# System Support for Partition-Aware Network Applications

Özalp Babaoğlu Renzo Davoli Alberto Montresor Roberto Segala

Department of Computer Science
University of Bologna
40127 Bologna, Italy

## Abstract

*Network applications and services need to be* environment-aware *in order to meet non-functional requirements in increasingly dynamic contexts. In this paper we consider* partition awareness *as an instance of environment awareness in network applications that need to be reliable and self-managing. Partition-aware applications dynamically reconfigure themselves and adjust the quality of their services in response to partitioning and merging of networks. As such, they can automatically adapt to changes in the environment so as to remain available in multiple partitions without blocking, albeit with reduced or degraded functionality. We propose a system layer consisting of group membership and reliable multicast services that provides systematic support for partition-aware application development. We illustrate the effectiveness of the proposed interface by solving several problems that represent different classes of realistic network applications.*

## 1. Introduction

Functional requirements, which define how output values are related to input values, are usually sufficient for specifying traditional applications. For modern network applications, however, non-functional requirements can be just as important as their functional counterparts: the services that these applications provide must not only be correct with respect to the functional requirements, they must also be delivered with acceptable "quality" levels. Reliability, timeliness and configurability are examples of non-functional requirements that are of particular interest to network applications.

A correct application satisfies its functional requirements in all possible operating environments: it just may take more or less time to do so depending on the characteristics of the environment. On the other hand, there may be operating environments in which it is impossible to achieve non-functional properties beyond certain levels. For this reason, non-functional requirements of network applications define quality *intervals* rather than absolute values that are considered acceptable. In order to deliver quality levels that are both feasible and acceptable, network applications need to be aware of the environment in which they are operating and use this information to dynamically modify their behavior. We call this capability of an application *environment awareness*.

By their nature, network applications for mobile computing, data sharing or collaborative work involve cooperation among multiple sites. For these applications, which are characterized by reliability and configurability requirements, partitionings of the communication network into several partitions is an extremely important aspect of the environment. The nature of a partitioning will determine the quality for the application in terms of which services are available where, and at what performance levels. In other words, partitionings may result in service *reduction* or service *degradation* but need not necessarily render application services completely unavailable. Informally, we define the class of *partition-aware* applications as those that are able to make progress in multiple concurrent partitions without blocking.

Service reduction and degradation that are unavoidable during partitionings depend heavily on the application semantics and establishing them is beyond the scope of this paper. For certain application classes with strong consistency requirements, it may be the case that all services have to be suspended completely in all but one partition. This situation corresponds to the so-called *primary-partition* model [**?**, **?**] that has traditionally characterized partitioned operation of network applications. In this paper we focus on *system services* that support partition awareness such that continued operation of network applications is not restricted to a single partition but may span multiple concurrent partitions. The system provides the necessary abstractions such that the application itself can decide which of its services will be available in each partition and at what quality levels.

Our methodology is based on the *process group* paradigm [**?**, **?**] suitably extended to partitionable systems. Members of a group cooperate in order to implement a given network application. The methodology unifies all relevant aspects of environment awareness (process crashes/recoveries, network partitionings/merges) in a single abstraction: group membership. Partition-aware applications are programmed so as to reconfigure themselves and adjust their behavior using the current membership of the group as input. The perception that each process has of the group's membership needs to be constructed with care, since otherwise, inconsistencies may compromise functional requirements or result in quality levels that are lower than what is feasible. In Section **??** we specify a *partitionable group membership service* (PGMS) that forms the basis of our support layer. As illustrated in Section **??**, PGMS alone may be sufficient for programming a small class of self-configuring network services. To support a broader class of applications that require closer cooperation, in Section **??** we augment PGMS with a reliable multicast communication service satisfying *view synchrony* semantics.

The contributions of this paper are to argue that partition awareness is an important attribute of future network applications and to show that the view synchrony service we propose is indeed useful for supporting it. We do so by developing partition-aware solutions to several abstract problems. Each problem is representative of a particular class of realistic applications. The advocated methodology not only results in simple solutions to non-trivial applications, it also admits simple proofs of their correctness.

