Large Scale Verification of MPI Programs Using Lamport Clocks with Lazy Update

Anh Vo, Ganesh Gopalakrishnan, and Robert M. Kirby
School of Computing
University of Utah
Salt Lake City, Utah 84112-9205
Email: {avo, ganesh, kirby}@cs.utah.edu*

Bronis R. de Supinski, Martin Schulz, and Greg Bronevetsky
Center for Applied Scientific Computing
Lawrence Livermore National Lab
Livermore, California 96678-2391
Email: {bronis, schulzm, bronevetsky}@llnl.gov*

Abstract—We propose a dynamic verification approach for large-scale message passing programs to locate correctness bugs caused by unforeseen nondeterministic interactions. This approach hinges on an efficient protocol to track the causality between nondeterministic message receive operations and potentially matching send operations. We show that causality tracking protocols that rely solely on logical clocks fail to capture all nuances of MPI program behavior, including the variety of ways in which nonblocking calls can complete. Our approach is hinged on formally defining the matches-before relation underlying the MPI standard, and devising lazy update logical clock based algorithms that can correctly discover all potential outcomes of nondeterministic receives in practice. We show that causality tracking protocols involving a thousand MPI processes, incurring only modest overheads in terms of communication bandwidth, latency, and memory consumption.

I. INTRODUCTION

MPI (Message Passing Interface, [7]) continues to be the most widely used API for writing parallel programs that run on large clusters. MPI allows programmers to employ many high performance communication constructs such as asynchronous (nonblocking) and synchronous communication calls, nondeterministic constructs, and persistent communication requests. Its feature-rich semantics provide performance and flexibility, but also create several debugging challenges.

Buggy MPI programs, especially under the presence of nondeterminism, are notoriously hard to debug. Bugs may lurk in nondeterministic execution paths that are not taken on a given testing platform. The key problem we solve in this paper is: how does one affordably detect missed nondeterministic choices and force executions to occur along them? To illustrate how difficult this problem becomes with MPI, we consider the program in Figure 1 in which an asynchronous nondeterministic receive posted in process $P_1$ can potentially match with messages sent by either $P_0$ and $P_2$. Under traditional testing, one may never successfully be able to force $P_2$’s message (which triggers ERROR) to match. While this option appears impossible due to its issuance after an MPI barrier, it is indeed a possible match because MPI semantics allow a nonblocking call to pend until its corresponding wait is posted. This example illustrates the need for more powerful verification techniques than ordinary random testing on a cluster in which $P_2$’s match may never manifest itself due to the absolute delays and, yet, it may show up when the code is ported to a different machine.

In this paper, we show how one can achieve coverage over all feasible nondeterministic matches by tracking causality (or, the happens-before order) instead of relying upon real-time delays. In addition, we show how the traditional notion of happens-before is not sufficient to capture causality within an MPI execution. Our key contribution is to precisely formulate the matches-before relation underlying MPI execution and exploit it to discover alternative nondeterministic matches. We show that our methods can work at large scale based on efficient runtime mechanisms that enforce feasible matches while avoiding impossible matches.

In addition to computing causal orderings correctly in the presence of nondeterminism for the wide range of MPI constructs, a practical MPI debugger must scale with the application, i.e., potentially to several thousands of MPI processes. This requirement is essential since many test inputs require a large process count to run. Downscaling is often impossible or can mask the bug.

In this paper, we detail how the causal ordering between MPI nondeterministic receive calls and MPI send calls is determined efficiently using a new protocol, the Lazy Lamport Clocks Protocol (LLCP). We do not detail how to handle other MPI call types (e.g., MPI nondeterministic probe commands), but the details are similar. Once matches-before is
correctly computed, a dynamic verifier can determine all feasible matches for a nondeterministic receive. The remainder of this paper focuses on how to determine matches-before (or causal ordering) in production-scale MPI programs.

