Passing Messages while Sharing Memory

Marcos K. Aguilera
VMware Research

Naama Ben-David
CMU

Irina Calciu
VMware Research

Rachid Guerraoui
EPFL

Erez Petrank
Technion

Sam Toueg
University of Toronto

ABSTRACT

We introduce a new distributed computing model called m&m that allows processes to both pass messages and share memory. Motivated by recent hardware trends, we find that this model improves the power of the pure message-passing and shared-memory models. As we demonstrate by example with two fundamental problems—consensus and eventual leader election—the added power leads to new algorithms that are more robust against failures and asynchrony. Our consensus algorithm combines the superior scalability of message passing with the higher fault tolerance of shared memory, while our leader election algorithms reduce the system synchrony needed for correctness. These results point to a wide new space for future exploration of other problems, techniques, and benefits.

ACM Reference Format:

1 INTRODUCTION

The distributed computing community has a dichotomy between shared-memory and message-passing models. Books, courses, and papers explicitly separate these models to present results and algorithms. These models differ based on how processes communicate. In the shared-memory model, processes can write and read data in a common area of memory. In the message-passing model, processes can send and receive messages to and from each other.

In this paper, we investigate the benefits of a hybrid model, called message-and-memory model or simply m&m model, where processes can both pass messages and share memory. We are motivated by recent technologies that permit exactly that, such as Remote Direct Memory Access (rdma) [39, 41, 65], disaggregated memory [55], and Gen-Z [33]. By using both methods of communication, one can devise a new genre of algorithms that could potentially combine the advantages of shared-memory and message-passing algorithms.

\*Work done in part while the author was at VMware Research.

It has been proven that the message-passing and shared-memory models are equivalent [11], by demonstrating that one model can simulate the other. If that is so, what could be the benefit of combining two models that are equivalent? Closer inspection of the equivalence result reveals that it holds only in a certain sense and under some assumptions. In particular, the equivalence is computational: it shows how algorithms for a problem in one model can be translated to the other using an emulation. However, the emulation does not preserve efficiency or synchrony (e.g., a timely process in one model can become untimely as it waits for other processes). Moreover, the emulation requires the assumption that the message-passing model has a majority of correct processes, and that processes know each other’s identities as well as the number of processes in the system. In many cases these assumptions do not hold.

In fact, we find that each model has its own advantages and they benefit algorithms in different ways. Our contribution is to propose the new m&m model that merges the capabilities of the pure models, to identify some of the advantages of each pure model, and to show that they can be combined in the m&m model.

To illustrate the advantages of combining the models, consider the fundamental problem of mutual exclusion [26]. In this problem, there is a doorway and a critical section. The goal is to ensure that at most one process in the doorway enters the critical section at a time. The traditional algorithms for this problem, such as the bakery algorithm or Dijkstra’s algorithm, have been designed in the shared-memory read-write model [51]. In that abstract model, these algorithms have a common drawback: while some process is in the critical section, other processes in the doorway must spin on one or more shared-memory locations to know when the critical section becomes empty again. Much work was devoted to mitigate this problem, by making the spin local to each process (e.g., [23, 46–48]).

By allowing an algorithm to share memory and pass messages in the m&m model, one finds very simple solutions that do not have to spin: upon leaving the critical section, a process could send a message to the processes in the doorway; those processes, rather than spinning, go to sleep and resume execution when a message arrives. This ability to react to data without spinning is a characteristic of message passing. In practice, by avoiding the spin, the CPU can be better utilized to run other processes. Such benefits escape the abstract equivalence of the message-passing and shared-memory models in [11].

Besides benefits, each model has limitations that the m&m model can overcome. Shared-memory systems have worse scalability than message-passing systems due to hardware limitations. For example, a typical shared-memory system today has tens to thousands of processes, while message-passing systems can be much larger (e.g.,...
map-reduce systems with tens of thousands of hosts [25], peer-to-peer systems with hundreds of thousands of hosts, the SMTP email system and DNS with hundreds of millions of hosts, etc).

On the other hand, message-passing systems have limitations on fault tolerance and synchrony. On fault tolerance, some fundamental problems in distributed computing—such as consensus and atomic storage—require a majority of correct processes to be solvable in the message-passing model, even under strong partial synchrony assumptions, whereas the same problems can be solved in shared memory with an arbitrary number of correct processes using wait-free algorithms under partial synchrony, randomization, or stronger hardware primitives. With respect to synchrony, due to engineering reasons, message-passing systems have larger variances in their timing than shared memory: the slowest and fastest message delays can be eight orders of magnitude apart (microseconds to tens of seconds), while the variance in the execution speeds of processes in shared memory is much smaller. As a result, partial synchrony bounds tend to be much worse in message-passing systems.