## 2. Related Work

Numerous systems have been proposed or built recently as infrastructures for supporting applications in partitionable environments. Most of these systems are based on the process group paradigm and include Horus [**?**], Transis [**?**], Totem [**?**], Relacs [**?**], and NavTech [**?**] as some of the more notable examples. There have also been several attempts at developing realistic applications based on the services provided by these systems: Transis used to implement services that guarantee global ordering of messages assuming that a majority of the processes are connected and alive [**?**, **?**]; Transis used to implement simple system management services [**?**]; Horus used to implement a replicated state machine that replies to all requests in the majority partition but replies only to a restricted class of requests in minority partitions [**?**]; NavTech used to manage fault tolerance in groupware applications [**?**]. In most cases, these applications are based on the primary-partition model that is created on top of a partitionable service. As such, progress in these applications is limited to a single partition. The only exception to this observation is the groupware application described in [**?**]. The work reported in this paper differs from the above proposals in several important ways. First, we consider systematic support for partition awareness in a wide range of application areas rather than one specific area. Second, the applications we consider take full advantage of partition awareness and remain available in all partitions rather than only a majority-based primary partition.

## 3. System Model

The system comprises a set of processes that communicate by exchanging messages through a network. The system is asynchronous in the sense that neither communication delays nor relative process speeds can be bounded. Practical distributed systems often have to be considered as being asynchronous since transient failures, unknown scheduling strategies and variable loads on the computing and communication resources make it impossible to bound delays.

In the absence of failures, the network is logically connected and each process can communicate with every other process. Processes may fail by *crashing* whereby they halt prematurely. Crashed processes may *recover* and resume their execution after repairs. Communication failures may provoke a *partitioning* of the network whereby processes are split across two or more *partitions*.[1] A partitioning is described by a set of partitions and characterizes the global communication state of the system. Each partition, on the other hand, is described by a set of processes with the property that processes within the same partition are able to communicate while those across different partitions cannot communicate. Upon repairs, multiple partitions may *merge* to form a single larger partition.

Intuitively, partitions correspond to maximally connected components of a logical graph representing the "reachable" relation among processes. As such, they can be defined only in the context of specific communication primitives. For example, two processes may appear to belong to two different partitions with respect to "ping" messages, but the same two processes may appear in the same partition when communicating through email. This is because the two communication services being considered have significantly different message buffering, timeout and retransmission properties. The support layer we propose is built on top of the underlying communication primitives, whatever they may be, and exports appropriate abstractions for contructing partition-aware applications.

[1]In addition to communication failures, other situations including incorrect or inconsistent routing tables may lead to partitionings. And if the network is not fully connected, process crashes may also result in partitionings.

## 4. Support for Partition-Aware Applications

Our methodology is based on the *process group* paradigm with suitable extensions to systems that admit partitioning. In this methodology, processes initiate their collaboration towards a given application by *joining* as members a named group. Later on, a process may decide to terminate its collaboration by explicitly *leaving* the group. In the absence of failures, the *membership* of a group comprises those processes that have joined but have not left the group. In addition to explicit joins and leaves, a group's membership will vary also due to failures and repairs. A *partitionable group membership service* (PGMS), specified in the next section, tracks all changes in the group's membership and installs them as *views* at processes through $vchg()$ upcalls. Installed views are abstractions of the environment with respect to process crashes/recoveries and network partitionings/merges. Ideally, there should be a common view of the group's membership that is shared by all of its members and this view should include exactly those members that are currently operational. This is clearly not feasible in a partitionable system where processes in different partitions will have different perceptions of the membership for a given group. Nevertheless, the PGMS we specify is useful and guarantees that each installed view is shared by all of its components and corresponds to an actual partition of a partitioning.

Group members communicate through reliable multicasts by invoking the primitive $mcast()$. Multicast messages are delivered to processes through $dlvr()$ upcalls. The relationship between multicast message deliveries and the sequence of installed views is formalized as *view synchrony* in Section **??**. Views and multicast messages are labeled in order to be globally unique. Given a view $v$, we write $\overline{v}$ to denote its composition as a set of process names. The *current view* of process $p$ at time $t$ is the last view to have been installed at $p$ before time $t$. Events are said to occur *in the view* that is current. View $w$ is called the *immediate successor of* $v$ *at* $p$ if $p$ installs $w$ in view $v$. View $w$ is called an *immediate successor* of $v$ if $w$ happens to be the immediate successor of $v$ at some process $p$. The transitive closure of the immediate successor relation is called the *successor* relation. Two views that are unrelated through successor relation are called *concurrent*.