Tracking causality for MPI is not a new idea [10], [15]. These algorithms use logical clocks to detect causality (or the lack thereof) between send and receive events. However, as detailed later, most algorithms can only handle a small subset of MPI and also do not scale well. We are not aware of any prior work that can handle the example in Figure 1, which is a fairly simple MPI communication pattern. Further, all surveyed work uses vector clocks, despite their high overhead because the much cheaper Lamport clocks are well known to not provide sufficient precision to determine causality.

We show that while our LLCP algorithm is based on Lamport clocks, it still exhibits an extremely low omission probability (none in all real benchmarks that we have studied) because it exploits two additional crucial pieces of information that are not exploited in traditional Lamport clock-based approaches: (i) the matches-before formal model of MPI, and (ii) observed matches during an initial MPI run to infer causalities within alternate runs. Our LLCP algorithm can also be implemented using vector clocks, should omissions become a concern on real examples; in that case, it still offers significant advantages over standard vector-clock-based algorithms that, by tracking all events, incur a high overhead.

In contrast, our LLCP algorithms track only relevant events which, in our case are nondeterministic receives. This further improves the scalability of our approach. While we express our ideas in terms of MPI, they apply to any computational model that satisfies the conditions that we describe in Section III-A.

To summarize, we contribute an algorithm based on Lamport clocks that uses lazy updating rules to model the behavior of MPI nonblocking calls properly and to enable detection of matches that prior work misses. We also provide a vector clocks extension of the LLCP algorithm (lazy vector clocks) and evaluate the performance of the two protocols. Finally, we formally characterize both algorithms.

II. BACKGROUND

The usage of logical clocks to track causality in distributed systems has been extensively studied and enhanced. Most of approaches are either based on Lamport clocks [11] or vector clocks [5], [12].

Lamport clocks: The Lamport clock mechanism is an inexpensive yet effective mechanism to capture the total order between events in distributed systems. We briefly summarize the algorithm. Each process \( P_i \) maintains a counter \( LC_i \) initialized to 0. When event \( e \) occurs, \( P_i \) increments \( LC_i \) and assigns the new value, \( e.LC \), to \( e \). If \( e \) is a send event, \( P_i \) attaches \( e.LC \) to the message. On the other hand, if \( e \) is a receive event, \( P_i \) sets \( LC_i \) to a value greater than or equal to its present value and greater than that which it receives.

Let \( e^i_n \) be the \( n^{th} \) event that occurs in process \( P_i \), \( send(P_i, msg) \) be the event corresponding to sending message \( msg \) to \( P_i \), and \( recv(P_i, msg) \) be the event corresponding to the reception of \( msg \) from \( P_i \). We now define the happens-before (hb) relationship over all events.

\[
e^i_n \xrightarrow{hb} e^j_m \text{ when:}
\]

- \( i = j \land n > m \) (local event order)
- \( e^i_n = send(P_j, msg) \land e^j_m = recv(P_i, msg) \)
- \( \exists e^r_k : e^i_n \xrightarrow{hb} e^r_k \land e^r_k \xrightarrow{hb} e^m_j \) (transitive order)

Given the Lamport clock algorithm described above, we can infer the clock condition: for any two events \( e_1 \) and \( e_2 \), if \( e_1 \xrightarrow{hb} e_2 \) then \( e_1.LC < e_2.LC \). Henceforth, we use \( \xrightarrow{hb} \) instead of \( \xrightarrow{hb} \) when the context is clear.

Vector clocks: The timestamping algorithm proposed by Lamport has one major limitation: it does not capture the partial order between events. To address this problem, Fidge [5] and Mattern [12] independently propose that the single counter on a port has one major limitation: it does not capture the partial order between events. Our LLCP algorithm tracks only the matches that prior work misses. We also provide vector clock compression, exist, the worst case behavior of vector algorithms is still unscalable.

Piggybacking: All causality tracking protocols discussed above rely on the ability of the system to transmit the clocks together with the intended user data on all outgoing messages. We refer to the transmission of the clocks as piggybacking and the extra data (i.e., the clocks) as piggyback data. We can transmit piggyback data in many different ways such as packing the piggyback data together with the message data, or sending the piggyback data as a separate message [16]. The details of those approaches are beyond the scope of this paper and are thus omitted.