In this paper, we make these advantages and drawbacks more precise. We show that the m&m model can enhance the robustness of algorithms. We consider two aspects of robustness: the number of process crashes that algorithms can tolerate and the synchrony assumptions required by them. We illustrate how we improve the first aspect through the consensus problem and the second through the eventual leader election problem. In short, we show that message-passing and shared memory complement each other when combined, allowing algorithms that have inherent advantages over the pure message-passing and shared-memory models.

For consensus, we give an algorithm called Hybrid Ben-Or or simply HBO that tolerates more than a majority of failures (like shared-memory algorithms can), while scaling to a large system size (like message-passing systems can). To scale to a large system, the algorithm limits the number of processes that share a given shared-memory location to a small constant number. We define an undirected shared-memory graph, whose nodes are processes and there is an edge between processes $p$ and $q$ if they share a memory location. Due to hardware limitations, to scale the system we must limit the maximum degree $d$ of this graph [28, 43]. Our algorithm employs expander graphs to tolerate a majority of crash failures—up to $f < \frac{1}{2d+1}$ of them—in a system with $n$ processes, where $h$ is the expansion of the graph as measured by the vertex expansion ratio. Roughly, this ratio indicates by how much a set of vertices expands each time we add their neighbors to the set. The higher the expansion, the more failures the algorithm can tolerate. Our algorithm is a simulation of a pure message-passing consensus algorithm that requires a majority of correct processes, without having that majority in reality. To do that, we use a wait-free shared-memory consensus algorithm among each local neighborhood in the shared-memory graph to emulate a virtual process in the larger message-passing algorithm. This virtual process fails only if all processes in its neighborhood fail. By taking a shared-memory graph with high expansion, we ensure that even with a small $d$, many processes can fail without affecting a majority of virtual processes. Here, the topology of the shared-memory graph determines the fault tolerance of consensus: graphs with higher expansion allow for higher fault tolerance, because correct processes are adjacent to (and thus can simulate) more processes. We show that this relation is inherent by giving an impossibility result relating graph expansion and fault tolerance.

We next turn our attention to the (eventual) leader election problem. This problem is traditionally considered in message-passing systems, and much effort has been devoted to find the weakest synchrony needed to solve it [5, 6, 38]. Prior algorithms required some timeliness on processes and communication links. We show that in the m&m model, we can do better by using both shared memory and message passing to obtain different benefits. We use shared memory to reduce the timeliness requirement to processes only: our leader election algorithms require only that a single process, the leader, increments its local heartbeat in a timely fashion; other processes can be asynchronous—in particular, they can be arbitrarily slow to read the heartbeat. We use message passing to provide a trade-off between message reliability and amount of work in steady state: we give two algorithms for different types of links. (1) With reliable links\(^1\), the only steady-state work is that the leader periodically increments a local heartbeat counter in shared memory, while other processes read the counter; (2) With fair lossy links\(^2\), in addition to the above, the leader also periodically reads a shared register. In either algorithm, no messages are exchanged in the steady state, and all the communication links can be asynchronous. We further prove that the two leader election algorithms are optimal in the following sense. For systems with a timely process and asynchronous links, any algorithm requires the leader to write a shared register periodically (as in both our algorithms). This result holds whether links are reliable or fair lossy. Moreover, with fair lossy links, there are more requirements: either the leader writes and reads shared registers periodically (as our second algorithm), or some process keeps sending messages forever.

To summarize, the contributions of this paper are the following:

- We motivate and introduce the m&m model of distributed computing, which allows processes to both share memory and pass messages.
- We study the consensus problem under the m&m model. We give a new algorithm that improves on the fault tolerance of message-passing algorithms, while limiting memory sharing to allow for scalability. We show that the algorithm’s fault tolerance is improved by a shared-memory graph with high expansion, and prove an impossibility result showing that this relationship to expansion is inherent.
- The new consensus algorithm introduces a simulation technique that combines shared memory and message passing. We believe this technique is interesting in its own right, as it could be used to improve other algorithms, to show impossibility results, etc.
- We study the leader election problem under the m&m model and give algorithms that reduce the synchrony required, while maintaining a low communication complexity. We show that these algorithms are tight in their communication.

\(^1\)I.e., links that do not drop messages.
\(^2\)I.e., links that can drop messages, but if a message is repeatedly sent then it is eventually received.
While we focus on two problems and some benefits of the m&m model, we believe the paper opens a large space for future research on other problems and benefits.

Due to space limitations, we omit proofs and some details. They will be included in the full version of this paper.