### 4.1. Partitionable Group Membership Service

The first layer of our support architecture governs view installations and view compositions related to group membership. Installed views represent the perception of the group's membership that is shared by all processes in the view's composition. In other words, there is agreement among processes on the composition of the view that they will install. Furthermore, processes that install a given view are exactly those within a given partition. We have formalized these ideas leading to a specification for partitionable group membership in asynchronous systems [**?**]. For sake of brevity, we state our specification informally, omitting the discussions that motivate it. For the same reason, the specification considers group membership changes due to failures and repairs, omitting those due to joins and leaves.

**GM1 (View Agreement)** *(i) If process $p$ installs view $v$ and its immediate successor $w$, both containing $q$, then $p$ installs view $w$ only after $q$ has installed $v$. (ii) Suppose correct process $p$ installs view $v$ containing some process $q$. If the current view of $q$ after some time becomes permanently distinct from $v$, then $p$ will eventually install a new view as an immediate successor to $v$.*

**GM2 (View Accuracy)** *If there is a time after which process $q$ remains reachable from some correct process $p$, then eventually the current view of $p$ will always include $q$.*

**GM3 (View Completeness)** *If there is a time after which all processes in some partition $\Theta$ cannot communicate with the rest of the group, then eventually the current view of every correct process not in $\Theta$ will never include any process in $\Theta$.*

**GM4 (View Integrity)** *Every view installed by a process includes the process itself.*

**GM5 (View Order)** *The order in which processes install views is such that the successor relation is a partial order.*

Properties GM**??**–GM**??** together define a *partitionable group membership service* in asynchronous systems.

### 4.2. View Synchrony Service

The class of partition-aware applications that can be programmed using PGMS alone can be characterized as configuration management. In general, network applications require closer cooperation that is facilitated through communication among group members. In this Section we extend the group membership service of the previous Section with a reliable multicast primitive. The resulting service is called *view synchrony* and integrates delivery of multicast messages with the installation of views. Informally, all processes that survive together from one view to another deliver the same set of messages. In other words, there is agreement among the processes on the set of messages that they should deliver in a view before they can install a common next view.

**VS1 (Message Agreement)** *Given two views $v$ and $w$ such that $w$ is an immediate successor of $v$, all processes belonging to both views deliver the same set of multicast messages in view $v$.*

**VS2 (Uniqueness)** *Each multicast message, if delivered at all, is delivered in exactly one view.*

**VS3 (Merging Rule)** *Two views that merge to a single view must have disjoint compositions.* [2]

**VS4 (Message Integrity)** *Each process delivers a message at most once and only if some process actually multicast it earlier.*

**VS5 (Local FIFO Order)** *If a process $p$ multicasts message $m$ before it multicasts message $m'$, then $p$ does not deliver $m'$ unless it has delivered $m$.*

**VS6 (Liveness)** *(i) A correct process always delivers its own multicast messages. (ii) Let $p$ be a correct process that delivers message $m$ in view $v$ that includes some other process $q$. If $q$ never delivers $m$, then $p$ will eventually install view $w$ that excludes $q$ as immediate successor to $v$.*

**VS7 (Immediate Local Delivery)** *Messages multicast by a correct process $p$ while handling the upcall $vchg(v)$ are delivered by $p$ in view $v$.*

Properties GM**??**–GM**??** together with VS**??**–VS**??** define a *view synchrony service* in asynchronous systems. Note that our specification of view synchrony is similar to *extended virtual synchrony* (EVS) [**?**], although there are some important differences. EVS distinguishes between two types of view changes: transitional configurations and regular ones. Transitional configurations are used to handle scenarios in which configurations corresponding to merging partitions have non empty intersections. Our specification has a single, unified notion of a view and ensures that no two merging views have overlapping compositions through Merging Rule. Furthermore, our specification excludes trivial solutions by including non-triviality properties that are not considered in EVS.

## 5. Programming Partition-Aware Applications

In this section, we first illustrate how group membership alone can be sufficient to program simple network applications that require no communication. We then consider more realistic applications that require cooperation through communication and show how view synchrony can be used to program them. The problems are stated abstractly in order to hide irrelevant details, and can be instantiated as numerous realistic partition-aware applications. The solutions we present deliberately avoid optimizations that are often possible but would only complicate their presentations. A more detailed description of the algorithms along with their proofs of correctness can be found in the extended version of this work [**?**].

