child component of the final state $\mathtt{s}'$ is $\mathtt{R_n}(\langle \mathtt{t}', \mathtt{t}'_\mathtt{n} \rangle)$.*

(e) α has the same trace as π.

In practice, one would exploit this theorem as follows: The simulation proof between the parent automata already provides a corresponding execution sequence of the parent specification for every step of the parent algorithm. It is typically the case that the same execution sequence, padded with new state variables, corresponds to the same step at the child algorithm. Thus, conditions 2a, 2c, and 2e of Theorem A.2 hold for this sequence. The only conditions that have to be checked are 2b, and 2d, that is, that every step of this execution sequence is consistent with the transition restriction placed on $\mathtt{S}$ by $\mathtt{S}'$ and that the values of the new state variables of $\mathtt{S}'$ in the final state of this execution match those obtained when $\mathtt{R_n}$ is applied to the post-state of the child algorithm.

A.4 Safety versus Liveness

Proving that one automaton implements another in the sense of trace inclusion constitutes only *partial correctness*, as it implies *safety* but not *liveness*. In other words, partial correctness ensures than "bad" things never happen, but it does not say anything whether some "good" thing eventually happens.

In this paper, we use invariant assertions and simulation techniques to prove that our algorithms satisfy safety properties, which are stated as I/O automata. For liveness proofs, we use a combination of invariant assertions and carefully proven operational arguments.

B Correctness Proof: Safety Properties

We now formally prove using invariant assertions and simulations that our algorithms satisfies the safety properties of Section 4.1. Proofs done with invariant assertions and simulations are easily verifiable (even by a computer) because they involve reasoning only about single steps of the algorithm. A review of the used in this section proof techniques appears in Appendix A.

The safety proof is *modular*: we exploit the inheritance-based structure of our specifications and algorithms to reuse proofs. In Section B.1 we prove correctness of the within-view reliable FIFO multicast service by showing a refinement mapping from WV_RFIFO to WV_RFIFO : SPEC. In Section B.2 we extend this refinement mapping to map the new state added in VS_RFIFO+TS to that in VSRFIFO : SPEC. In Section B.3 we prove that VS_RFIFO+TS also simulates TS : SPEC. Finally, in Section B.4 we extend the refinement above to map the new state of GCS to that of SELF : SPEC. The proof-extension theorem of [KKLS00] (also reviewed in Appendix A) implies that the GCS automaton satisfies WV_RFIFO : SPEC, VSRFIFO : SPEC, TS : SPEC, and SELF : SPEC.

B.1 Within-view reliable FIFO multicast

Intuitively, in order to simulate WV_RFIFO : SPEC with WV_RFIFO, we need to show that WV_RFIFO satisfies Self Inclusion and Local Monotonicity for delivered views, and we need to show that the i'th message delivered by $\mathtt{q}$ from $\mathtt{p}$ in view $\mathtt{v}$ is the i'th message sent in view $\mathtt{v}$ by the client at $\mathtt{p}$. In order to prove this, we need to show that the algorithm correctly associates messages with the views in which they were sent and with their indices in the sequences of messages sent in these views. We split the proof into three parts: Section B.1.1 states key invariants, but defers the proof of one of them to Section B.1.3; Section B.1.2 contains the simulation proof.

B.1.1 Key Invariants

The following invariant captures the Self Inclusion property.

Invariant B.1 (Self-Inclusion) *In every reachable state* $\texttt{s}$ *of* WV_RFIFO*, for all* $\texttt{Proc p}$*,* $\texttt{p} \in \texttt{s[p].mbrshp\_view.set}$ *and* $\texttt{p} \in \texttt{s[p].current\_view.set}$.

Proof B.1: Immediate from the MBRSHP specification. ■

The Local Monotonicity property follows directly from the precondition, $\texttt{v.id} > \texttt{mbrshp\_view}$, of the MBRSHP.$\texttt{view}_\texttt{p}(\texttt{v})$ actions.

The following invariant relates application messages at different end-points' queues to the corresponding messages on the original senders' queues.

Invariant B.2 (Message Consistency) *In every reachable state* $\texttt{s}$ *of* WV_RFIFO*, for all* $\texttt{Proc p}$ *and* $\texttt{Proc q}$*, if* $\texttt{s[q].msgs[p][v][i] = m}$*, then* $\texttt{s[p].msgs[p][v][i] = m}$.

This proposition is vacuously true in the initial state because all message queues are empty. For the inductive step, we have to consider actions $\texttt{co\_rfifo.deliver}_{\texttt{q,p}}(\texttt{tag} = \textbf{app\_msg}, \texttt{m})$ and $\texttt{co\_rfifo.deliver}_{\texttt{q,p}}(\texttt{tag} = \textbf{fwd\_msg}, \texttt{r}, \texttt{v}, \texttt{m}, \texttt{i})$, and have to argue that the message $\texttt{m}$ they deliver is placed in the right place in $\texttt{q}$'s $\texttt{msgs}$ buffer. The proof of this invariant appears in Section B.1.3, after the simulation proof.

B.1.2 Simulation

Lemma B.1 *The following function* $\texttt{R()}$ *is a* refinement mapping *from automaton* WV_RFIFO *to automaton* WV_RFIFO : SPEC *with respect to their reachable states.*

$$\begin{array}{lrcl}
\multicolumn{4}{l}{\texttt{R(s} \in \texttt{ReachableStates(WV\_RFIFO))} = \texttt{t} \in \texttt{ReachableStates(WV\_RFIFO : SPEC), where}} \\
\quad \texttt{For each Proc p, View v:} & \texttt{t.msgs[p][v]} & = & \texttt{s[p].msgs[p][v]} \\
\quad \texttt{For each Proc p, Proc q:} & \texttt{t.last\_dlvrd[p][q]} & = & \texttt{s[q].last\_dlvrd[p]} \\
\quad \texttt{For each Proc p:} & \texttt{t.current\_view[p]} & = & \texttt{s[p].current\_view}
\end{array}$$

Proof B.1:

Action Correspondence: Automaton WV_RFIFO : SPEC has three types of actions. Actions of the types $\texttt{view}_\texttt{p}(\texttt{v})$, $\texttt{send}_\texttt{p}(\texttt{m})$, and $\texttt{deliver}_\texttt{p}(\texttt{q},\texttt{m})$, are simulated when WV_RFIFO takes the corresponding $\texttt{view}_\texttt{p}(\texttt{v})$, $\texttt{send}_\texttt{p}(\texttt{m})$, and $\texttt{deliver}_\texttt{p}(\texttt{q},\texttt{m})$ actions. Steps of WV_RFIFO involving other actions correspond to empty steps of WV_RFIFO : SPEC.

Simulation Proof: In the most part the simulation proof is straightforward. Here, we present only the interesting steps:

The fact that the corresponding step of WV_RFIFO : SPEC is enabled when WV_RFIFO takes a step involving $\texttt{view}_\texttt{p}(\texttt{v})$ relies on $\texttt{p} \in \texttt{mbrshp\_view.set}$ (Invariant B.1).

For steps involving $\texttt{deliver}_\texttt{p}(\texttt{q},\texttt{m})$, to deduce that the corresponding step of WV_RFIFO : SPEC is enabled, we need to know that the message at index $\texttt{s[p].last\_dlvrd[q] + 1}$ at end-point $\texttt{p}$'s $\texttt{s[p].msgs[q][s[p].current\_view]}$ is the same message that end-point $\texttt{q}$ has on its corresponding queue at the same index. This property is implied by Invariant B.2.

Steps that involve receiving original and forwarded application messages from the network simulate empty steps of WV_RFIFO : SPEC. Among these steps the only critical ones are those that deliver a message from p to p because they may affect s[p].msgs[p][p] queue. Since end-points do not send messages to themselves (CO_RFIFO.send$_p$(set, tag = **app_msg**, m) is preconditioned by set = s[p].current_view.set − {p}, and co_rfifo.send$_p$(set, tag = **fwd_msg**, r, v, m, i) is preconditioned by p $\notin$ set), such steps may not happen. ■

From Lemma B.1 and Theorem A.1 we conclude the following:

Theorem B.1 WV_RFIFO *implements* WV_RFIFO : SPEC *in the sense of trace inclusion.*

B.1.3 Auxiliary Invariants

We now state and prove a number of auxiliary invariants necessary for the proof of the key message consistency invariant (Invariant B.2).

In any view, before an end-point sends a view_msg to others (and hence before it sends any application message to others) it tells CO_RFIFO to maintain reliable connection to every member of its current view. The following invariant captures this property.

Invariant B.3 (Connection Reliability) *In every reachable state* s *of* WV_RFIFO*, for all* Proc p*, if* s[p].current_view = s[p].view_msg[p]*, then* s[p].current_view.set $\subseteq$ s[p].reliable_set.

Proof B.3: By induction on the length of the execution sequence; follows directly from the code. ■

After an end-point delivers a new view to its client, it sends a view_msg to other members of the view. The stream of view_msgs that an end-point sends to others is monotonic because the delivered views satisfy Local Monotonicity. The following invariant captures this property. It states that the subsequence of messages in transit from end-point p to end-point q consisting solely of the view_msgs is monotonically increasing. It also relates the current view of an end-point p to the view contained in the p's latest view_msg to q.

Invariant B.4 (Monotonicity of View Messages) *Let* s *be a reachable state of* WV_RFIFO*. Consider the subsequence of messages in* s.channel[p][q] *for which* m.tag=**view_msg***. We examine the sequence of views included in these view messages, and construct a new sequence* seq *of views by pre-pending this view sequence with the element* s[q].view_msg[p]*. For all* Proc p, Proc q*, the following propositions are true:*

1. *The sequence* seq *is (strictly) monotonically increasing.*
2. *If* s[p].current_view $\neq$ s[p].view_msg[p]*, then* s[p].current_view *is strictly greater then the last (largest) element of* seq.
3. *If* s[p].current_view = s[p].view_msg[p]*, and if* q $\in$ s[p].current_view.set*, then* s[p].current_view *is equal to the last (largest) element of* seq.