2 RELATED WORK

The m&m model is motivated by emerging technologies, such as Remote Direct Memory Access (RDMA) [39, 41, 65], disaggregated memory [31, 55], Gen-Z [33], and OmniPath [40]. These technologies provide remote memory [4] and can be unified under higher-level abstractions [3]. RDMA permits a process to access the memory of a remote host without interrupting the remote processor. It has been widely used in high-performance computing [72] and is now being adopted in modern data centers [34]. Work on RDMA shows how it can improve the performance of important applications, such as key-value storage systems [28, 43, 60], database systems [13, 73], distributed file systems [58], and more [29, 36, 67, 68]. Recent work uses RDMA to improve performance of consensus [64, 70] assuming a majority of processes are correct. Disaggregated memory separates compute and memory, and connects them using a fast network; prior work proposes new architectures for disaggregated memory [9, 22, 62] and studies the network [35] and system [4, 56] requirements for a practical implementation. Gen-Z and OmniPath are commercial technologies under development that offer memory semantics and low-latency access to remote data.

The shared-memory and the message-passing models are well studied in academic research, and have been compared under both theoretical and practical considerations [17, 18, 49, 54]. The two models have been shown to be computationally equivalent [11], though for efficiency, simplicity, or hardware availability, one might prefer one model over the other. For instance, Barrelfish [14] uses message passing to improve performance on a shared-memory multicore machine [24]. Conversely, distributed shared-memory systems [8, 16, 66] offer the abstraction of shared memory on top of a message-passing system. Recent work improves the performance of such systems using RDMA [45, 61]. Integrating message passing and shared memory in hardware has been explored in the MIT Aline machine [50]. Our work differs in that (1) we propose a new abstract distributed computing model, which encapsulates low-latency remote access technologies, such as RDMA and disaggregated memory, and (2) we show that this model can improve the robustness of algorithms, rather than the performance or simplicity of applications.

Consensus is a fundamental problem in distributed computing. Following the well-known FLP result [32] showing that it cannot be solved in an asynchronous crash-prone message-passing system, much work has focused on getting around the impossibility by using randomization [10, 12, 15], partial synchrony [27, 30], or unreliable failure detectors [19, 20]. In fact, the eventual leader election problem, Ω, is the weakest failure detector that can solve consensus, and is used in several algorithms [53, 63]. This is the problem that we study in the second part of the paper. The eventual leader election problem is known to require some synchrony to implement. We show that the m&m model permits solutions with less synchrony than before, while minimizing the work done.

2 RELATED WORK

The m&m model is motivated by emerging technologies, such as Remote Direct Memory Access (RDMA) [39, 41, 65], disaggregated memory [31, 55], Gen-Z [33], and OmniPath [40]. These technologies provide remote memory [4] and can be unified under higher-level abstractions [3]. RDMA permits a process to access the memory of a remote host without interrupting the remote processor. It has been widely used in high-performance computing [72] and is now being adopted in modern data centers [34]. Work on RDMA shows how it can improve the performance of important applications, such as key-value storage systems [28, 43, 60], database systems [13, 73], distributed file systems [58], and more [29, 36, 67, 68]. Recent work uses RDMA to improve performance of consensus [64, 70] assuming a majority of processes are correct. Disaggregated memory separates compute and memory, and connects them using a fast network; prior work proposes new architectures for disaggregated memory [9, 22, 62] and studies the network [35] and system [4, 56] requirements for a practical implementation. Gen-Z and OmniPath are commercial technologies under development that offer memory semantics and low-latency access to remote data.

The shared-memory and the message-passing models are well studied in academic research, and have been compared under both theoretical and practical considerations [17, 18, 49, 54]. The two models have been shown to be computationally equivalent [11], though for efficiency, simplicity, or hardware availability, one might prefer one model over the other. For instance, Barrelfish [14] uses message passing to improve performance on a shared-memory multicore machine [24]. Conversely, distributed shared-memory systems [8, 16, 66] offer the abstraction of shared memory on top of a message-passing system. Recent work improves the performance of such systems using RDMA [45, 61]. Integrating message passing and shared memory in hardware has been explored in the MIT Aline machine [50]. Our work differs in that (1) we propose a new abstract distributed computing model, which encapsulates low-latency remote access technologies, such as RDMA and disaggregated memory, and (2) we show that this model can improve the robustness of algorithms, rather than the performance or simplicity of applications.