III. MATCHES-BEFORE RELATIONSHIP IN MPI

A. Our Computational Model

A message passing program consists of sequential processes \( P_1, P_2, \ldots, P_n \) communicating by exchanging messages through some communication channels. The channels are assumed to be reliable and support the following operations:

- \( send(\text{dest}, T) \) - send a message with tag \( T \) to process \( \text{dest} \). This operation has similar asynchronous semantics to \( \text{MPI\_Send} \): it can complete before a matching receive has been posted.
We briefly discuss how one could apply the traditional vector clock algorithm to the example in Figure 2 to conclude that the first wildcard receive in \( P_0 \) can match either send from \( P_1 \) and \( P_2 \) (and also why Lamport clocks fail to do so).

We assume that the first receive from \( P_0 \) matches the send from \( P_1 \) and that the second receive from \( P_0 \) matches the send from \( P_2 \). We must determine whether the vector clock algorithm could determine that the first receive from \( P_0 \) could have received the message that \( P_2 \) sent. Using the clock updating rules from the vector clock algorithm, the vector timestamp of \( P_0 \)’s first receive is \([1,1,0]\) while that of the send from \( P_2 \) is \([0,2,2]\). Clearly, the send and the receive are concurrent and thus, the send could match the receive.

In contrast, if we apply the original Lamport clock algorithm to this example, \( P_0 \)’s first receive event would have a clock of 1 while the send from \( P_2 \) would have a clock of 3. Thus, the algorithm cannot determine that the two events have no causal relationship and cannot safely identify the send from \( P_2 \) as a potential match to the first receive from \( P_0 \). One can observe that the communication between \( P_1 \) and \( P_2 \) in this example has no effect on \( P_0 \), yet the matching causes a clock increase that disables \( P_0 \)’s ability to determine the causality between the first wildcard receive and \( P_2 \)’s send.

Now consider the example shown in Figure 1. Assuming the \texttt{irecv} by \( P_1 \), denoted as \( r \), matches the \texttt{isend} from \( P_0 \), we apply the vector clock algorithm to determine if the \texttt{isend} from \( P_2 \), denoted as \( s \), can match \( r \). By using the vector clock updating rules and considering \texttt{barrier} as a synchronization event in which all processes synchronize their clocks to the global maximum, the clocks for \( r \) and \( s \) are \([1,0,0]\), and \([1,0,1]\), respectively. Thus, \( r \overset{hbvc}{\rightarrow} s \) and the algorithm fails to recognize \( s \) as a potential match to \( r \).

Clearly, the notion of \textit{happening} and the corresponding \textit{happens-before} relationship are insufficient in capturing all behaviors of MPI programs. We need a new model that completely captures the ordering of all events within an MPI program execution.

C. The Matches-before Relationship

We first consider all possible states of an MPI operation \( op \) after a process invokes \( op \):

- \texttt{issued} - \( op \) attains this state immediately after the process invokes it. All MPI calls are issued in program order.
- \texttt{matched} - We define this state in Definition 3.1.
- \texttt{returned} - \( op \) reaches this state when the process finishes executing the code of \( op \).
- \texttt{completed} - \( op \) reaches this state when \( op \) no longer has any visible effects on the local program state. All blocking calls reach this state immediately after they return while nonblocking calls reach this state after their corresponding waits return.

Of these states, only the \texttt{issued} and \texttt{matched} states have significant roles in our algorithms; nonetheless, we included all possible states for completeness. The \texttt{matched} state is central to our protocols. We describe it in detail below.