[2] At first sight, this property might appear relevant for specifying PGMS rather than view synchrony since it concerns only view compositions. Yet, its consequences are relevant only if there is communication among the group members [**?**].

```
procedure $PartitionableServiceActivator()$
  $server\_at \leftarrow p$
  $ActivateService()$

  while true do
    wait-for $vchg(v)$
    if $server\_at = p$ and $Min(\overline{v}) \neq p$ then
      $DeactivateService()$
    if $server\_at \neq p$ and $Min(\overline{v}) = p$ then
      $ActivateService()$
    $server\_at \leftarrow Min(\overline{v})$
  od
```

Figure 1. Partitionable Service Activator.

### 5.1. Partitionable Service Activator Application

We describe a simple network application for which PGMS alone is sufficient as the support layer. Consider a network service for diffusing a continuous stream of data (e.g., audio, video, stock quotes, news headlines) to a collection of subscribers. The diffusion can be provided by any one of a set of *servers* that have access to the data source. The service should be available in every partition that contains at least one server; furthermore, to minimize wasted resources, multiple active servers within the same partition should be avoided. New servers may be added and existing ones removed at will by an administrator. The goal is to devise a *service activator* algorithm such that a server can decide when it should be active and when it should be passive. A solution must activate a new server if the current one is removed from the system, if it crashes or if it ends in another partition. And when a new server is added, a crashed server recovers or when partitions merge, redundant instances of active servers should be deactivated. During transition periods, it is possible that some partitions contain zero or more than one active server. However, such periods should have bounded duration.

Figure **??** illustrates a partition-aware solution to the service activator problem based on PGMS. The collection of servers form a group. In this way, all relevant events for service activation/deactivation are transformed into view changes. Thus, the algorithm reduces to the management of $vchg()$ upcalls for the group. We assume there is a total ordering among server names such that the function $Min(S)$ returns the smallest name among the set $S$. At server $p$, local variable $server\_at$ identifies the server that $p$ believes to be active. A server starts out in the active state and blocks waiting for view change upcalls. This choice for the initial

```
procedure $PartitionableChat()$
  $view\_comp \leftarrow \{p\}$   % Current view composition
  $threadset \leftarrow \emptyset$   % Set of open threads
  $nthread \leftarrow 0$   % Number of open threads

  while true do
    wait-for $event$
    case $event$ of

      $vchg(v)$:
        if ($\overline{v} \not\subseteq view\_comp$) then
          foreach $id \in threadset$ do
            $mcast(\langle \text{UPDATE}, p, id, thread[id].vc \rangle)$
        $output($"Old participants:", $\overline{v} \cap view\_comp)$
        $output($"New participants:", $\overline{v})$
        $view\_comp \leftarrow \overline{v}$
        foreach $id \in threadset$ do
          $thread[id].msg \leftarrow \emptyset$

      $dlvr(\langle \text{UPDATE}, q, id, ts \rangle)$:
        $CheckNewId(id)$
        foreach ($q \in view\_comp - \{p\}$) do
          if $ts[q] > thread[id].vc[q]$ then
            $thread[id].vc[q] \leftarrow ts[q]$
        $CausalDelivery(id)$

      $newthread(m)$:
        $nthread \leftarrow nthread + 1$
        $id \leftarrow CreateNewId(p, nthread)$
        $CheckNewId(id)$
        $thread[id].lsent \leftarrow 1$
        $ts \leftarrow thread[id].vc$
        $ts[p] \leftarrow thread[id].lsent$
        $mcast(\langle \text{MESSAGE}, id, m, ts, p \rangle)$

      $shout(id, m)$:
        if $id \in threadset$ then
          $thread[id].lsent \leftarrow thread[id].lsent + 1$
          $ts \leftarrow thread[id].vc$
          $ts[p] \leftarrow thread[id].lsent$
          $mcast(\langle \text{MESSAGE}, id, m, ts, p \rangle)$
        fi

      $dlvr(\langle \text{MESSAGE}, id, m, ts, q \rangle)$:
        $CheckNewId(id)$
        $thread[id].msg \leftarrow thread[id].msg \cup \{(m, ts, q)\}$
        $CausalDelivery(id)$
    esac
  od

  procedure $CheckNewId(id)$
    if $id \notin threadset$ then
      $threadset \leftarrow threadset \cup \{id\}$
      $thread[id].vc \leftarrow (0, \ldots, 0)$
      $thread[id].lsent \leftarrow 0$
      $thread[id].msg \leftarrow \emptyset$
    fi
```