Consensus is a fundamental problem in distributed computing. Following the well-known FLP result [32] showing that it cannot be solved in an asynchronous crash-prone message-passing system, much work has focused on getting around the impossibility by using randomization [10, 12, 15], partial synchrony [27, 30], or unreliable failure detectors [19, 20]. In fact, the eventual leader election problem, Ω, is the weakest failure detector that can solve consensus, and is used in several algorithms [53, 63]. This is the problem that we study in the second part of the paper. The eventual leader election problem is known to require some synchrony to implement. We show that the m&m model permits solutions with less synchrony than before, while minimizing the work done.
In practice, a shared register is provided by the hardware, but not all processes may be able to access all registers, as there may be limits on the number of processes that can share the same memory [28, 43, 44, 69]. We define the shared-memory domain \( S \) as a set of process subsets; intuitively, \( S \) determines what subsets of processes can share memory. More precisely, for each set \( S \subseteq S \), the model permits having any number of registers shared among processes in \( S \). In general, \( S \) can be arbitrary. However, in practice memory sharing is simpler, as the hardware technology naturally imposes a structure on \( S \); for example, a process might be able to share memory only with processes that connect to it over the underlying hardware. We say that \( S \) is uniform if it can be represented by an undirected graph \( G_{SM} \) of processes, such that registers can be shared by a process and its neighbors in \( G_{SM} \); intuitively, \( G_{SM} \) is the graph of connections of the underlying hardware that implements the shared memory. Formally, \( G_{SM} \) is a graph \( G_{SM} = (\Pi, E_{SM}) \) and the sets in \( S \) are exactly the sets consisting of a process \( p \) and its neighbors in \( G_{SM} \). That is, if we let \( S_p = \{ p \} \cup \{ q : (p, q) \in E_{SM} \} \) then \( S = \{ S_p : p \in \Pi \} \). For a uniform \( S \), we say that \( G_{SM} \) is its shared-memory graph. Figure 1 gives an example. In this paper, we are interested in the uniform model, and all our results work with the graph \( G_{SM} \). The broader model based on \( S \) is provided to allow for future theoretical work and potential new hardware platforms. Note that, while the model does not constrain the number or size of registers that can be shared, algorithms may choose to reduce their shared-memory usage for efficiency.

In systems with few processes (e.g., in the tens), \( G_{SM} \) could be a fully connected graph, but systems with lots of processes may have to limit the maximum degree of \( G_{SM} \) (e.g., limit the connections over the hardware).

We assume that the shared memory does not fail, as in the pure shared-memory model. This assumption can be supported by the hardware: with RDMA, the shared memory can be registered with the kernel so that it remains accessible after processes crash; disaggregated memory can similarly preserve memory accesses after process crashes.

**Synchrony.** For some results, we make some partial synchrony assumptions about the relative execution speed of processes. We first define what it means for a process \( p \) to be timely with respect to another process \( q \):

- **[Pairwise timeliness]:** We say that \( p \) is \( q \)-timely (in a run) if \( p \) is correct and there is an integer \( i \geq 1 \) such that every time interval containing \( i \) steps of \( q \) has at least one step of \( p \).

  The timeliness bound \( i \) above is not known to processes: it may depend on each run and each pair of processes \( p \) and \( q \). We now define what it means for a process \( p \) to be timely:

- **[Timeliness]:** We say that \( p \) is timely (in a run) if \( p \) is \( q \)-timely for every process \( q \in \Pi \) (in this run).

  Intuitively, timeliness means that eventually the process executes within a bounded rate relative to other processes. This is a weak requirement in many ways. First, it is relative to the speed of other processes, so a process can be timely even if it slows down arbitrarily in real time. Second, the bound on execution rate need not be known. Third, the bound is arbitrary and not fixed a priori, so timeliness can be satisfied even if a process is initially arbitrarily slow relative to others.

We consider two types of systems: (1) asynchronous systems, which might not have any timely processes, and (2) systems with little synchrony, that is, systems where at least one process is timely; this process can vary from run to run and is not known to the processes. We make no assumptions about the timeliness of messages.

**Consensus problem.** In the consensus problem, each process begins with an input value \( v \in \{0, 1\} \) which it proposes, and it decides on an output value in the end. The output values must satisfy three properties:

- **[Uniform Agreement]:** No two processes decide on a different value.
- **[Validity]:** If a process decides on a value \( v \) then \( v \) was proposed by some process.
- **[Termination]:** Every correct process eventually decides.

We allow randomized solutions, where the above Termination property must hold with probability 1 under a strong adversary—one that can schedule processes based on their current state and past history.

A consensus object is a shared-memory object with one operation, \( \text{propose}(v) \), which takes a value \( v \) and returns the first value that was proposed to the object.

**The eventual leader election problem.** This problem is formally defined in [20] as the \( \Omega \) failure detector. Informally, each process \( p \) outputs a single process denoted \( \text{leader}_p \), such that the following property holds:

- There is a correct process \( \ell \) and a time after which, for every correct process \( p \), \( \text{leader}_p = \ell \).

  Note that, at any given time, processes do not know if there is a commonly agreed leader; they only know that eventually there will be a common leader.

4 CONSENSUS

We show that the m&m model can be used improve the fault tolerance of algorithms compared to a pure message-passing system. It is known that consensus cannot be solved deterministically in an asynchronous system subject to failures, even if processes can only fail by crashing and at most one process may fail; this is true in both message-passing and shared-memory models [32, 57]. We thus consider asynchronous systems where processes can toss coins. Then, in a shared-memory system, consensus can be solved (with probability 1) with up to \( n-1 \) crash failures [1] (\( n \) is the number of processes in the system), whereas in a message-passing system, a consensus algorithm can tolerate at most \([(n-1)/2] \) crash failures [15]. We show that the m&m model can strike a balance between shared memory and message passing.