\begin{definition}
A event \( e \) in an MPI execution attains the \texttt{matched} state if it satisfies one of these conditions:
\begin{itemize}
  \item \( e \) is the send event of message \( m \) and the destination process has started to receive \( m \) through some event \( e' \). \( e \) is said to have matched with \( e' \). The receiving process is considered to have started to receive \( m \) when we can (semantically) determine from which send event it will receive data. The timing of the completion of the receiving process is up to the MPI runtime and is not relevant to this discussion. In this case, we consider \( e \) and \( e' \) to be in a send-receive match-set.
  \item \( e \) is a receive event that marks the start of reception.
  \item \( e \) is a \texttt{wait} \((h)\) call and the pending receive request associated with \( h \) has been matched. For an \texttt{isend}, the \texttt{wait} can attain the \texttt{matched} state upon completion while the \texttt{isend} still has not matched (i.e., it is buffered by the MPI runtime). A \texttt{matched} \texttt{wait} is the only element in its match-set (a \texttt{wait match-set}).
\end{itemize}
\end{definition}
e is a barrier and all processes have reached their associated barrier. All participating barriers belong to the same match-set (a barrier match-set). We consider e to have matched $e$ if they are in the same barrier match-set.

While determination of the matching point of $\text{recv}$ and barrier calls is straightforward, the situation is more complex with nonblocking calls. The assumption that all nonblocking calls attain the matched state exactly at their corresponding wait call returns, which could be much earlier than the completion of its corresponding wait. This interval arises from the nonovertaking rule of the MPI standard, which we describe later in this section.

Let $E$ be the set of events produced in an execution $P = \langle E, \text{mb} \rangle$, where each $e \in E$ is a match event and $\text{mb}$ is the matches-before relation over $E$ defined as follows: consider two distinct events, $e_1, e_2 \in E$, $e_1 \xrightarrow{\text{mb}} e_2$ if and only if one of the following conditions holds:

- **C1.** $e_1$ is a blocking call (send, recv, wait, barrier) and $e_2$ is an event in the same process that is issued after $e_1$.
- **C2.** $e_2$ is the corresponding wait of nonblocking call $e_1$.
- **C3.** $e_1$ and $e_2$ are send events from the same process $i$ with the same tag, targeting the same process $j$ and $e_1$ is issued before $e_2$. This is the nonovertaking rule of MPI for sends. The sends can be blocking or nonblocking.
- **C4.** $e_1$ and $e_2$ are receive events from the same process $i$ with the same tag, either $e_1$ is a wildcard receive or both are receiving from the same process $j$, and $e_1$ is issued before $e_2$. This is the nonovertaking rule of MPI for receives. The receives can be blocking or nonblocking.
- **C5.** $e_1$ and $e_2$ are from two different processes and $\exists e_3, e_4 \in E$ in the same match-set and $e_3$ is not a receive event (i.e., $e_3$ is a send, isend, or barrier) such that $e_1 \xrightarrow{\text{mb}} e_3$, $e_4 \xrightarrow{\text{mb}} e_2$. Figure 4 illustrates this transitivity condition. The two shaded areas in the figure show a send-receive match-set and a barrier match-set while the dashed arrows show the matches-before relationship between events in the same processes. Condition C5 implies that $e_1 \xrightarrow{\text{mb}} e_2$ and $e_2 \xrightarrow{\text{mb}} e_3$.
- **C6.** $\exists e_3 \in E$ such that $e_1 \xrightarrow{\text{mb}} e_3$ and $e_3 \xrightarrow{\text{mb}} e_2$ (transitive order). In Figure 4, condition C6 combines with C5 to imply that $e_1 \xrightarrow{\text{mb}} e_3$.

**Corollary 3.1:** If $e_1$ and $e_2$ are two events in the same match-set, neither $e_1 \xrightarrow{\text{mb}} e_2$ nor $e_2 \xrightarrow{\text{mb}} e_1$ holds.

**Corollary 3.2:** If $e_1$ and $e_2$ are two events in the same process and $e_1$ is issued before $e_2$, then $e_2 \xrightarrow{\text{mb}} e_1$ is false. This is directly based on the definition of matches-before.