Figure 2. Partitionable Chat (Part a).

state is arbitrary since it will last only until the first view change reporting the current membership of the group. To handle a view change, the smallest server in the view composition is chosen as the new active server. If $p$ results as the new active server while it had been passive in the previous view, it is activated through $ActivateService()$. If, on the other hand, $p$ had been active in the previous view but some other server is designated to be active in the new one, then $p$ is made passive through $DeactivateService()$. In all other cases, the state of $p$ remains unchanged with respect to being active or passive.

Given the simplicity of the algorithm and the properties of PGMS, it is easy to argue that under stable conditions, each partition containing at least one server will eventually have exactly one active server.

## 5.2. Partitionable Chat

Consider a service, not unlike Internet Relay Chat (IRC), for holding a discussion among a collection of users. Users may contribute to the discussion by *creating* a new thread or by *shouting* messages in an existing thread. Messages are potentially addressed to every user who has joined the discussion. Upon a partitioning, the discussion may continue among users in each of the partitions. Users are informed about others with whom they are currently chatting. In some sense, this application extends the notion of partition awareness all the way up to the user level. Shouted messages have to satisfy agreement, integrity, uniqueness and liveness properties of view synchrony messages. Furthermore, messages shouted within the same partition and belonging to the same discussion thread should be seen in an order that is consistent with causal precedence. No requirements are placed on message threads that span multiple partitions. In other words, upon merging, a user may miss some messages that were shouted in other partitions. For this application, we consider it unreasonable to require causal order at a global level since this would force a user returning to the discussion after having been isolated to listen to the entire discussion that occurred during his absence before being able to resume.

Figures **??**-**??** contain the code of our partitionable chat algorithm. A user invokes the primitive $newthread(m)$ to start a new thread of discussion whose first message is $m$, and the primitive $shout(id, m)$ to shout message $m$ relative to the thread identified by $id$. Users receive messages shouted by others and control information from the system through $output()$ events. In order to distinguish between messages related to distinct threads, all multicast messages are tagged with a thread identifier; moreover, each user maintains a different set of variables for each known thread. Causal delivery within each thread is guaranteed through a system of vector clocks [**?**]. User $u$ maintains a sepa-

```
function $Deliverable(vc, ts, q)$
  return $(vc[q] = ts[q] - 1)$ and
          $(\forall r \in view\_comp, r \neq q : vc[r] \geq ts[r])$

procedure $CausalDelivery(id)$
  foreach $(m, ts, q) \in thread[id].msg$ do
    if $Deliverable(thread[id].vc, ts, q)$ then
      $output(id, m)$
      $thread[id].msg \leftarrow thread[id].msg - \{(m, ts, q)\}$
      $thread[id].vc[q] \leftarrow ts[q]$
    fi
```

Figure 3. Partitionable Chat (Part b).

rate vector clock for each thread and increments its local component whenever $u$ delivers one of its own messages in that thread. Each multicast message is timestamped with the current vector clock of the sender, except that the entry corresponding to the sender is replaced by the total number of messages the sender has multicast. This is motivated by the fact that the sender of a message will itself output the message not immediately after the multicast, but only after having delivered it (otherwise messages could be output in different views, due to the fact that messages are not guaranteed to be delivered in the view in which they have been multicast).

When a message is delivered, it is stored in a buffer until it can be output so as to maintain causal precedence within the discussion. Upon a view change, the user is informed about others that are currently in the discussion and those with which the communication is impossible. Two types of view changes need to be considered: if the new view is a contraction of the previous one, no further action is needed; users that survived from the previous view to the new one have delivered and output the same set of messages, and thus maintain the same vector clock values for each other user in the new view. If the view change represents an expansion, an additional merging protocol is needed: two merging partitions may have different sets of active threads, or may have output different messages for the same thread. For these reasons, upon a view expansion, users multicast for each known thread a message containing the thread identifier and the corresponding vector clock. When a user delivers such a message, the thread is added to the local list of threads (if unknown) and the vector clock is updated. In this manner, users will know exactly the same set of threads, and will be able to output all messages delivered during the new view without incurring in causal inconsistencies due to some past partitioning.