First note that if the shared-memory graph \( G_{SM} \) is fully connected then any fault-tolerant shared-memory algorithm also works in the m&m model—the algorithm simply never sends messages. Thus, there are algorithms in the m&m model that can tolerate up to \( n-1 \) crash failures. However, in a large system, it is impractical to connect all processes over shared memory (§3). When fewer

---

1Henceforth, when we say that there is a time after which some property \( C \) holds, we mean that there is a time \( t \) such that, for every time \( t' \geq t \), property \( C \) holds at time \( t' \).
In the message-passing model, this is one of the simplest consensus algorithms [10, 12], which work in the m&m model because neighbors in $G_M$ share memory.

We call this algorithm the Hybrid Ben-Or or HBO algorithm. Figure 2 shows the pseudocode. There, processes do not terminate after deciding, but it is easy to modify the algorithm so that they do. This algorithm always satisfies the safety properties of consensus, irrespective of the number of crash failures:

**Theorem 4.1.** The HBO algorithm in Figure 2 satisfies the Validity and Uniform Agreement properties of consensus in the m&m model with reliable links.

The proof of this theorem is given in the full paper. There, we also show that the HBO algorithm terminates as long as a majority of the processes are represented.

**Theorem 4.2.** The HBO algorithm in Figure 2 satisfies the Termination property of consensus with probability 1 in the m&m model with reliable links where a majority of the processes are represented.

In the next section, we consider how many failures may occur while still ensuring that a majority of processes are represented.

### 4.2 Shared-memory expanders

In this section, we consider the fault tolerance of the HBO algorithm: how many crash failures can it tolerate while ensuring that
processes decide. In the algorithm, correct processes represent their neighbors in \( G_{SM} \), so the fault tolerance depends on \( G_{SM} \) and how many neighbors correct processes have. We show that, by choosing \( G_{SM} \) to be a graph with high expansion, we obtain the best trade-off between maintaining low degree and achieving high fault tolerance. Having a low degree is important because the degree indicates how many neighbors correct processes have. We show that, by choosing \( G_{SM} \) to have a low degree, the system is related to the minimum cut that separates a large subgraph from the rest of the \( G_{SM} \) graph. In graphs with high expansion, the size of such a cut it guaranteed to be large, and thus many failures can be tolerated.

To establish the impossibility, we extend the well-known partitioning argument [29] to the m&m model. Basically, if two processes cannot communicate during the execution of an algorithm, then they cannot decide the same value. Thus, if the adversary can partition the system into two disjoint subgraphs \( A \) and \( B \), each of size \( \geq n - f \), where processes in \( A \) do not communicate with processes in \( B \), then agreement cannot hold. This argument works in message-passing models, where the adversary can arbitrarily delay messages on the network, but it breaks in a shared-memory model, in which communication between processes cannot be delayed without blocking the processes themselves. Thus, in the m&m model, to create such a partition the adversary must get rid of all shared-memory edges of \( G_{SM} \) on the cut between \( A \) and \( B \).

We now formalize the intuition to arrive at the impossibility. Given a graph \( G = (V, E) \), we say that \( C = (B, S, T) \) is an SM-cut in \( G \) if \( B, S, T \), and are disjoint subsets of \( V \) such that \( B \cup S \cup T = V \), and there is a way to partition \( B \) into two disjoint subsets \( B_1 \) and \( B_2 \) such that \( (B_1 \cup S, B_2 \cup T) \) is a cut of the graph \( G \), and for every \( b_1 \in B_1, b_2 \in B_2, s \in S \) and \( t \in T \), \( \{s, t\} \notin E \), \( \{b_1, t\} \notin E \), and \( \{b_2, s\} \notin E \). Intuitively, \( B \) is the set of vertices on the boundary of the cut, and \( S \) and \( T \) are the remaining vertices on each side.

**Theorem 4.4.** Consider the m&m model with shared-memory graph \( G_{SM} \), where links are reliable and \( f \) processes may crash. Consensus cannot be solved if \( G_{SM} \) has an SM-cut \((B, S, T)\) with \( |S| \geq n - f \) and \( |T| \geq n - f \).

We prove this theorem in the full paper. Note that, in a graph with high expansion, there are no SM-cuts \((B, S, T)\) with \( |S| \geq n - f \) and \( |T| \geq n - f \). Intuitively, this is because if we want to build an SM-cut and we start with some set \( S \) with \( |S| \geq n - f \), we must include \( S \) in \( B_1 \), and then include \( S \cup B_1 \) in \( B_2 \). As these sets expand quickly, we are then left with fewer than \( n - f \) vertices to put in \( T \). In the full paper, we formalize the above intuition to relate the impossibility to the expansion properties of \( G_{SM} \).