In addition to the $\text{mb}$ relationship for two events, we also define the $\text{mb}$ relationship between $X$ and $Y$ where either $X, Y$, or both are match-sets: $X \xrightarrow{\text{mb}} Y$ if and only if one of the following conditions holds:

- **C7.** $X$ is an event $e_1$, $Y$ is a send-receive match-set, $e_2$ is the send event in $Y$, and $e_1 \xrightarrow{\text{mb}} e_2$.
- **C8.** $X$ is an event $e_1$, $Y$ is a barrier match-set or a wait match-set and for each event $e_2 \in Y$, $e_1 \xrightarrow{\text{mb}} e_2$.
- **C9.** $X$ is a send-receive match-set, $e_1$ is the receive event in $X$, $Y$ is an event $e_2$, and $e_1 \xrightarrow{\text{mb}} e_2$.
- **C10.** $X$ is a barrier match-set or a wait match-set and $Y$ is an event $e_2$, and $\exists e_1 \in X$ such that $e_1 \xrightarrow{\text{mb}} e_2$.
- **C11.** $X$ and $Y$ are match-sets and $\exists e_1 \in X$ such that $e_1 \xrightarrow{\text{mb}} Y$.

**Concurrent events:** We consider $e_1$ and $e_2$ concurrent if they are not ordered by $\xrightarrow{\text{mb}}$. Let $e_1 \xrightarrow{\text{mb}} e_2$ denote that $e_1$ is not required to match before $e_2$; then $e_1$ and $e_2$ are concurrent if and only if $e_1 \xrightarrow{\text{mb}} e_2 \land e_2 \xrightarrow{\text{mb}} e_1$.

**IV. Lazy Update Algorithms**

Since we have adapted the matches-before relationship from the traditional Lamport happens-before, we need new clock updating algorithms that capture the causality information between events in an MPI execution based on $\xrightarrow{\text{mb}}$. To this end, we propose two algorithms: the Lazy Lamport Clocks Protocol (LLCP) and the Lazy Vector Clocks Protocol (LVCP). These are our design goals:

- **scalable** - the protocol should scale to large process counts;
- **sound** - the protocol should not indicate that impossible matches are potential matches;
- **complete** - the protocol should find all potential matches.

We require a sound protocol while the other goals are desirable. An unsound protocol can cause a deadlock in an otherwise deadlock-free MPI program. Both LLCP and LVCP are sound. Many MPI applications today require large scale runs in order to test important input sets due to memory size and other limits. Also, many bugs, including nondeterminism
related bugs, are only manifest at large scales. Thus, a protocol must scale to support finding many important errors. LLCP is scalable compared to LVCP as our experimental results demonstrate. However LVCP is complete while LLCP is not. The design of a complete and scalable protocol represents a significant challenge. We require a scalable protocol that maintains completeness for common usage. In our testing with real MPI programs, LLCP is complete, that is we did not discover any extra matches when we ran the same program under LVCP. When completeness is absolutely required, one should use LVCP at the cost of scalability.

A. The Lazy Lamport Clocks Protocol

Relevant events: Clearly, not all events in an MPI execution are relevant to our purposes of tracking causality. However, one cannot naively track only relevant events and ignore irrelevant events because these events can indirectly introduce causality between relevant events. In our protocol, we consider wildcard receives (whether blocking or nonblocking) as relevant events. While we do not consider sends as relevant, our algorithm still allows us to track causality between sends and relevant events.

Algorithm overview: LLCP maintains the matches-before relationship between events by maintaining a clock in each process and associates each event with a clock value so we can order these events according to when they attain the matched state. Since the matches-before relationship describes the ordering for events within and across processes, the algorithm needs must provide such coverage. More specifically, given a wildcard receive \( r \) from process \( P_i \) and a send \( s \) targeting \( P_i \) that did not match \( r \), the protocol should allow us to determine if \( r \) and \( s \) have any matches-before relationship. If, for example, the successful completion of \( r \) triggers the issue of \( s \), then \( s \) could never match \( r \). The intuitive way to achieve this soundness is for the protocol to maintain the clock such that if \( r \) triggers the issue of some event \( e \) then the clock of \( r \) must be smaller than the clock of \( e \). Basically this condition requires all outgoing messages after \( r \) from \( P_i \) to carry some clock value (as piggyback data) higher than \( r \).