Proof B.4: All three propositions are true in the initial state. We now consider steps involving the critical actions:
CO_RFIFO.lose(p, q): The first two propositions remain true because this action throws away only the last message from the CO_RFIFO s.channel[p][q].
The third proposition is vacuously true because q can not be in s[p].current_view.set. If it were, the CO_RFIFO.lose(p, q) action would not be enabled because Invariant B.3 would imply that s[p].current_view.set is a subset of s[p].reliable_set, which would then imply that q $\in$ s.reliable_set[p] (because s[p].reliable_set = s.reliable_set[p], as can be shown by straightforward induction).

$\mathtt{view}_{\mathtt{p}}(\mathtt{v})$: The first proposition is unaffected. The second proposition follows from the inductive hypothesis and the precondition $\mathtt{v.id} > \mathtt{s[p].current\_view.id}$. The third proposition is vacuously true because $\mathtt{s[p].current\_view} \neq \mathtt{s[p].view\_msg[p]}$ as follows from the precondition $\mathtt{v.id} > \mathtt{s[p].current\_view.id}$ and the fact that, in every reachable state s, s[p].current_view $\geq$ s[p].view_msg[p] (can be proved by straightforward induction).

CO_RFIFO.$\mathtt{send}_{\mathtt{p}}(\mathtt{set}, \mathtt{tag} = \textbf{view\_msg}, \mathtt{v})$: The first proposition is true in the post-state because of the inductive hypothesis of the second proposition. The second proposition is vacuously true in the post-state. The third proposition is true in the post-state because of the effect of this action.

CO_RFIFO.$\mathtt{deliver}_{\mathtt{p,q}}(\mathtt{tag} = \textbf{view\_msg}, \mathtt{v})$: It is straightforward to see that all three propositions remain true in the post-state.

■

History Tags

In order to reason about original application messages traveling on CO_RFIFO channels we need a way to reference, for each of these messages, the view in which it was originally sent and its index in the FIFO-ordered sequence of messages sent in that view. To this end, we augment each original application message ⟨ tag=**app_msg**, m ⟩ with two *history tags*, Hv and Hi, that are set to current_view and last_sent + 1 respectively when CO_RFIFO.$\mathtt{send}_{\mathtt{p}}(\mathtt{set}, \mathtt{tag} = \textbf{app\_msg}, \mathtt{m})$ occurs. (See Appendix A for details on history variables).

```
OUTPUT co_rfifo.send_p(set, tag=app_msg, m, Hv, Hi)
pre: ...
     Hv = current_view
     Hi = last_sent + 1
 eff: ...
```

With the addition of these history tags, the interface between WV_RFIFO and CO_RFIFO for handling original application messages becomes CO_RFIFO.$\mathtt{send}_{\mathtt{p}}(\mathtt{set}, \mathtt{tag} = \textbf{app\_msg}, \mathtt{m}, \mathtt{Hv}, \mathtt{Hi})$ and CO_RFIFO.$\mathtt{deliver}_{\mathtt{p,q}}(\mathtt{tag} = \textbf{app\_msg}, \mathtt{m}, \mathtt{Hv}, \mathtt{Hi})$.

The goal of the next three invariants is to show that, when end-point q receives an application message m tagged with a history view Hv and a history index Hi, the current value of q's view_msg[p] equals Hv and that of last_rcvd[p] + 1 equals Hi.

Invariant B.5 (History View Consistency) *In every reachable state* s *of* WV_RFIFO*, for all* Proc p, Proc q*, the following is true: For all messages* ⟨ tag=**app_msg**, m, Hv, Hi ⟩ *on the* CO_RFIFO s.channel[p][q]*, view* Hv *equals either the view of the closest preceding view message on* s.channel[p][q] *if there is such, or* s[q].view_msg[p] *otherwise.*

Proof B.5: By induction. The step involving a CO_RFIFO.$\mathtt{send}_{\mathtt{p}}(\mathtt{set}, \mathtt{tag} = \textbf{app\_msg}, \mathtt{m}, \mathtt{Hv}, \mathtt{Hi})$ action follows directly from Invariant B.4 Part 3. The proposition is not affected by steps involving CO_RFIFO.$\mathtt{lose}(\mathtt{p}, \mathtt{q})$ because those may only remove the last messages from the CO_RFIFO s.channel[p][q]. The other steps are straightforward. ■

The following invariant states that the value of s[p].last_sent equals to the number of application messages that p sent in its current view and that are either still in transit on the CO_RFIFO s.channel[p][q] or are already received by q.

Invariant B.6 *In every reachable state* s *of* WV_RFIFO*, for all* Proc p *and for all* Proc q $\in$ s[p].current_view.set $-$ {p}*, the following is true:*

$$\begin{aligned}&\texttt{s[p].last\_sent} = \\ &\qquad \left|\{\texttt{msg} \in \texttt{s.channel[p][q]} \ : \ \texttt{msg.tag=}\textbf{app\_msg} \texttt{ and msg.Hv = s[p].current\_view}\}\right| + \\ &\qquad + \begin{cases} \texttt{s[q].last\_rcvd[p]} & \textit{if } \texttt{s[q].view\_msg[p] = s[p].current\_view} \\ 0 & \textit{otherwise.} \end{cases}\end{aligned}$$

Proof : By induction. Consider steps involving the following critical actions:

CO_RFIFO.lose(p, q): Assume that the last message on s.channel[p][q] is an application message msg with msg.Hv = s[p].current_view. If a step involving CO_RFIFO.lose(p, q) action could occur, then the proposition would be false. However, as we are going to argue now, q $\in$ s.reliable_set[p], so such a step cannot occur.

We can prove by straightforward induction that msg $\in$ s.channel[p][q] implies s[p].view_msg[p] = s[p].current_view. By invariant B.3, s[p].current_view.set $\subseteq$ s[p].reliable_set. Since q $\in$ s[p].current_view.set and s[p].reliable_set = s.reliable_set[p], it follows that q $\in$ s.reliable_set[p].

$\texttt{view}_\texttt{p}(\texttt{v})$: The proposition remains true for steps involving $\texttt{view}_\texttt{p}(\texttt{v})$ action because its effect sets s′[p].last_sent to 0 and because both summands of the right hand side of the equation also becomes 0. Indeed, the first summand becomes 0 because CO_RFIFO channels never have messages tagged with views that are larger then the current views of the messages' senders (as can be shown by a simple inductive proof); the second summand becomes 0 because Invariant B.4 Part 2 implies that s′[q].view_msg[p] $\neq$ s′[p].current_view.

CO_RFIFO.$\texttt{deliver}_{\texttt{p,q}}$(tag = **view_msg**, v): The proposition remains true for steps involving this action because s[q].view_msg[p] $\neq$ s[p].current_view, as follows immediately from Invariant B.4.

CO_RFIFO.$\texttt{send}_\texttt{p}$(set, tag = **app_msg**, m, Hv, Hi) and

CO_RFIFO.$\texttt{deliver}_{\texttt{p,q}}$(tag = **app_msg**, m, Hv, Hi): For steps involving these actions the truth of the proposition follows immediately from the effects of these actions, the inductive hypotheses, and Invariant B.5.

■

The history index attached to an original application message m sent in a view Hv that is in transit on a CO_RFIFO channel to end-point q is equal to the number of such messages (including m) that precede m on that channel, plus those (if any) that q has already received.

Invariant B.7 (History Indices Consistency) *In every reachable state* s *of* WV_RFIFO*, for all* Proc p *and* Proc q*, if* ⟨ tag=**app_msg**, m, Hv, Hi ⟩ = s.channel[p][q][j] *for some index* j*, then*

$$\begin{aligned}\texttt{Hi} \ = \ &\left|\{\texttt{msg} \in \texttt{s.channel[p][q][.. j]} \ : \ \texttt{msg.tag=}\textbf{app\_msg} \texttt{ and msg.Hv = Hv}\}\right| + \\ &+ \begin{cases} \texttt{s[q].last\_rcvd[p]} & \textit{if } \texttt{s[q].view\_msg[p] = Hv} \\ 0 & \textit{otherwise.} \end{cases}\end{aligned}$$

Proof B.7: In the initial state s.channel[p][q] is empty. For the inductive step, we consider steps involving the following critical actions:

CO_RFIFO.lose(p, q): The proposition remains true since CO_RFIFO.lose(p, q) discards only the last messages from the CO_RFIFO s.channel[p][q].

CO_RFIFO.$\mathtt{deliver}_{\mathtt{p,q}}$(tag = **view_msg**, v): We have to consider the effects on two types of application messages: those associated with view s[q].view_msg[p], and those associated with view Hv. Invariants B.4 Part 1 and B.5 imply that there are no application messages with msg.Hv = s[q].view_msg[p] on the CO_RFIFO channel[p][q]. Thus, the proposition does not apply for such messages. For those messages that have msg.Hv = Hv, the proposition remains true because s′[q].last_rcvd[p] is set to 0 as a result of this action.

CO_RFIFO.$\mathtt{deliver}_{\mathtt{p,q}}$(tag = **app_msg**, m, Hv, Hi): Follows immediately from the effect of this action, the inductive hypothesis, and Invariant B.5.

CO_RFIFO.$\mathtt{send}_{\mathtt{p}}$(set, tag = **app_msg**, m, Hv, Hi): The inductive step follows immediately from the inductive hypothesis and Invariant B.6. ■

We now prove a generalization of Invariant B.2, which relates application messages either in transit on the CO_RFIFO channels or at end-points' queues to their corresponding messages on the senders' queues.