## 5 LEADER ELECTION

We now show that the m&m model allows us to not only improve the fault tolerance of message-passing systems, but also to reduce the synchrony needed to solve certain problems. To demonstrate that, we turn our attention to the (eventual) leader election problem. In this problem, each process has a leader, and the goal is for all correct processes to eventually have the same correct leader (this is also known as the \( \Omega \) failure detector [29]). Leader election is used in several well-known consensus algorithms, such as Paxos [52], Raft [63], and CT [20]. To be solvable, leader election requires the system to have some partial synchrony (because it can be used to solve consensus, and consensus is impossible in asynchronous systems [32]). Finding the weakest models of synchrony for solving this problem was the goal of several papers, but all known leader election algorithms for message-passing systems require some synchrony on at least some of the network communication links. In practice, it can be hard to guarantee small bounds on network delays, thus leading to high recovery time when a leader crashes.

We show that the m&m model permits solutions with almost no synchrony: the only requirement is that some process be timely.
give two algorithms: one assumes reliable links (§5.1), and the other relaxes that requirement and assumes only fair lossy links (§5.2). In both algorithms, the leader regularly increments a heartbeat counter in shared memory, and other processes verify that the leader is alive by monitoring and timing out on this counter. With fair lossy links, the leader in addition periodically reads a register. We can make it easier for the leader to be timely, by placing the shared registers so that eventually the leader accesses only local registers (§5.3). We show that our algorithms are tight in efficiency (§5.4). In this section, we assume that \( G_{SM} \) is the complete graph.

5.1 Algorithm for reliable links

The basic idea of the first algorithm is that each process \( p \) has a “badness” counter that it shares with other processes. Intuitively, this badness counter represents the number of times that other processes suspected \( p \) of having crashed. To pick its leader, \( p \) keeps a set of processes, called contenders, that are contending for leadership; this set always includes \( p \) and initially contains no other processes. If there are no other contenders, \( p \) picks itself as the leader; otherwise, it picks the process with the smallest badness counter. Note that at this point, different processes could pick different leaders. We show that our algorithm always eventually realizes and corrects such situations.

When \( p \) becomes its own leader, it announces its leadership to other processes using a notification mechanism; here, this mechanism simply sends a message to the other processes (in our next algorithm, the mechanism is more complex because messages can be lost). If \( p \) thinks it is the leader, it sets an active bit in shared memory to indicate that it believes itself to be the leader. Then, it periodically increments a heartbeat counter in shared memory to tell others that it is alive. Process \( p \) also periodically checks whether it got any notifications from another process \( q \); if it did, this means that \( q \) also wants to be the leader. So, \( p \) adds \( q \) to the contenders set, and \( p \) starts a timer on \( q \); in effect, \( p \) now monitors \( q \) to see if it remains timely. Lastly, upon adding \( q \) to its contenders set, \( p \) also notifies \( q \) that \( p \) is also a contender, in case \( q \) does not know.

Whether \( p \) is a leader or not, \( p \) monitors the processes in its contender set other than itself. Intuitively, \( p \) expects each contender \( q \neq p \) to periodically increment \( q \)'s heartbeat counters in shared memory. If \( q \) fails to do so within a timeout,\(^4\) \( p \) removes \( q \) from its contenders set; \( p \) then checks whether \( q \) has the active bit set in shared memory; if it does, that means \( q \) thinks it is the leader, so \( p \) sends an accusation to \( q \) and increments the timeout value. While \( p \) is its own leader, it also checks whether it received any accusation messages. If \( p \) receives an accusation, it increments its badness counter. If \( p \) stops thinking it is the leader, it clears the active bit in shared memory. The active bit is critical for correctness; it prevents processes from sending an accusation to \( p \) after it relinquishes leadership and stops incrementing its heartbeat.

Processes who believe themselves to be the leader fight among themselves for leadership: as described above, they notify each other about their desire to be the leader, adding each other to their contender set, and they pick the contending process with the smallest badness counter as the winning leader. We now explain how all correct processes eventually choose the same leader forever.

First note that every timely process eventually stops being accused (because it increases its heartbeat in a timely fashion when it thinks it is the leader, and it clears its active bit when it thinks it is not the leader). By assumption, the system has at least one timely process, thus there is at least one correct process that stops receiving accusations, and so its badness counter stops growing. Let \( \ell \) be such a correct process whose badness counter is smallest. Therefore, for every process \( q \neq \ell \), either \( q \) has a badness counter that eventually grows larger than \( \ell \)'s badness counter, or \( q \) crashes (and so every correct process eventually times out on \( q \) and removes \( q \) from its contender set). Since the contender set of \( \ell \) contains \( \ell \) forever, it is clear that \( \ell \) eventually selects itself as the leader forever. When \( \ell \) becomes leader, it notifies all the correct processes. Eventually other processes stop thinking they are leader: if a process \( q \neq \ell \) thought it were leader infinitely often, it would check notifications infinitely often and eventually get a notification from \( \ell \), add \( \ell \) to its contender set, see that \( \ell \) has a smaller badness counter, and then pick \( \ell \) rather than itself as its leader. This implies that every correct process eventually selects \( \ell \) as a leader forever.