The challenge of the protocol lies in the handling of nonblocking wildcard receives. As explained earlier in the example in Figure 1, a nonblocking wildcard receive can potentially be pending (not yet have reached the matched state) until its corresponding wait is posted. However, we have also shown in Figure 3 that the receive could also attain the matched state due to the nonovertaking semantics (which could be earlier than the posting of the wait). The protocol must precisely determine the status of the receive to avoid sending the wrong piggyback data, which could lead to incorrect matching decisions. To achieve this precision, we use the following set of clock updating rules:

- **R1.** Each process \( P_i \) keeps a clock \( LC_i \), initialized to 0.
- **R2.** When a nonblocking wildcard receive event \( e \) occurs, assign \( LC_i \) to \( e.LC \) and add \( e \) to the set of pending receives: \( \text{Pending} \leftarrow \text{Pending} \cup \{e\} \).
- **R3.** When \( P_i \) sends a message \( m \) to \( P_j \), it attaches \( LC_i \) (as piggyback data) to \( m \) (denoted \( m.LC \)).
- **R4.** When \( P_i \) completes a receive event \( r \) (either forced by a blocking receive or at the wait of a nonblocking receive as in Figure 3), it first constructs the ordered set \( \text{CompleteNow} \) as follows: \( \text{CompleteNow} = \{ e \mid e \in \text{Pending} \land e \text{mb } r \} \). The set \( \text{CompleteNow} \) is ordered by the event’s clock, where \( \text{CompleteNow}[i] \) denotes the \( i \)-th item of the set and \( \text{CompleteNow}[j] \) denotes the total items in the set. Intuitively, these pending nonblocking receives have matched before \( r \) due to the nonovertaking rule. Since they have all matched, we must update their clocks. The ordering of the events in \( \text{CompleteNow} \) is important since all receives in \( \text{CompleteNow} \) are also ordered by the nonovertaking semantics. We can update the clocks using the following loop:

\[
\text{for } i = 1 \text{ TO } |\text{CompleteNow}| \text{ do}
\]

\[
\text{CompleteNow}[i].LC = LC_i
\]

\[
LC_i \leftarrow LC_i + 1
\]

\text{end for}

\( \text{Pending} \leftarrow \text{Pending} \setminus \text{CompleteNow} \)

- **R5.** At barrier events, all clocks are synchronized to the global maximum of the individual clocks.

Lazy Update: Rules R2 and R4 form the lazy basis of the protocol in the sense that a nonblocking wildcard receive \( r \) has a temporary clock value when it initially occurs and obtains its final clock value when it finishes (either by its corresponding wait or by another receive \( r' \) for which \( r \text{mb } r' \)).

\textbf{Lemma 4.1:} If \( e_1 \text{mb } e_2 \text{ then } e_1.LC \leq e_2.LC \)

\textbf{Proof:} We first consider the case when \( e_1 \) and \( e_2 \) are from the same process. Based on our definition of matches-before, event \( e_2 \) will always occur after event \( e_1 \). Since our algorithm never decreases the clock, it follows that \( e_1.LC \leq e_2.LC \)

Now assume \( e_1 \) and \( e_2 \) are events from different processes. From the definition of matches-before, events \( e_3 \) and \( e_4 \) exist such that \( e_1 \text{mb } e_3, e_3 \text{mb } e_2, e_2 \text{ and } e_4 \text{ are in a match-set, and } e_3 \text{ is an isend, a send, or a barrier. We recursively apply this process to } (e_1, e_3) \text{ and } (e_4, e_2) \text{ to construct the set } S = s_1, s_2, ..., s_n \text{ in which } s_1 = e_1, s_n = e_2, \text{ and other elements are events or match-sets that satisfy } s_i \text{mb } s_{i+1} \).

Also, \( S \) must satisfy the following rule: for any pair of adjacent elements \( (s_i, s_{i+1}) \), no event \( e \) exists such that \( s_i \text{mb } e \) and \( e \text{mb } s_{i+1} \). The construction of \( S \) is possible based on our
definition of $\text{mb}$. Intuitively, $S$ represents all events between $e_1$ and $e_2$ along the $\text{mb}$ event chain.