We outline how our solution exploits view synchrony. Properties GM**??**, GM**??** and VS**??** guarantee that each user will be informed of other users with whom discussion is impossible. And Property GM**??** ensures that this will happen only for those users who are effectively partitioned. Properties VS**??** and VS**??** are used in the vector clock merging protocol. Users surviving from one view to the same next view obtain identical vector clocks after the installation of the new view. Without this guarantee, it would be complex to reconstruct the new vector clocks necessary for users to deliver new messages. Furthermore, Property VS**??** guarantees that surviving users will deliver (and consequently output) the same set of messages during the first view. By Property VS**??**, all output messages have been previously multicast. Property VS**??**, combined with Property GM**??**, is used to avoid scenarios in which different overlapping views merge to a single common view. If such scenarios were admitted, two users $p$ and $q$ could output different sets of messages even though user $q$ always installed views including $p$.

## 5.3. Partitionable Parallel Computation

Consider a time-intensive computation such as ray tracing, prime factorization or weather forecasting. The computation can be decomposed into a number of jobs that can be carried out independently by a collection of *workers*. New workers may be added and existing ones removed at will. The computation and all relevant input data are known ahead of time to all possible workers. The goal of the *parallel computation* problem is to conclude the computation in as short a time as possible despite crashes, recoveries, partitionings and merges.

Our solution to the problem is illustrated in Figures **??** and **??**. Workers that will perform a job on behalf of the computation form a group. Within each view, the total work is equally distributed among the workers. During normal operation, each worker carries out the jobs assigned to it and diffuses the results through multicast messages. There are two types of view changes that need to be considered. If the new view represents a contraction of the previous one due to partitionings or process crashes, and the workers in the previous view know the same set of results, view synchrony guarantees that at the beginning of the new view all workers know the same set of results as well. Thus, it is sufficient to perform a redistribution of the uncompleted work. If, on the other hand, the new view represents an expansion due to network merges or process recoveries, then a "reconciliation" protocol is needed so as not to repeat jobs that may have already been completed by other workers. The reconciliation protocol may also be necessary in case workers disagree on the set of completed results due to an incomplete execution of a previous reconciliation protocol. Upon a view expansion, each worker $p$ locally assigns the remaining work among other workers that know the same set of results as itself (note that work cannot be reassigned among *all* workers of the new group membership since some of them may know different sets of results, leading to incon-

```
procedure $PartitionableParallelComputation(total)$
  $my\_work \leftarrow total$   % Subcomputations assigned to $p$
  $view\_comp \leftarrow \{p\}$   % Current view composition
  $lset \leftarrow \{p\}$   % Proc. that know less results
  $mset \leftarrow \{p\}$   % Proc. that know more results
  $res \leftarrow (\emptyset, \ldots, \emptyset)$   % Array of known results
  $new\_res \leftarrow \emptyset$   % Results delivered in this view
  $dist\_res \leftarrow \emptyset$   % Results merged in this view

  cobegin

    || Task 1:
    while true do

      wait-for $event$
      case $event$ of

        $vchg(v)$:
          foreach $q \in view\_comp \cap \overline{v}$ do
            $res[q] \leftarrow res[q] \cup (new\_res \cup dist\_res)$
          $view\_comp \leftarrow \overline{v}$
          $new\_res \leftarrow \emptyset$
          $dist\_res \leftarrow \emptyset$
          $lset \leftarrow lset \cap \overline{v}$
          $mset \leftarrow mset \cap \overline{v}$
          $my\_work \leftarrow Redist(ToDo(total, res[p]),$
                $lset \cap mset, p)$
          if $mset \neq \overline{v}$ and $Min(mset) = p$ then
            $mcast(\langle \text{UPDATE}, \bigcup_{r \in \overline{v}}(res[p] - res[r]), lset\rangle)$

        $dlvr(\langle \text{UPDATE}, rset, pset\rangle)$:
          foreach $q \in pset$ do
            $res[q] \leftarrow res[q] \cup rset$
          $dist\_res \leftarrow dist\_res \cup rset$
          $lset \leftarrow lset \cup pset$
          if $p \in pset$ then
            $mset \leftarrow view\_comp$
          if $mset = lset = view\_comp$ then
            $my\_work \leftarrow Redist(ToDo(total,$
                    $res[p] \cup dist\_res), view\_comp, p)$

        $dlvr(\langle \text{RESULT}, w, r\rangle)$:
          $new\_res \leftarrow new\_res \cup \{(w, r)\}$

      esac
    od

    || Task 2:
    $ComputeResults()$
  coend

  function $ToDo(T, C)$
    return $\{w | w \in T \wedge \nexists r : (w, r) \in C\}$
```