This algorithm performs little work in steady state, as eventually the following happens: (1) processes stop sending notifications because they do so only when they become a leader or when they are leader and receive a notification; (2) processes stop sending accusations, because eventually \( \ell \) is their only contender (other than themselves), and no process times out on \( \ell \); (3) \( \ell \) is the only process who writes to a shared register, because only a process who thinks it is the leader does so; (4) \( \ell \) does not read any shared register, because processes stop sending notifications; (5) processes \( p \neq \ell \) read only the shared register written by \( \ell \), because \( \ell \) is their only contender. Also note that each shared register in this algorithm is written only by a single process (single-writer multi-reader shared register).

Figures 3 and 4 show the detailed algorithm. In the text below, when necessary for clarity, we use subscripts on a local variable to denote its process, and possibly superscripts to denote a time (e.g., \( state_p \) is \( state \) variable of \( p \), while \( state_p^\ell \) is the value of this variable at time \( \ell \)).

Figure 4 shows the notification mechanism, which is very simple in this case (where links are reliable): the Notify\( (q) \) procedure just sends a notification message to \( q \), while the Get Notifications() procedure returns the processes from which \( p \) got a notification message since the last invocation. Figure 3 has the main pseudocode. Each process \( p \) has a \( STATE[p] \) shared register, which is a triple with \( p \)'s heartbeat counter, badness counter, and active bit. Process \( p \) has a local variable \( state_p[q] \) containing \( p \)'s local view of \( STATE[q] \). Process \( p \) executes forever in a loop. In this loop, \( p \) first picks its leader (line 9). Then \( p \) checks if it has just become the leader and, if so, notifies others (lines 10–11). Next \( p \) checks if it has just lost leadership and, if so, clears its active bit (lines 12–14). Next, if \( p \) is the leader (line 15), it increments its heartbeat counter and sets the active bit in shared memory (lines 16–18), and checks from whom it received notifications (line 19). For each such process \( q \), it adds \( q \) to

\(^4\)In the code, \( p \) checks the timer of all processes for expiration, but it is easy to see that only processes in \( p \)'s contender set have an ongoing timer at \( p \).
Shared objects:
1. \( \text{STATE}[p] \leftarrow (0, 0, \text{false}) \): register accessible by all, \( \forall p \in \Pi \)

Variables of process \( p \):
1. \( \text{state}[q] \leftarrow (0, 0, \text{false}), \forall q \in \Pi \{ \text{local} \ \text{var} \ \text{with fields} \ (\text{hh}, \text{counter}, \text{active}) \} \)
2. \( \text{hbtimer}[q] \leftarrow \eta + 1, \forall q \in \Pi \setminus \{ p \} \)
3. \( \text{notifiers} \leftarrow \emptyset \), competitors \( \leftarrow \emptyset \)
4. \( \text{contenders} \leftarrow \{ p \} \)
5. \( \text{leader} \leftarrow \perp \)

Code for each process \( p \):
1. \( \text{repeat} \)
2. \( \text{previous_leader} \leftarrow \text{leader} \)
3. \( \text{leader} \leftarrow \ell \ \text{s.t.} \)
4. \( \text{(state}(\ell), \text{counter}, \ell) = \min ((\text{state}[q], \text{counter}, q) : q \in \text{competitors}) \)
5. \( \text{if previous_leader} \neq p \ \text{and leader} = p \text{ then } \}
6. \( \text{for each } q \in \Pi \setminus \{ p \} \) \( \text{do} \)
7. \( \text{NOTIFICATIONS} \leftarrow \text{false} \)

8. \( \text{NOTIFICATIONS[q]} \leftarrow \text{false} \)
9. \( \text{for each } q \in \Pi \setminus \{ p \} \) \( \text{do} \)
10. \( \text{NOTIFICATIONS} \leftarrow \text{false} \)

11. \( \text{notifiers} \leftarrow \{ q \} \)
12. \( \text{return} \)

13. \( \text{send notification to } q \)
14. \( \text{receive new notification message from } q \)
15. \( \text{return} \)

Figure 3: Leader election algorithm.

Figure 4: Notification mechanism (reliable links).

Figure 5: Notification mechanism (fair lossy links).

contenders, starts a timer on \( q \), reads \( \text{STATE}[q] \), and notifies \( q \) that it is also a leader contender (lines 20–24). Then \( p \) checks whether it received accusations and, if so, increments its badness counter (lines 25–27).