Invariant B.8 (General Message Consistency) *In every reachable state* s *of* WV_RFIFO*, for all* Proc p *and* Proc q*, the following are true:*

1. *If* ⟨ tag=**app_msg**, m, Hv, Hi ⟩ ∈ s.channel[p][q]*, then* s[p].msgs[p][Hv][Hi] = m*.*
2. *If* ⟨ tag=**fwd_msg**, r, m, v, i ⟩ ∈ s.channel[p][q]*, then* s[r].msgs[r][v][i] = m*.*
3. *If* s[q].msgs[p][v][i] = m*, then* s[p].msgs[p][v][i] = m*.*

Proof B.8:

Basis: In the initial state all message queues are empty.

Inductive Step: The following are the critical actions:

```
send_p(m),
co_rfifo.send_p(set, tag=app_msg, m, Hv, Hi),
co_rfifo.deliver_q,p(tag=app_msg, m, Hv, Hi),
co_rfifo.send_p(set, tag=fwd_msg, r, v, m, i),
co_rfifo.deliver_q,p(tag=fwd_msg, r, v, m, i).
```

For steps involving CO_RFIFO.$\mathtt{deliver}_{\mathtt{q,p}}$(tag = **app_msg**, m, Hv, Hi), we use Invariants B.5 and Invariant B.7, which respectively imply that history view Hv equals s[p].view_msg[q] and that history index Hi equals s[p].last_rcvd[q] + 1. Inductive steps involving each of the other actions are straightforward. ■

Invariant B.2 is a private case of this invariant.

B.2 Virtual Synchrony

We now show that automaton VS_RFIFO+TS simulates VSRFIFO : SPEC. We prove this by extending the refinement above using the Proof Extension Theorem of [KKLS00] (see Appendix A for details).

B.2.1 Invariants

We prove that end-points that move together from one view to the next consider the same synchronization messages and thus compute the same transitional sets and use the same cuts from the members of the transitional set.

Invariant B.9 *In every reachable state* s *of* VS_RFIFO+TS*, for all* Proc p*,* Proc q*, and for every* StartChangeId cid*,*

if $\texttt{s[q].sync\_msg[p][cid]} \neq \bot$*, then* s[q].sync_msg[p][cid] = s[p].sync_msg[p][cid]*.*

Proof B.9: The proposition is true in the initial state s_0 as all $s_0[q].\texttt{sync\_msg}[p][\texttt{cid}] = \bot$. The inductive step involving a $\texttt{set\_cut}_p()$ action is trivial, for it only affects the case q = p. The inductive step involving a CO_RFIFO.$\texttt{deliver}_{p,q}(\texttt{tag} = \textbf{sync\_msg}, \texttt{cid}, \texttt{v}, \texttt{cut})$ action follows immediately from the following proposition:

$$\langle \texttt{tag=}\textbf{sync\_msg}\texttt{, cid, v, cut}\rangle \in \texttt{s.channel[p][q]} \Rightarrow \texttt{s[p].sync\_msg[p][cid]} = \langle \texttt{v, cut}\rangle,$$

which can be proved by straightforward induction. Indeed, there are two critical actions:
CO_RFIFO.$\texttt{send}_p(\texttt{set}, \texttt{tag} = \textbf{sync\_msg}, \texttt{cid}, \texttt{v}, \texttt{cut})$ – immediate from the code, and
CO_RFIFO.$\texttt{deliver}_{p,p}(\texttt{tag} = \textbf{sync\_msg}, \texttt{cid}, \texttt{v}, \texttt{cut})$ – may not occur because end-points do not send synchronization messages to themselves. ■

Corollary B.1 *End-points that move together from one view to the next, use the same sets of synchronization messages to calculate transitional sets and message cuts.*

Proof : Consider two end-points that deliver view v′ while in view v. At the time of delivering view v′, each of these end-points has synchronization messages from all end-points in the intersection of these views (second precondition), and these synchronization messages are the same as those at their original end-points (Invariant B.9). Thus, the two end-points calculate the same transitional sets, and use the same cuts from the members of this transitional set. ■

B.2.2 Simulation

We augment VS_RFIFO+TS with a *global* history variable H_cut that keeps track of the cuts used for moving between views.

```
For each View v, v′: (Proc → Int)⊥ H_cut[v][v′], initially ⊥
OUTPUT view_p(v, T) modifies wv_rfifo.view_p(v)
pre: ...
eff: ...
     (∀ q ∈ Proc) H_cut[current_view][v](q) ← max_{r∈T} (sync_msg[r][v.startId(r)].cut(q))
```

Variable H_cut[v][v′] is updated every time *any* end-point is delivering view v′ while in view v. Corollary B.1 implies that whenever this happens after H_cut[v][v′] is set for the first time the value of H_cut[v][v′] remains unchanged.

We now extend the refinement mapping R() of Lemma B.1 with the new mapping $R_n()$:

$$\texttt{For each View v, View v}': R_n(\texttt{s.H\_cut}[v][v']) = \texttt{cut}[v][v'].$$

We call the resulting mapping R′(). We exploit the Proof Extension Theorem of [KKLS00] (see Appendix A) in order to prove that R′() is a refinement mapping from VS_RFIFO+TS to VSRFIFO : SPEC.

Lemma B.2 R′() *defined above is a refinement mapping from* VS_RFIFO+TS *to* VSRFIFO : SPEC*.*

Proof B.2:

Action Correspondence: The action correspondence is the same as that of WV_RFIFO, except for the steps of the type $(\mathtt{s}, \mathtt{view_p}(\mathtt{v}', \mathtt{T}), \mathtt{s}')$ which involve VS_RFIFO+TS delivering views to the application clients. Among these steps, those that are the first to set variable $\mathtt{H\_cut}[\mathtt{v}][\mathtt{v}']$ (when $\mathtt{s}.\mathtt{H\_cut}[\mathtt{v}][\mathtt{v}'] = \perp$) simulate two steps of VSRFIFO : SPEC: $\mathtt{set\_cut}(\mathtt{v}, \mathtt{v}', \mathtt{s}'.\mathtt{H\_cut}[\mathtt{v}][\mathtt{v}'])$ followed by $\mathtt{view_p}(\mathtt{v}')$. The rest (when $\mathtt{s}.\mathtt{H\_cut}[\mathtt{v}][\mathtt{v}'] \neq \perp$) simulate single steps that involve just $\mathtt{view_p}(\mathtt{v}')$.

Simulation Proof:

First, we show that the refinement mapping of WV_RFIFO (presented in Lemma B.1) is still preserved after the modifications introduced by VSRFIFO : SPEC to WV_RFIFO : SPEC. Automaton VSRFIFO : SPEC adds the following preconditions to $\mathtt{view_p}(\mathtt{v})$ actions of WV_RFIFO : SPEC:

```
cut[current_view[p]][v] ≠ ⊥
(∀ q)  last_dlvrd[q][p] = cut[current_view[p]][v](q)
```

The first precondition holds since action $\mathtt{set\_cut}(\mathtt{current\_view}[\mathtt{p}], \mathtt{v}, \mathtt{s}'.\mathtt{H\_cut}[\mathtt{current\_view}[\mathtt{p}]][\mathtt{v}])$ is simulated before action $\mathtt{view_p}(\mathtt{v})$. The second one follows immediately from the precondition on VS_RFIFO+TS.$\mathtt{view_p}(\mathtt{v}, \mathtt{T})$, and the extended mapping $\mathtt{R}'()$.

Second, we show that the mapping $\mathtt{R_n}()$ used to extend $\mathtt{R}()$ to $\mathtt{R}'()$ is also a refinement. For those steps $(\mathtt{s}, \mathtt{view_p}(\mathtt{v}', \mathtt{T}), \mathtt{s}')$ that are the first to set variable $\mathtt{H\_cut}[\mathtt{v}][\mathtt{v}']$, the action correspondence implies that the mapping is preserved. For those steps that are not the first to set variable $\mathtt{H\_cut}[\mathtt{v}][\mathtt{v}']$, the mapping is preserved because $\mathtt{s}'.\mathtt{H\_cut}[\mathtt{v}][\mathtt{v}'] = \mathtt{s}.\mathtt{H\_cut}[\mathtt{v}][\mathtt{v}']$, by Corollary B.1. ■

From Lemmas B.1 and B.2 and from Theorem A.1 we conclude the following:

Theorem B.2 VS_RFIFO+TS *implements* VSRFIFO : SPEC *in the sense of trace inclusion.*

B.3 Transitional Set

We now show that VS_RFIFO+TS simulates TS : SPEC. The proofs makes use of *prophecy variables*. A simulation proof that uses prophecy variables implies only finite trace inclusion, but this is sufficient for proving safety properties, (see Appendix A).

B.3.1 Invariants

Invariant B.10 *In every reachable state* s *of* VS_RFIFO+TS*, for all* Proc p *and* StartChangeId id*,*
if id > s[MBRSHP].start_change[p].id*, then* s[p].sync_msg[p][id] = ⊥.

Proof B.10: The proposition is true in the initial state. It remains true for the inductive step involving MBRSHP.$\mathtt{start\_change_p}(\mathtt{id}, \mathtt{set})$ because s[mbrshp].start_change[p].id is increased as a result of this action. For the step involving $\mathtt{set\_cut_p}()$, the proposition remains true because s[p].start_change.id = s[MBRSHP].start_change[p].id, as implied by the following invariant, which can be proved by straightforward induction:

In every reachable state s of VS_RFIFO+TS, for all Proc p, if s[p].start_change.id ≠ ⊥, then s[MBRSHP].start_change[p].id = s[p].start_change.id. This invariant holds in the initial state. Critical action MBRSHP.$\mathtt{start\_change_p}(\mathtt{id}, \mathtt{set})$ makes it true; Critical action $\mathtt{view_p}(\mathtt{v}, \mathtt{T})$ makes it vacuously true.