Figure 4. Partitionable Parallel Computation (Part a).

```
  procedure $ComputeResults()$
    $R \leftarrow new\_res \cup dist\_res$
    while $ToDo(total, res[p] \cup S) \neq \emptyset$ do
      $w \leftarrow Min(ToDo(my\_work, res[p] \cup S))$
      $my\_work \leftarrow my\_work - \{w\}$
      if $w \neq$ null then
        $mcast(\langle \text{RESULT}, w, PerformWork(w)\rangle)$
    od
    $res[p] \leftarrow res[p] \cup S$
```

Figure 5. Partitionable Parallel Computation (Part b).

sistencies). Then, each worker $p$ elects a leader $c$ among the set of all workers that know a superset of the results known to $p$ itself. Leader $c$ acts as a representative for all workers that know a subset of the results known to itself, and multicasts a message containing the results that could be unknown to other workers in the view. The choice of electing a leader is motived by efficiency: only one worker from each merging partition performs reconciliation. When the reconciliation protocol is concluded (each worker has delivered one message for each of the partitions that have merged in the new view), a new reassignment of the uncompleted work is performed among all workers in the new membership. Otherwise, in case of further failures (a leader crashes or ends in a different partition before multicasting the reconciliation message), the reconciliation protocol may have to continue in subsequent views.

We conclude by sketching how our solution exploits properties of view synchrony. Property GM**??** guarantees that work is always distributed among all workers within a given partition, thus achieving the maximum possible speedup. Properties GM**??**, VS**??**, VS**??** and VS**??** allow a worker to reason locally about other workers that know the same set of results as itself. In particular, Property VS**??** guarantees that two workers surviving from a view to the same next view will deliver the same set of messages (i.e., both new results and results exchanged during the reconciliation protocol). Property VS**??** ensures that each reconciliation message will be delivered in the same view in which it has been multicast, simplifying the algorithm since workers do not have to worry about obsolete messages multicast during previous views. Properties GM**??**, GM**??** and VS**??** guarantee that a worker will eventually deliver the results of all jobs assigned to other workers, or it will install a new view and thus reassign the work. Finally, Property VS**??** guarantees that no spurious messages will be delivered.

## 6. Conclusions

Specifications for services in asynchronous distributed systems require a delicate balance between two conflicting goals: they must be strong enough to be useful and exclude trivial solutions, yet they must be weak enough to be implementable [**?**]. The support layer we have specified in this paper has benefited from extended reflection and several revisions. The main objective of the current work has been to argue that realistic and interesting partition-aware applications can indeed be developed with ease on top of such a layer. The question of how our specification can be implemented on top of a typical operating system and an unreliable datagram communication service is treated in another work [**?**].

Information about group compositions conveyed through views and view changes happens to be appropriate for achieving partition awareness in network applications. Through a single mechanism (view change), we are able to abstract a large number of environment characteristics resulting from complex scenarios due to administrative intervention, crashes, recoveries, partitionings and merges. Group membership services alone are sufficient for only the simplest of network applications. More realistic network applications typically require closer cooperation among their components through communication. For this purpose, we have extended PGMS with a communication service based on view synchrony. The strong semantics provided by view synchrony regarding the composition and installation of views, delivery of messages, and most importantly, the integration between them, allow sophisticated global reasoning to be accomplished through local information alone and without having to resort to complex communication protocols.

The partition-aware applications we have developed reconfigure themselves in order to provide the "best" quality that is possible for their services in each partition. Obviously, no support layer can accomplish miracles and guarantee services at constant quality levels despite partitionings. Partition-aware applications we have considered have the simplifying characteristic that it is always possible to restore a consistent global state after recoveries or merges. In general, applications that admit conflicting operations (e.g., updates to replicated data) in concurrent partitions will require additional system support for restoring a meaningful global state (if at all possible) after recoveries and merges [**?**].