Next, whether \( p \) is leader or not, it checks whether the timer on each process \( q \neq p \) has expired (lines 28–29). If it has, \( p \) checks whether \( q \)’s heartbeat increased since \( p \) started the timer (lines 30–33). If so, \( p \) restarts the timer (line 34). Otherwise, \( p \) removes \( q \) from its contender set; if \( q \) is active, \( p \) sends an accusation to \( q \) and increments its timeout value (lines 35–39).

THEOREM 5.1. The algorithm in Figures 3 and 4 solves the eventual leader election problem in the \( m \times m \) model where links are reliable and at least one process is timely. Eventually, no messages are sent, and the only accesses to shared memory is that the leader periodically writes a shared register and other processes periodically read this register.

The proof is in the full paper.

5.2 Algorithm for fair lossy links

We now present an algorithm that works with fair lossy links. This algorithm has an added cost: the leader not only keeps writing a shared register, but also keeps reading a shared register. This cost is necessary, as we see later (§5.4).

The algorithm is similar to the previous one, except that it uses a different notification mechanism. Instead of sending a message, a process \( p \) notifies \( q \) by setting a bit in a \( \text{NOTIFIES}[q][p] \) matrix in shared memory; \( p \) also sets a bit in a \( \text{NOTIFICATIONS}[q] \) vector; this is an optimization so that \( q \) can monitor just one bit instead of a row of the matrix. More precisely, to know whether \( p \) has any notifications, \( p \) first examines the \( \text{NOTIFICATIONS}[p] \) bit; if it is  

5The timer is a local variable with a counter that is decremented at each step of \( p \) (this is not shown in the code for clarity), until it reaches 0 (it “expires”).
clear, it need not check anything else; otherwise, it examines the row $\text{NOTIFY}_r[p] = \text{-}1$ of the matrix to find out which process sent the notification. Figure 5 has the detailed code. By combining it with in Figure 3, we obtain the leader election algorithm.

**Theorem 5.2.** The algorithm in Figures 3 and 5 solves the eventual leader election problem in the m&m model where links are fair lossy and at least one process is timely. Eventually, no messages are sent, and the only accesses to shared memory is that the leader periodically writes a shared register and periodically reads a shared register, and other processes periodically read a shared register.

### 5.3 Locality

Our leader election algorithms are also efficient in terms of the locality, where each register is local to some process. This model corresponds well to the reality of an RDMA network, in which each process and each register are on some machine. In this case, if a process $p$ is on the same machine as some register $r$, we say that $r$ is local to $p$, and that $p$ owns $r$. We say that the owner can read and write the register locally, while others read and write remotely.

Both leader election algorithms ensure that eventually the leader $\ell$ accesses registers only locally, by writing $\text{STATE}()$ (first algorithm), or by writing $\text{STATE}()$ and reading $\text{NOTIFICATIONS}()$ (second algorithm). This property further decreases the synchrony needed in practice: recall that we require only one process to be timely (the leader). By making the leader accesses local, the algorithms make it easier for the leader to be timely, since its steps involve only local accesses. On the other hand, the other processes access shared registers remotely, which is slower, but they are not required to be timely.

### 5.4 Impossibility result

In the full paper, we show that our algorithms are tight in the following sense. For systems with a timely process and asynchronous links, the leader must write shared registers forever. This result holds even if links are reliable and, a fortiori, if links are fair lossy. If, however, links are fair lossy, there is an additional requirement: either the leader writes and reads shared registers forever, or some process keeps sending messages forever.

**Theorem 5.3.** Let $A$ be any eventual leader election algorithm for the m&m model with $n \geq 2$ processes, where links are reliable and at least one process is timely. There is a run of $A$ whose leader writes shared registers infinitely often.

**Theorem 5.4.** Let $A$ be any eventual leader election algorithm for the m&m model with $n \geq 3$ processes, where links are fair lossy and at least one process is timely. Either (1) there is a run of $A$ whose leader writes and reads shared registers infinitely often, or (2) there is a run of $A$ in which some process sends messages infinitely often.

### 6 CONCLUSION

The m&m model provides some inherent advantages over the pure message-passing and shared-memory models. In this paper, we demonstrated advantages in two aspects, fault tolerance and synchrony, and we focused on two problems, consensus and leader election. There are many exciting directions for future work in this space: discovering other benefits of the m&m model, developing better algorithms, studying other problems beyond consensus and leader election, evaluating algorithms in practice, and considering more failure models. On the last point, we addressed only process crashes, but it would be interesting to consider Byzantine failures, where the ability to pass messages and share registers selectively could overcome the difficulties of dealing with Byzantine processes in shared memory. Also interesting is to consider failures of the shared memory, especially if only parts of the memory fails [2, 42].

### REFERENCES