Finally, a step involving CO_RFIFO.$\texttt{deliver}_{\texttt{q,p}}(\texttt{tag} = \textbf{sync\_msg}, \texttt{cid}, \texttt{v}, \texttt{cut})$ does not affect the proposition because the case q=p can not happen since end-points do not send synchronization messages to themselves. ■

Lemma B.3 *For any step* $(\texttt{s}, \text{MBRSHP}.\texttt{start\_change}_{\texttt{p}}(\texttt{id}, \texttt{set}), \texttt{s}')$ *of* VS_RFIFO+TS*,*

$$\texttt{s[p].sync\_msg[p][start\_change.id]} = \bot.$$

Proof B.3: Follows immediately from the precondition $\texttt{id} > \texttt{s[}\text{MBRSHP}\texttt{].start\_change[p].id}$ and Invariant B.10. ■

Invariant B.11 *In every reachable state* s *of* VS_RFIFO+TS*, for all* Proc p*, if* $\texttt{s[p].start\_change} \neq \bot$ *and* $\texttt{s[p].sync\_msg[p][s[p].start\_change.id]} \neq \bot$*, then*

$$\texttt{s[p].sync\_msg[p][s[p].start\_change.id].view} = \texttt{s[p].current\_view}.$$

Proof B.11: The proposition is vacuously true in the initial state. For the inductive step, consider the following critical actions:

MBRSHP.$\texttt{start\_change}_{\texttt{p}}(\texttt{id}, \texttt{set})$: The proposition remains vacuously true because $\texttt{s}'[\texttt{p}].\texttt{sync\_msg}[\texttt{p}][\texttt{start\_change.id}] = \texttt{s}[\texttt{p}].\texttt{sync\_msg}[\texttt{p}][\texttt{start\_change.id}] = \bot$ (Lemma B.3).

$\texttt{set\_cut}_{\texttt{p}}()$: Follows immediately from the code.

CO_RFIFO.$\texttt{deliver}_{\texttt{q,p}}(\texttt{tag} = \textbf{sync\_msg}, \texttt{cid}, \texttt{v}, \texttt{cut})$: The proposition is unaffected because the case q=p can not happen since end-points do not send synchronization messages to themselves.

$\texttt{view}_{\texttt{p}}(\texttt{v})$: The proposition becomes vacuously true because $\texttt{s}'[\texttt{p}].\texttt{start\_change} = \bot$.

■

B.3.2 Simulation

We augment VS_RFIFO+TS with a prophecy variable P_legal_views(p)(id) for each Proc p, and each StartChangeId id. At the time a start_change id is delivered to an end-point p, this variable is set to a *predicted* finite set of future views that are allowed to contain id as p's start_change id.

```
Prophecy Variable:
For each Proc p, StartChangeId id: SetOf(View) P_legal_views(p)(id), initially arbitrary

INTERNAL mbrshp.start_change$_p$(id, set)   hidden parameter V, a finite set of views
pre: ...
     choose V such that $\forall$ v $\in$ V: (p $\in$ v.set) $\wedge$ (v.startId(p) = id)
eff: ...
     P_legal_views(p)(id) $\leftarrow$ V

OUTPUT view$_p$(v, T)
pre: ...
     ($\forall$ q $\in$ v.set) v $\in$ P_legal_views(q)(v.startId(q))
eff: ...
```

The VS_RFIFO+TS automaton augmented with the prophecy variable has the same traces as those of the original automaton because, it is straightforward to show that the following conditions required for adding a prophecy variable hold:

1. Every state has at least one value for P_legal_views(p)(id).
2. No step is disabled in the *backward direction* by new preconditions involving P_legal_views.
3. Values assigned to state variables do not depend on the values of P_legal_views.
4. If $\texttt{s}_0$ is an initial state of VS_RFIFO+TS, and $\langle \texttt{s}_0, \texttt{P\_legal\_views} \rangle$ is a state of the automaton VS_RFIFO+TS augmented with the prophecy variable, then this state is an initial state.

Invariant B.12 *In every reachable state* s *of* VS_RFIFO+TS, *for all* Proc p, *if* s[p].start_change $\neq \perp$, *then, for all* View v $\in$ P_legal_views(p)(s[p].start_change.id), *it follows that* p $\in$ v.set *and* v.startId(p) $=$ s[p].start_change.id.

Proof B.12: By induction. The only critical actions are MBRSHP.$\texttt{start\_change}_\texttt{p}(\texttt{id}, \texttt{set})$ and $\texttt{view}_\texttt{p}(\texttt{v}, \texttt{T})$. The proposition is true after the former, and is vacuously true after the latter. ■

Lemma B.4 *The following function* TS() *is a* refinement mapping *from automaton* VS_RFIFO+TS *to automaton* TS : SPEC *with respect to their reachable states.*

TS(s $\in$ ReachableStates(VS_RFIFO+TS)) $=$ t $\in$ ReachableStates(TS : SPEC), where

For each Proc p: t.current_view[p] $=$ s[p].current_view

For each Proc p, View v: t.prev_view[p][v] $=$

$$= \begin{cases} \perp & \textit{if } \texttt{v} \notin \texttt{s.P\_legal\_views[p][v.startId(p)]} \\ \texttt{s[p].sync\_msg[p][v.startId(p)].view} & \textit{otherwise} \end{cases}$$

Proof B.4:

Action Correspondence: A step (s, $\texttt{set\_cut}_\texttt{p}()$, s$'$) of VS_RFIFO+TS simulates a sequence of steps of TS : SPEC that involve one $\texttt{set\_prev\_view}_\texttt{p}(\texttt{v}')$ for each v$'$ $\in$ s.P_legal_views(p)(cid), where cid $=$ s[p].start_change.id. A step (s, $\texttt{view}_\texttt{p}(\texttt{v}, \texttt{T})$, s$'$) of VS_RFIFO+TS simulates (TS(s), $\texttt{view}_\texttt{p}(\texttt{v}, \texttt{T})$, TS(s$'$)) of TS : SPEC.

Simulation Proof: Consider the following critical actions:

MBRSHP.$\texttt{start\_change}_\texttt{p}(\texttt{id}, \texttt{set})$: A step involving this action simulates an empty step of TS : SPEC. The simulation holds because s$'$[p].sync_msg[p][id] $=$ s[p].sync_msg[p][id] $= \perp$ (Lemma B.3).

$\texttt{set\_cut}_\texttt{p}()$: simulates a sequence of steps of TS : SPEC that involve one $\texttt{set\_prev\_view}_\texttt{p}(\texttt{v}')$ for each v$'$ $\in$ s.P_legal_views(p)(cid), where cid $=$ s[p].start_change.id. Each such step is enabled as can be seen from the following derivation:

$$\begin{aligned} &\texttt{TS(s).prev\_view[p][v']} = \\ &\quad = \texttt{s[p].sync\_msg[p][v'.startId(p)].view} \text{ (Refinement mapping)} \\ &\quad = \texttt{s[p].sync\_msg[p][cid].view} \text{ (Invariant B.12)} \\ &\quad = \perp. \text{ (Precondition of } \texttt{set\_cut}_\texttt{p}()) \end{aligned}$$

In the post-state, s'[p].sync_msg[p][cid].view and all TS(s').prev_view[p][v'] are equal to s[p].current_view, thus the simulation step holds.

CO_RFIFO.$\texttt{deliver}_\texttt{q,p}(\texttt{tag} = \textbf{sync\_msg}, \texttt{cid}, \texttt{v}, \texttt{cut})$: A step involving this action does not affect any of the variables of the refinement mapping and thus simulates an empty step of TS : SPEC. In particular, note that the case of q=p may not happen because end-points do not send synchronization messages to themselves.

view$_{p}$(v, T): A step involving this action simulates a step of TS : SPEC that involves view$_{p}$(v, T). The key thing is to show that it is enabled (since it is straightforward to see that, if it is, the refinement is preserved). Action view$_{p}$(v, T) of TS : SPEC has three preconditions. The fact that they are enabled follows directly from the inductive hypothesis, the code, the refinement mapping, and Invariants B.11 and B.12. ■

From Lemma B.4 and Theorem A.1 we conclude the following:

Theorem B.3 VS_RFIFO+TS *implements* TS : SPEC *in the sense of* finite *trace inclusion.*

B.4 Self Delivery

We now prove that the complete GCS end-point automaton simulates SELF : SPEC. In order to prove this, we need to formalize our assumptions about the behavior of the clients of a GCS end-point: we assume that a client eventually responds to every block request with a block_ok response and subsequently refrains from sending messages until a view is delivered to it. We formalize this requirement by specifying an abstract client automaton in Figure B.4. In this automaton, each locally controlled action is defined to be a task by itself, which means that it eventually happens if it becomes enabled unless it is subsequently disabled by another action.

```
AUTOMATON CLIENT_p : SPEC

Signature:
 Input:   deliver_p(q, m), Proc q, AppMsg m          Output:  send_p(m), AppMsg m
          view_p(v), View v                                   block_ok_p()
          block_p()

State:    block_status ∈ {unblocked, requested, blocked}, initially unblocked

Transitions:
 INPUT  block_p()                                  OUTPUT  send_p(m)
 eff: block_status ← requested                     pre: block_status ≠ blocked
                                                   eff: none
 OUTPUT  block_ok_p()
 pre: block_status = requested                     INPUT  deliver_p(q, m)
 eff: block_status ← blocked                       eff: none

                                                   INPUT  view_p(v)
                                                   eff: block_status ← unblocked
```

Figure 16: Abstract specification of a blocking client at end-point p

B.4.1 Invariants

The following invariant states that GCS end-points and their clients have the same perception of what their block_status is.