## References

[1] O. Amir, Y. Amir, and D. Dolev. A highly available application in the transis environment. In *Proc. of the Hardware and Software Architectures for Fault Tolerance Workshop*, Le Mont Saint-Michel, France, June 1993.

[2] Y. Amir, D. Breitgand, G. Chockler, and D. Dolev. Group communication as an infrastructure for distributed system management. In *Proc. of the 3rd Int. Workshop on Services in Distributed and Networked Environments*, Macau, June 1996.

[3] E. Anceaume, B. Charron-Bost, P. Minet, and S. Toueg. On the formal specifi cation of group membership services. Tech. Rep. TR95-1534, Dept. of Computer Science, Cornell Univ., Aug. 1995.

[4] Ö. Babaoğlu, A. Bartoli, and G. Dini. Enriched view synchrony: A programming paradigm for partitionable asynchronous distributed systems. *IEEE Trans. on Computers*, 46(6):642–658, June 1997.

[5] Ö. Babaoğlu, R. Davoli, L. Giachini, and M. Baker. Relacs: A communications infrastructure for constructing reliable applications in large-scale distributed systems. In *Proc. of the 28th Hawaii Int. Conf. on System Sciences*, pages 612–621, Maui, Hawaii, Jan. 1995.

[6] Ö. Babaoğlu, R. Davoli, and A. Montresor. Group membership and view synchrony in partitionable asynchronous systems: Specifi cation and algorithms. Tech. Rep. UBLCS-98-1, Dept. of Computer Science, Univ. of Bologna, Mar. 1998.

[7] Ö. Babaoğlu, R. Davoli, A. Montresor, and R. Segala. System support for partition-aware network applications. Tech. Rep. UBLCS-97-8, Dept. of Computer Science, Univ. of Bologna, Oct. 1997.

[8] K. Birman. The process group approach to reliable distributed computing. *Commun. ACM*, 36(12):36–53, Dec. 1993.

[9] F. Cosquer, P. Antunes, and P. Verissimo. Enhancing dependability of cooperative applications in partitionable environments. In *Proc. of the 2nd European Dependable Computing Conf.*, Taormina, Italy, Oct. 1996.

[10] F. Cosquer and P. Verissimo. Large scale distribution support for cooperative applications. In *Proc. of the European Reasearch Seminar on Advances in Distributed Systems (ERSADS)*, L'Alpe d'Houez, France, Apr. 1995.

[11] D. Dolev, D. Malki, and R. Strong. A framework for partitionable membership service. Tech. Rep. CS95-4, Institute of Computer Science, The Hebrew Univ. of Jerusalem, 1995.

[12] R. Friedman and A. Vaysburd. Implementing replicated state machines over partitionable networks. Tech. Rep. TR96-1581, Dept. of Computer Science, Cornell Univ., 1996.

[13] F. Kaashoek and A. Tanenbaum. Group communication in the Amoeba distributed operating system. In *Proc. of the 12th IEEE Symp. on Reliable Distributed Systems*, pages 222–230, Arlington, TX, May 1991.

[14] I. Keidar and D. Dolev. Effi cient message ordering in dynamic networks. In *Proc. of the 15th ACM Symp. on Principles of Distributed Computing*, Philadelphia, PA, May 1996.

[15] F. Mattern. Virtual time and global states of distributed systems. In M. C. et. al., editor, *Proc. of the Int. Workshop on Parallel and Distributed Algorithms*, pages 215–226. North-Holland, Oct. 1989.

[16] L. Moser, Y. Amir, P. Melliar-Smith, and D. Agarwal. Extended virtual synchrony. In *Proc. of the 14th Int. Conf. on Distributed Computing Systems*, Poznan, Pol, June 1994.

[17] A. Ricciardi and K. Birman. Using process groups to implement failure detection in asynchronous environments. In *Proc. of the 10th ACM Symp. on Principles of Distributed Computing*, pages 341–352, Aug. 1991.

[18] R. van Renesse, K. Birman, R. Cooper, B. Glade, and P. Stephenson. The horus system. In K. Birman and R. van Renesse, editors, *Reliable Distributed Computing with the Isis Toolkit*, pages 133–147. IEEE Computer Society Press, 1993.