Invariant B.13 *In every reachable state* s *of* GCS*, for all* Proc p*,*
s[GCS$_{p}$].block_status = s[client$_{p}$].block_status*.*

Proof B.13: Trivial induction. ■

Invariant B.14 *In every reachable state* $\mathtt{s}$ *of* GCS*, for all* $\mathtt{Proc\ p}$*, if* $\mathtt{s[p].start\_change} \neq \bot$ *and* $\mathtt{s[p].block\_status} \neq \mathtt{blocked}$*, then* $\mathtt{s[p].sync\_msg[p][s[p].start\_change.id]} = \bot$.

Proof B.14: The proposition is vacuously true in the initial state $\mathtt{s}_0$ because $\mathtt{s}_0[\mathrm{p}].\mathtt{start\_change} = \bot$. For the inductive step, consider the following critical actions:

$\underline{\text{MBRSHP}.\mathtt{start\_change}_\mathtt{p}(\mathtt{id}, \mathtt{set})}$: The proposition remains true because of Lemma B.3.

$\underline{\mathtt{block}_\mathtt{p}()}$: The proposition is true in the post-state if it is true in the pre-state.

$\underline{\mathtt{block\_ok}_\mathtt{p}()}$: The proposition becomes vacuously true because $\mathtt{s}'[\mathrm{p}].\mathtt{block\_status} = \mathtt{blocked}$.

$\underline{\mathtt{set\_cut}_\mathtt{p}()}$: The proposition remains vacuously true because $\mathtt{s[p].block\_status} = \mathtt{s}'[\mathrm{p}].\mathtt{block\_status} = \mathtt{blocked}$.

$\underline{\text{CO\_RFIFO}.\mathtt{deliver}_{\mathtt{q,p}}(\mathtt{tag} = \mathbf{sync\_msg}, \mathtt{cid}, \mathtt{v}, \mathtt{cut})}$: The proposition is unaffected because the case $\mathtt{q{=}p}$ can not happen since end-points do not send synchronization messages to themselves.

$\underline{\mathtt{view}_\mathtt{p}(\mathtt{v}, \mathtt{T})}$: The proposition becomes vacuously true because $\mathtt{s}'[\mathrm{p}].\mathtt{start\_change} = \bot$.

■

Invariant B.15 *In every reachable state* $\mathtt{s}$ *of* GCS*, for all* $\mathtt{Proc\ p}$*, if* $\mathtt{s[p].start\_change} \neq \bot$ *and* $\mathtt{s[p].sync\_msg[p][s[p].start\_change.id]} \neq \bot$*, then*
$\mathtt{s[p].sync\_msg[p][s[p].start\_change.id].cut[p]} =$
$=\mathtt{LastIndexOf(s[p].msgs[p][s[p].current\_view])}$.

Proof B.15: The proposition is vacuously true in the initial state $\mathtt{s}_0$ because $\mathtt{s}_0[\mathrm{p}].\mathtt{start\_change} = \bot$. For the inductive step, consider the following critical actions:

$\underline{\mathtt{send}_\mathtt{p}(\mathtt{m})}$: The proposition is vacuously true because $\mathtt{s}'[\mathrm{p}].\mathtt{sync\_msg}[\mathrm{p}][\mathtt{s}[\mathrm{p}].\mathtt{start\_change.id}] = \bot$, as follows from the precondition $\mathtt{s}[\mathtt{client}_\mathtt{p}].\mathtt{block\_status} \neq \mathtt{blocked}$ on this action at $\mathtt{client}_\mathtt{p}$, and from Invariants B.13 and B.14.

$\underline{\text{MBRSHP}.\mathtt{start\_change}_\mathtt{p}(\mathtt{id}, \mathtt{set})}$: The proposition is vacuously true because $\mathtt{s}'[\mathrm{p}].\mathtt{sync\_msg}[\mathrm{p}][\mathtt{id}] = \mathtt{s}[\mathrm{p}].\mathtt{sync\_msg}[\mathrm{p}][\mathtt{id}]$ which by Lemma B.3 is $\bot$.

$\underline{\mathtt{set\_cut}_\mathtt{p}()}$: Follows from $\mathtt{p} \in \mathtt{current\_view.set}$ (Invariant B.1) and the precondition $(\forall \mathtt{q} \in \mathtt{current\_view.set})\ \mathtt{cut(q)} = \mathtt{LongestPrefixOf(msgs[q][v])}$.

$\underline{\text{CO\_RFIFO}.\mathtt{deliver}_{\mathtt{q,p}}(\mathtt{tag} = \mathbf{sync\_msg}, \mathtt{cid}, \mathtt{v}, \mathtt{cut})}$: The proposition is unaffected because the case $\mathtt{q{=}p}$ can not happen since, as can be proved by straightforward induction, end-points do not send synchronization messages to themselves.

$\underline{\mathtt{view}_\mathtt{p}(\mathtt{v}, \mathtt{T})}$: The proposition becomes vacuously true because $\mathtt{s}'[\mathrm{p}].\mathtt{start\_change} = \bot$. ■

B.4.2 Simulation

Lemma B.2 in Section B.2 on page 43 establishes function $\mathtt{R}'()$ as a refinement mapping from automaton VS_RFIFO+TS to automaton VSRFIFO : SPEC. We now argue that $\mathtt{R}'()$ is also a refinement mapping from automaton GCS to automaton SELF : SPEC.

Lemma B.5 *Refinement mapping* $\mathtt{R}'()$ *from automaton* VS_RFIFO+TS *to automaton* VSRFIFO : SPEC *(given in Lemma B.2) is also a refinement mapping from automaton* GCS *to automaton* SELF : SPEC*, under the assumption that clients at each end-point* $\mathtt{p}$ *satisfy the* CLIENT : SPEC$_p$ *specification for blocking clients.*

Proof : Automaton SELF : SPEC modifies automaton WV_RFIFO : SPEC by adding a precondition, $\texttt{last\_dlvrd[p][p]} = \texttt{LastIndexOf(msgs[p][current\_view[p]])}$, to the steps involving $\texttt{view}_\texttt{p}()$ actions. We have to show that this precondition is enabled when a step of GCS involving $\texttt{view}_\texttt{p}(\texttt{v}, \texttt{T})$ attempts to simulate a step of SELF : SPEC involving $\texttt{view}_\texttt{p}(\texttt{v})$. Indeed:

$$\begin{aligned}
\texttt{s[p].last\_dlvrd[p]} &= \max_{\texttt{r}\in\texttt{T}}\texttt{sync\_msg}[\texttt{r}][\texttt{v.startId}(\texttt{r})].\texttt{cut}[\texttt{p}] \text{ (a precondition)}\\
&= \texttt{s[p].sync\_msg[p][v.startId(p)].cut[p]} \text{ (Invariant B.9.)}\\
&= \texttt{s[p].sync\_msg[p][s[p].start\_change.id].cut[p]} \text{ (a precondition)}\\
&= \texttt{LastIndexOf(s[p].msgs[p][s[p].current\_view])} \text{ (Invariant B.15).}
\end{aligned}$$

Thus, $\texttt{R}'(\texttt{s}).\texttt{last\_dlvrd}[\texttt{p}][\texttt{p}] = \texttt{LastIndexOf}(\texttt{R}'(\texttt{s}).\texttt{msgs}[\texttt{p}][\texttt{R}'(\texttt{s}).\texttt{current\_view}[\texttt{p}]])$ and the precondition is satisfied. ■

From Lemmas B.1, B.2, and B.5 and Theorem A.1 we conclude the following:

Theorem B.4 *GCS implements SELF : SPEC in the sense of trace inclusion, under the assumption that clients at each end-point* p *satisfy the CLIENT : SPEC$_p$ specification for blocking clients.*

As a child of VS_RFIFO+TS, GCS also satisfies all the safety property that VS_RFIFO+TS does, in particular TS : SPEC. Thus, from Theorems B.3, and B.4 we conclude the following:

Theorem B.5 *GCS implements WV_RFIFO : SPEC, VSRFIFO : SPEC, TS : SPEC, and SELF : SPEC in the sense of trace inclusion, under the assumption that clients at each end-point* p *satisfy the CLIENT : SPEC$_p$ specification for blocking clients.*

C Correctness Proof: Liveness Property

In this section we prove that fair executions of our group communication service GCS satisfy Liveness property 4.1 of Section 4.2. In order to show that a certain action eventually happens, we argue that the preconditions on this action eventually become and stay satisfied, and thus the action eventually occurs, by fairness of the execution. Subsection C.1 below presents a number of invariant that are used in the proof of Liveness property 4.1 in subsection C.2.

C.1 Invariants

The following invariant captures the fact that, before an end-point computes who the members of its transitional set are, it does not deliver to its client application messages other than those committed by its own synchronization message. Afterwards, the end-point delivers only the messages committed to delivery by the members of the transitional set.

Invariant C.1 *In every reachable state* s *of GCS, for all* Proc p, *if* $\texttt{s[p].start\_change} \neq \bot$ *and* $\texttt{s[p].sync\_msg[p][s[p].start\_change.id]} \neq \bot$, *then for all* $\texttt{Proc q} \in \texttt{s[p].current\_view.set}$,

1. *If* $\texttt{s[p].start\_change.id} \neq \texttt{s[p].mbrshp\_view.startId(p)}$, *then* $\texttt{s[p].last\_dlvrd[q]} \leq \texttt{s[p].sync\_msg[p][s[p].start\_change.id].cut[q]}$.

2. *Otherwise, let* $\texttt{v} = \texttt{s[p].current\_view}$, $\texttt{v}' = \texttt{s[p].mbrshp\_view}$, *and let* $\texttt{T} = \{\texttt{q} \in \texttt{v}'\texttt{.set} \cap \texttt{v.set} \mid \texttt{sync\_msg[q][v}'\texttt{.startId(q)].view} = \texttt{v}\}$, *then* $\texttt{s[p].last\_dlvrd[q]} \leq max_{\texttt{r}\in\texttt{T}}\ \texttt{s[p].sync\_msg[r][v}'\texttt{.startId(r)].cut[q]}$.

Proof C.1: The proposition is true in the initial state s_0, since $s_0[p].\mathtt{start\_change} = \perp$. For the inductive step, consider the following critical actions:
$\mathtt{deliver}_p(q, m)$: The proposition remains true because the precondition on this action mimics the statement of this proposition.

MBRSHP.$\mathtt{start\_change}_p(\mathtt{id}, \mathtt{set})$: The proposition is vacuously true because $s'[p].\mathtt{sync\_msg}[p][\mathtt{id}] = s[p].\mathtt{sync\_msg}[p][\mathtt{id}]$, which by Lemma B.3 is equal to $\perp$.

MBRSHP.$\mathtt{view}_p(v)$: In the post-state, $s[p].\mathtt{start\_change.id} = s[p].\mathtt{mbrshp\_view.startId}(p)$, so we must consider the second proposition. Its truth follows from the inductive hypothesis and the fact that $p \in T$, as implied by Invariant B.1.

$\mathtt{set\_cut}_p()$: The proposition holds since index $s[p].\mathtt{last\_dlvrd}[q]$ is bounded by $\mathtt{LongestPrefixOf}(s[p].\mathtt{msgs}[q][s[p].\mathtt{current\_view}])$ in every reachable state of the system for any $\mathtt{Proc}\ q \in s[p].\mathtt{current\_view.set}$ (this fact can be straightforwardly proved by induction), and from the precondition, $(\forall q \in s[p].\mathtt{current\_view.set})$
$\mathtt{cut}(q) = \mathtt{LongestPrefixOf}(s[p].\mathtt{msgs}[q][s[p].\mathtt{current\_view}])$.

CO_RFIFO.$\mathtt{deliver}_{q,p}(\mathtt{tag} = \textbf{sync\_msg}, \mathtt{cid}, v, \mathtt{cut})$: The proposition is unaffected because the case $q = p$ is impossible since end-points do not send cuts to themselves.

$\mathtt{view}_p(v, T)$: The proposition becomes vacuously true because $s'[p].\mathtt{start\_change} = \perp$.

■

The following Invariant states that if an end-point p has end-point q's cut committing certain messages sent by end-point r in view v, then end-point q has those messages buffered.

Invariant C.2 *In every reachable state* s *of* GCS*, for all* Proc p*,* Proc q*,* Proc r*, and* StartChangeId cid*, if* $s[p].\mathtt{sync\_msg}[q][\mathtt{cid}] \neq \perp$*, then, for every integer* i *between* 1 *and* $s[p].\mathtt{sync\_msg}[q][\mathtt{cid}].\mathtt{cut}[r]$*,* $s[q].\mathtt{msgs}[r][s[p].\mathtt{sync\_msg}[q][\mathtt{cid}].\mathtt{view}][i] \neq \perp$.

Proof C.2: Let us first argue that an end-point's cut commits to deliver only those messages that it has on its message queue. Formally, this means that, in every reachable state s of GCS, for all Proc q, if $s[q].\mathtt{start\_change} \neq \perp$ and $s[q].\mathtt{sync\_msg}[q][s[q].\mathtt{start\_change.id}] \neq \perp$, then, for all Proc r and all Int i such that $1 \leq i \leq s[q].\mathtt{sync\_msg}[q][s[q].\mathtt{start\_change.id}].\mathtt{cut}[r]$, $s[q].\mathtt{msgs}[r][s[q].\mathtt{current\_view}][i] \neq \perp$. This proposition can be straightforwardly proved by induction: The only interesting action is $\mathtt{set\_cut}_q()$. The truth of the proposition after this action is taken follows immediately from the precondition: $(\forall r \in s[q].\mathtt{current\_view.set})\ \mathtt{cut}(r) = \mathtt{LongestPrefixOf}(s[q].\mathtt{msgs}[r][s[q].\mathtt{current\_view}])$. Given this property, the truth of the invariant follows from Invariant B.9. ■

Invariant C.3 *In every reachable state* s *of* GCS*, for all* Proc p *and* Proc q*, if* $q \in s[p].\mathtt{sync\_set}$ *then (a)* $q \in s[p].\mathtt{start\_change.set}$ *and (b)* $q \in s[p].\mathtt{reliable\_set}$.

Proof C.3: The proposition is vacuously true in the initial state, where $s[p].\mathtt{sync\_set}$ is empty. The inductive steps for the critical actions MBRSHP.$\mathtt{start\_change}_p(\mathtt{id}, \mathtt{set})$, GCS.$\mathtt{view}_p(v, T)$, and CO_RFIFO.$\mathtt{send}_p(\mathtt{set}, \mathtt{tag} = \textbf{sync\_msg}, \mathtt{cid}, v, \mathtt{cut})$ follow immediately from their code in Figure 12. The inductive step for the action CO_RFIFO.$\mathtt{reliable\_set}_p(\mathtt{set})$ follows straightforwardly from the precondition-effect code in Figures 10 and 12. The inductive step for the critical action GCS.$\mathtt{set\_cut}_p()$ follows from the code, which sets sync_set to $\{p\}$, and from the fact that p is always in its own reliable_set and start_change.set (provided $\mathtt{start\_change} \neq \perp$), which can be straightforwardly proved by induction. ■

C.2 Liveness Proof

The following lemma states that, in any execution of GCS, every GCS.$\mathtt{view_p}$ event is preceded with the right MBRSHP.$\mathtt{view_p}$ event, which itself is preceded with the right MBRSHP.$\mathtt{start\_change_p}$ event.

Lemma C.1 *In every execution sequence α of* GCS*, the following are true:*

1. *For every* GCS.$\mathtt{view_p(v,T)}$ *event, there is a preceding* MBRSHP.$\mathtt{view_p(v)}$ *event. Moreover, neither a* MBRSHP.$\mathtt{start\_change_p}$ *nor a* MBRSHP.$\mathtt{view_p}$ *event occurs between* MBRSHP.$\mathtt{view_p(v)}$ *and* GCS.$\mathtt{view_p(v,T)}$*.*

2. *For every* MBRSHP.$\mathtt{view_p(v)}$ *event, there is a preceding* MBRSHP.$\mathtt{start\_change_p(id,set)}$ *event with* $\mathtt{id} = \mathtt{v.startId(p)}$ *and* $\mathtt{set} \supseteq \mathtt{v.set}$*, such that neither a* MBRSHP.$\mathtt{start\_change_p}$*, a* MBRSHP.$\mathtt{view_p}$*, nor a* GCS.$\mathtt{view_p}$ *event occurs in α between* MBRSHP.$\mathtt{start\_change_p(id,set)}$ *and* MBRSHP.$\mathtt{view_p(v)}$*.*

Proof C.1:

1. Assume that GCS.$\mathtt{view_p(v,T)}$ occurs in α. Two of the preconditions on GCS.$\mathtt{view_p(v,T)}$ are $\mathtt{v} = \mathtt{p.mbrshp\_view}$ and $\mathtt{v.startId(p)} = \mathtt{p.start\_change.id}$, which can only become satisfied as a result of a preceding MBRSHP.$\mathtt{view_p(v)}$ event, followed by no MBRSHP.$\mathtt{start\_change_p}$ and MBRSHP.$\mathtt{view_p}$ events.

2. Assume that MBRSHP.$\mathtt{view_p(v)}$ occurs in α. Then a MBRSHP.$\mathtt{start\_change_p(id,set)}$ event with $\mathtt{id} = \mathtt{v.startId(p)}$ and $\mathtt{set} \supseteq \mathtt{v.set}$ must precede MBRSHP.$\mathtt{view_p(v)}$ because, by the MBRSHP specification, it is the only possible event that can cause the preconditions for MBRSHP.$\mathtt{view_p(v)}$ to become true, and since these preconditions do not hold in the initial state of MBRSHP. There maybe several MBRSHP.$\mathtt{start\_change_p(id,set)}$ events with the same $\mathtt{id}$ and different $\mathtt{set}$ arguments. After the last such event, an occurrence of a different MBRSHP.$\mathtt{start\_change_p}$ event or a MBRSHP.$\mathtt{view_p}$ event would violate one of the preconditions of MBRSHP.$\mathtt{view_p(v)}$; thus, such events may not happen. As a corollary from this and part 1 of this Lemma, a GCS.$\mathtt{view_p(v',T')}$ event cannot occur between the last MBRSHP.$\mathtt{start\_change_p(id,set)}$ and MBRSHP.$\mathtt{view_p(v)}$.

■

Lemma C.2 (Liveness) *Let* $\mathtt{v}$ *be a view. Let α be a fair execution of a group communication service* GCS *in which, for every* $\mathtt{p} \in \mathtt{v.set}$*, the action* MBRSHP.$\mathtt{view_p(v)}$ *occurs and is followed by neither* MBRSHP.$\mathtt{view_p}$ *nor* MBRSHP.$\mathtt{start\_change_p}$ *actions. Then at each end-point* $\mathtt{p} \in \mathtt{v.set}$*,* GCS.$\mathtt{view_p(v,T)}$*, with some* $\mathtt{T}$*, eventually occurs. Furthermore, for every* GCS.$\mathtt{send_p(m)}$ *that occurs after* GCS.$\mathtt{view_p(v,T)}$ *and for every* $\mathtt{q} \in \mathtt{v.set}$*,* GCS.$\mathtt{deliver_q(p,m)}$ *also occurs.*

Proof C.2:

Part I We first prove that GCS.$\mathtt{view_p(v,T)}$ eventually occurs. Our task is to show that, for each $\mathtt{p} \in \mathtt{v.set}$ and some transitional set $\mathtt{T}$, action GCS.$\mathtt{view_p(v,T)}$ becomes enabled at some point after $\mathtt{p}$ receives MBRSHP.$\mathtt{view_p(v)}$ and that it stays enabled forever thereafter unless it is executed. The fact that α is a fair execution of GCS then implies that GCS.$\mathtt{view_p(v,T)}$ is in fact executed.

In order for GCS.$\mathtt{view_p(v,T)}$ to become enabled, its preconditions (see Figures 10 and 12) must eventually become and stay satisfied until GCS.$\mathtt{view_p(v,T)}$ is executed. We now consider each of these preconditions:

$\mathtt{v} = \mathtt{p.mbrshp\_view} \neq \mathtt{current\_view}$: This precondition ensures that view v that is attempted to be delivered to the client at p is the latest view produced by MBRSHP and has not yet been delivered to the client. The precondition becomes satisfied as a result of MBRSHP.$\mathtt{view}_\mathtt{p}(\mathtt{v})$. Since in any reachable state of the system MBRSHP.$\mathtt{mbrshp\_view} = \mathtt{p.mbrshp\_view} \geq \mathtt{p.current\_view}$ (Local Monotonicity), this precondition remains satisfied forever, unless GCS.$\mathtt{view}_\mathtt{p}(\mathtt{v}, \mathtt{T})$ is executed. This is because, by our assumption, α does not contain any subsequent MBRSHP.$\mathtt{view}_\mathtt{p}(\mathtt{v}')$, and hence, by contrapositive of part 1 of Lemma C.1, it also does not contain any subsequent GCS.$\mathtt{view}_\mathtt{p}(\mathtt{v}', \mathtt{T}')$ with $\mathtt{v}' \neq \mathtt{v}$.

$\mathtt{v.startId(p)} = \mathtt{p.start\_change.id}$: This precondition prevents delivery of obsolete views: it ensures that the MBRSHP service has not issued a new start_change notification since the time it produced view v. If this condition is not already satisfied before the last MBRSHP.$\mathtt{start\_change}_\mathtt{p}(\mathtt{id}, \mathtt{set})$ event with $\mathtt{id} = \mathtt{v.startId(p)}$ and $\mathtt{set} \supseteq \mathtt{v.set}$, then it becomes satisfied as a result of this event, which, by part 2 of Lemma C.1, must precede MBRSHP.$\mathtt{view}_\mathtt{p}(\mathtt{v})$ in α.

This condition stays satisfied from the time of the last MBRSHP.$\mathtt{start\_change}_\mathtt{p}(\mathtt{id}, \mathtt{set})$ at least until GCS.$\mathtt{view}_\mathtt{p}(\mathtt{v}, \mathtt{T})$ occurs because the only two types of actions, MBRSHP.$\mathtt{start\_change}_\mathtt{p}(\mathtt{id}', \mathtt{set}')$ and GCS.$\mathtt{view}_\mathtt{p}(\mathtt{v}', \mathtt{T}')$ with $\mathtt{v}' \neq \mathtt{v}$ that may affect the value of p.start_change cannot occur in α after MBRSHP.$\mathtt{start\_change}_\mathtt{p}(\mathtt{id}, \mathtt{set})$, as implied by the assumption on this lemma and Lemma C.1.

$\mathtt{v.set} - \mathtt{sync\_set} = \{\}$: This precondition ensures that prior to delivering view v, end-point p sends out its synchronization message to every member of v.

If this precondition is satisfied any time after the last MBRSHP.$\mathtt{start\_change}_\mathtt{p}(\mathtt{id}, \mathtt{set})$ event with $\mathtt{id} = \mathtt{v.startId(p)}$ and $\mathtt{set} \supseteq \mathtt{v.set}$ occurs, then it stays satisfied from then on until GCS.$\mathtt{view}_\mathtt{p}(\mathtt{v}, \mathtt{T})$ is executed. If it is not already satisfied right after the MBRSHP.$\mathtt{start\_change}_\mathtt{p}$ action, it becomes satisfied as a result of CO_RFIFO.$\mathtt{send}_\mathtt{p}(\mathtt{set}, \mathtt{tag} = \mathbf{sync\_msg}, \mathtt{v.startId(p)}, \mathtt{v}, \mathtt{cut})$ with $\mathtt{set} = \mathtt{p.start\_change.set} - \mathtt{p.sync\_set}$. This CO_RFIFO.$\mathtt{send}_\mathtt{p}$ action must eventually occurs in α because its two preconditions, ($\mathtt{p.sync\_msg[p][id]} \neq \bot$) and ($\mathtt{set} \subseteq \mathtt{reliable\_set}$), eventually become satisfied, for the following reasons:

- If the first precondition is satisfied any time after the last MBRSHP.$\mathtt{start\_change}_\mathtt{p}(\mathtt{id}, \mathtt{set})$ event with $\mathtt{id} = \mathtt{v.startId(p)}$ and $\mathtt{set} \supseteq \mathtt{v.set}$ occurs, then it stays satisfied from that point on. If it is not already satisfied right after the MBRSHP.$\mathtt{start\_change}_\mathtt{p}$ action, it becomes satisfied as a result of $\mathtt{set\_cut}_\mathtt{p}()$. In order for $\mathtt{set\_cut}_\mathtt{p}()$ to occur, its precondition, $\mathtt{block\_status} = \mathtt{blocked}$, has to becomes satisfied (see Figure 13). This occurs as a result of a $\mathtt{block\_ok}_\mathtt{q}()$ input from the client at q. If block_status equals blocked at anytime after MBRSHP.$\mathtt{start\_change}_\mathtt{q}(\mathtt{v.startId(q)}, \mathtt{set})$, then it remains such until GCS.$\mathtt{view}_\mathtt{q}(\mathtt{v})$ happens because $\mathtt{block}_\mathtt{q}()$ is not enabled after that, and because GCS.$\mathtt{view}_\mathtt{q}(\mathtt{v})$ is the only possible GCS view event (by the contrapositive of part 2 of Lemma C.1). To see that block_status does in fact become blocked consider the three possible values of block_status right after MBRSHP.$\mathtt{start\_change}_\mathtt{q}(\mathtt{v.startId(q)}, \mathtt{set})$ occurs:

  1. $\mathtt{block\_status} = \mathtt{blocked}$: We are done.
  2. $\mathtt{block\_status} = \mathtt{requested}$: By Invariant B.13, $\mathtt{client.block\_ok}_\mathtt{q}()$ is enabled. It stays enabled until it is executed because the actions, $\mathtt{block}_\mathtt{q}()$ and GCS.$\mathtt{view}_\mathtt{q}()$, which would disable it, cannot occur. When it is executed, the precondition becomes satisfied.
  3. $\mathtt{block\_status} = \mathtt{unblocked}$: When MBRSHP.$\mathtt{start\_change}_\mathtt{q}(\mathtt{v.startId(q)}, \mathtt{set})$ occurs, $\mathtt{block}_\mathtt{q}()$ becomes and stays enabled until it is executed. After that, block_status becomes requested and the same reasoning as in the previous case applies.

- The second precondition, $\mathtt{set} \subseteq \mathtt{reliable\_set}$, holds after $\textsc{co\_rfifo}.\mathtt{reliable}_\mathtt{q}(\mathtt{set})$ with $\mathtt{set} = \mathtt{current\_view.set} \cup \mathtt{start\_change.set}$ occurs. This action eventually occurs because it becomes enabled when q receives $\textsc{mbrshp}.\mathtt{start\_change}_\mathtt{q}(\mathtt{v.startId(q)}, \mathtt{set})$. Note that although $\textsc{co\_rfifo}.\mathtt{reliable}_\mathtt{q}(\mathtt{set})$ may subsequently occur multiple times, reliable_set remains unchanged until $\textsc{gcs}.\mathtt{view}_\mathtt{q}(\mathtt{v})$ occurs, since q's current_view and start_change remain unchanged.

  When $\textsc{co\_rfifo}.\mathtt{send}_\mathtt{p}(\mathtt{set}, \mathtt{tag} = \mathbf{sync\_msg}, \mathtt{v.startId(p)}, \mathtt{v}, \mathtt{cut})$ occurs, p.sync_set is set to p.start_change.set. Since v.set is a subset of p.start_change.set, this implies that v.set $-$ p.sync_set eventually becomes and stays $\{\}$.

$\underline{(\forall \mathtt{q} \in \mathtt{v.set} \cap \mathtt{p.current\_view.set})\ \mathtt{p.sync\_msg[q][v.startId(q)]} \neq \perp}$: This precondition ensures that p has received the right synchronization message from every q in $\mathtt{v.set} \cap \mathtt{p.current\_view.set}$. The argument above implies that q eventually sends to p a synchronization message tagged with v.startId(q) and, at the same time, adds p to q.sync_set, where p remains forever, unless $\textsc{gcs}.\mathtt{view}_\mathtt{p}(\mathtt{v}, \mathtt{T})$ with some T occurs. In order to conclude that CO_RFIFO eventually delivers this synchronization message to p, we argue that, from the time the last synchronization message from q to p is placed on CO_RFIFO.channel[q][p] and at least until it is delivered to p, end-point p is in both CO_RFIFO.reliable_set[q] and CO_RFIFO.live_set[q]. The former implies that CO_RFIFO does not lose any messages (in particular, this synchronization message) from q to p. In conjunction with α being a fair execution, the latter implies that CO_RFIFO eventually delivers every message (in particular, this synchronization message) on the channel from q to p.

- From the time q sends to p the last synchronization message tagged with v.startId(q) until $\textsc{gcs}.\mathtt{view}_\mathtt{q}(\mathtt{v}, \mathtt{T})$ occurs, p is included in q.sync_set. Invariant C.3 implies that in that period p is included in CO_RFIFO.reliable_set[q]. After $\textsc{gcs}.\mathtt{view}_\mathtt{q}(\mathtt{v}, \mathtt{T})$ occurs, p is still included in CO_RFIFO.reliable_set[q], since $\mathtt{p} \in \mathtt{v.set}$.
- End-point p becomes a member of CO_RFIFO.live_set[q] at the time of $\textsc{mbrshp}.\mathtt{view}_\mathtt{q}(\mathtt{v})$, because $\textsc{mbrshp}.\mathtt{view}_\mathtt{q}(\mathtt{v})$ is linked to $\textsc{co\_rfifo}.\mathtt{live\_set}_\mathtt{q}(\mathtt{v.set})$ and because $\mathtt{p} \in \mathtt{v.set}$. This property remains true afterward because α does not contain any subsequent MBRSHP events at end-point q.

Thus, end-point p eventually receives the right synchronization messages from every q in
$\mathtt{v.set} \cap \mathtt{p.current\_view.set}$.

$\underline{\mathtt{last\_sent} \geq\ \mathtt{sync\_msg[p][v.startId(p)].cut(p)}}$: This precondition ensures that before delivering view v, p sends to others all of its own messages indicated in its own cut. This precondition eventually becomes satisfied because sending of of application messages via $\textsc{co\_rfifo}.\mathtt{send}_\mathtt{p}$, which increments p.last_sent, is enabled at least until p.last_sent reaches sync_msg[p][v.startId(p)].cut(p), as implied by Invariant C.2.

$\underline{(\forall \mathtt{q} \in \mathtt{current\_view.set})\ \mathtt{p.last\_dlvrd[q]} = \mathtt{max}_{\mathtt{r} \in \mathtt{T}}\mathtt{p.sync\_msg[r][v.startId(r)].cut[q]}}$: This precondition verifies that p has delivered to its client exactly the application messages that it needs to deliver in order for Virtually-Synchronous Delivery to be satisfied. By Invariant C.1, p.last_dlvrd[q] never exceeds $\mathtt{max}_{\mathtt{r} \in \mathtt{T}}\ \{\mathtt{p.sync\_msg[r][v.startId(r)].cut[q]}\}$ for any q. It is therefore left to show that p.last_dlvrd[q] does not remain smaller than $\mathtt{max}_{\mathtt{r} \in \mathtt{T}}$.

We have shown above that all the other preconditions for delivering view v by p eventually become and remain satisfied until the view is delivered. Consider the part of α after all of these preconditions hold. Let q be an end-point in current_view.set such that $\mathtt{p.last\_dlvrd[q]} < \mathtt{max}_{\mathtt{r} \in \mathtt{T}}\mathtt{p.sync\_msg[r][v.startId(r)].cut[q]}$. Let $\mathtt{i} = \mathtt{p.last\_dlvrd[q]} + 1$. We now argue that p.last_dlvrd[q] eventually becomes i, that is, that p eventually delivers the next message from q. Applying this argument inductively, implies that p.last_dlvrd[q] eventually reaches $\mathtt{max}_{\mathtt{r} \in \mathtt{T}}$ {p.sync_msg[r][v.startId(r)].cut[q]}.

All the preconditions (except perhaps p.msgs[q][p.current_view][i] $\neq \perp$) for delivering the i'th message from q are eventually satisfied because they are the same as the preconditions for p delivering view v, which we have shown to be satisfied. Thus, if the i'th message is already on p.msgs[q][p.current_view][i], then delivery of this message eventually occurs by fairness, resulting in p.last_dlvrd[q] being incremented; in this case, we are done.

Therefore, consider the case when p lacks the i'th message, m, from q. There are two possibilities:

1. If end-point q is in p's transitional set T for view v, then we know the following:
   - q's view prior to installing view v is the same as p's current view (by definition of T and Invariant B.11).
   - q's reliable_set contains p starting before q sent any messages in that view and continuing for the rest of α.
   - Invariant C.2 implies that q has this message and all the messages that precede it in q.msgs[q][p.current_view].
   - End-point q is enabled to send these messages to p in FIFO order. The only event that could prevent q from sending these messages is gcs.view$_{\mathtt{q}}$(v), as it would change the value of q.current_view. However, as we argued above, q must send all of the messages it committed in its cut before delivering view gcs.view$_{\mathtt{q}}$(v). Self Delivery (Invariant B.15) implies that q's cut includes all of the messages q sent while in v. Thus, q would eventually send m to p.
   - The fact that the connection between q and p is live at least after MBRSHP.view$_{\mathtt{q}}$(v) occurs implies that CO_RFIFO eventually delivers this message to p.
2. Otherwise, if end-point q is not in p's transitional set T for view v, we know by the fact that i is $\leq$ $\mathtt{max}_{\mathtt{r} \in \mathtt{T}}$ {p.sync_msg[r][v.startId(r)].cut[q]}, that there exist some end-points in T whose synchronization messages commit to deliver the i'th message from q in view p.current_view. Let r be an end-point with a smallest identifier among these end-points. Here is what we know:
   - Invariant C.2 implies that r has this message on its r.msgs[r][p.current_view] queue.
   - r's reliable_set contains p starting before r sent any messages in that view and continuing for the rest of α.
   - Upon examination of each of the ForwardingStrategyPredicates in Section 5.2.1, we see that the preconditions for r forwarding the i'th message of q to a set including p eventually become and stay satisfied.
   - Since in both forwarding strategies there is only a finite number of messages from q sent in this view that can be forwarded, fairness implies that the i's message is eventually forwarded to p.
   - The fact that the connection between r and p is live at least after MBRSHP.view$_{\mathtt{q}}$(v) occurs implies that CO_RFIFO eventually delivers this message to p.

Therefore, the i'th message from q is eventually delivered to end-point p, and since, as a result of this, the preconditions on delivering this message to the client at p are satisfied, this delivery eventually occurs, and p.last_dlvrd[q] is incremented. Applying this argument inductively, we conclude that p.last_dlvrd[q] eventually reaches $\max_{r \in T}$ p.sync_msg[r][v.startId(r)].cut[q] for every q in current_view.set.

We have shown that each precondition on p delivering GCS.view$_p$(v, T) eventually becomes and stays satisfied. Fairness implies that GCS.view$_p$(v, T) eventually occurs.

Part II We now consider the second part of the lemma. The following argument proves that, after GCS.view$_p$(v, T) occurs at p, for every subsequent GCS.send$_p$(m) event at p, there is a corresponding GCS.deliver$_q$(p, m) event that occurs at every $q \in$ v.set:

1. For the rest of α, after GCS.view$_p$(v, T) occurs, CO_RFIFO.live_set[p] is equal to v.set.

   This is true because CO_RFIFO.live_set[p] is set to v.set when MBRSHP.view$_p$(v) occurs and remains unchanged thereafter because of the assumption that α does not contain any subsequent MBRSHP events at end-point p.

2. After GCS.view$_p$(v, T) occurs and before any CO_RFIFO.send$_p$ event involving a ViewMsg or an AppMsg occurs, p eventually executes CO_RFIFO.reliable$_p$(v.set). Moreover, after that and forever thereafter, both p.reliable_set and CO_RFIFO.reliable_set[p] equal v.set.

   This is true because GCS.view$_p$(v, T) sets p.start_change to $\perp$ and p.current_view.set to v.set, thus enabling CO_RFIFO.reliable$_p$(v.set). This action eventually happens because α is a fair execution and because for the rest of α there are no subsequent MBRSHP.start_change$_p$ and GCS.view$_p$(v$'$, T$'$) events. Moreover, since p.start_change and p.current_view.set remain unchanged because of the latter reason, whenever CO_RFIFO.reliable$_p$ occurs subsequently, both p.reliable_set and CO_RFIFO.reliable_set[p] remain equal to v.set.

   From the above argument and from fairness, it follows that any kind of message that end-point p sends subsequently to q via CO_RFIFO will eventually reach end-point q.

3. After CO_RFIFO.reliable$_p$(v.set) occurs, CO_RFIFO.send$_p$(v.set $- \{p\}$, tag $=$ **view_msg**, v) eventually occurs, as follows from the code in Figure 12. By the reasoning above, CO_RFIFO delivers this ViewMsg to every end-point $q \in$ v.set $- \{p\}$, resulting in q.view_msg[p] being set to v for the remainder of α (Invariant B.4).

4. When GCS.send$_p$(m) event occurs at p, m is appended to p.msgs[p][v].

5. After sending the ViewMsg, for the rest of α, if p.msgs[p][v][p.last_sent $+ 1$] contains a message (say m$'$), action CO_RFIFO.send$_p$(v.set $- \{p\}$, tag $=$ **app_msg**, m$'$) is enabled, and hence eventually occurs by fairness. Since p.last_sent is incremented after each application message is sent using CO_RFIFO.send$_p$, any message on p.msgs[p][v] is eventually sent to v.set $- \{p\}$. As was argued above, these messages are eventually delivered to every end-point $q \in$ v.set $- \{p\}$. Since q.view_msg[p] $=$ v at the time q receives m$'$, q puts m$'$ in q.msgs[p][v][q.last_rcvd $+ 1$] (Invariant B.5) and increments q.last_rcvd. Therefore, all messages that end-point p sends in view v are eventually inserted with no gaps in the end-point q's queue, q.msgs[p][v], for every $q \in$ v.set $- \{p\}$.

6. Once GCS.view$_q$(v, T) happens (by Part I of the lemma), end-point $q \in$ v.set is continuously enabled to deliver a message, m$'$, from q.msgs[p][v][q.last_dlvrd $+ 1$]; by fairness, such delivery eventually occurs, resulting in q.last_dlvrd[p] being incremented. Therefore, every messages on q.msgs[p][v] is eventually delivered to client at p, including the case of q $=$ p.

It follows from this argument that every GCS.send$_p$(m) event at p that occurs after GCS.view$_p$(v, T) in α is eventually followed by a GCS.deliver$_q$(p, m) at every $q \in$ v.set. ■