Now consider any pair of events $(s_i, s_{i+1})$. They must both be events from the same process, in which case $s_i.LC \leq s_{i+1}.LC$, or at least one is a match-set, in which case our piggyback ensures that $s_i.LC \leq s_{i+1}.LC$. Thus, the set $S$ has the property that $s_1.LC \leq s_2.LC \leq \ldots \leq s_n.LC$, so $e_1.LC \leq e_2.LC$.

**Lemma 4.2:** Assuming $r$ is either a blocking receive or a nonblocking receive that is not pending, if $r \xrightarrow{\text{mb}} e$ then $r.LC < e.LC$.

**Proof:** If $e$ is an event in the same process as $r$ then rules $R2$ and $R4$ ensure that $r.LC < e.LC$. If $e$ is not in the same process as $r$ then based on the definition of $\text{mb}$, an event $f$ from the same process as $r$ must exist such that $r \xrightarrow{\text{mb}} f \land f \xrightarrow{\text{mb}} e$, which means $r.LC < f.LC$ and by Lemma 4.1, $f.LC \leq e.LC$. Thus, $r.LC < e.LC$. ■

We now introduce the concept of late messages, which is essential for the protocol to determine if an incoming send can match an earlier wildcard receive. One can think of late messages as in-flight wildcard messages in the sense that these messages have already been issued when a receive reaches the matched state. Consider the MPI program shown in Figure 5. The first wildcard receive of $P_1$ matched with the $\text{send}$ from $P_0$ while the second wildcard receive matches the $\text{send}$ from $P_2$. The clock value associated with each event according to our protocol is displayed in the square bracket. The shaded area represents the set of all events that are triggered by $r$ (i.e., for all events $e$ in that shaded area, $r \xrightarrow{\text{mb}} e$). The message from $P_2$ was not triggered by the matching of the first wildcard receive in $P_1$, despite being received within the shaded area. We call the message from $P_2$ a late message. At the reception of a late message, the protocol determines if it is a potential match for receives that matched earlier (in this figure, the late message from $P_2$ is a potential match).

**Definition 4.1:** A message $m$ originating from process $m.src$ with timestamp $m.LC$ is considered late with respect to a wildcard receive event $r$ (which earlier did not match with $m.src$) iff $m.LC \leq r.LC$. If message $m$ is late with respect to $r$, it is denoted as $\text{late}(m, r)$.

We can now present our match detection rule.

*Theorem 4.3:* An incoming send $s$ carrying message $m$ with tag $\tau$ that is received by event $r'$ in process $P_1$, is a potential match to a wildcard receive event $r$ with tag $\tau$ issued before $r'$ if $(m.LC < r'.LC \land \text{late}(m, r))$.

**Proof:** In order to prove that $s$ is a potential match of $r$, we prove that $s$ and $r$ are concurrent, that is, $r \xrightarrow{s} s \land s \rightarrow r$. First, note that $r \xrightarrow{\text{mb}} r'$, which means $r$ cannot be a pending event (due to rule $R4$). We also have $r.LC \geq s.LC$ since $s$ is a late message. Using the contrapositive of Lemma 4.2, we infer that $r \xrightarrow{\text{mb}} s$. Also, $s \xrightarrow{\text{mb}} r$ because if $s \xrightarrow{\text{mb}} r$ then $s \xrightarrow{\text{mb}} r'$ due to the transitive order rule of matches-before. This conclusion violates Corollary 3.1: $\text{mb}$ does not order two events in the same match-set. ■

We now revisit the example in Figure 1. Using the LLCP clock update rules, $P_1$ has a clock of 0 when issues its $\text{irecv}$, which it adds to $\text{Pending}$. The barrier calls synchronize all clocks to the global maximum, which is 0. At the $\text{recv}$ call, $P_1$ applies rule $R4$ and constructs the CompleteNow set consisting of $\text{irecv}$. Upon the completion of this step, the $\text{irecv}$ has a clock of 0 and the $\text{recv}$ has a clock of 1. Assuming that the $\text{isend}$ from $P_0$ matches the $\text{irecv}$, which means the $\text{recv}$ call matches the $\text{isend}$ from $P_2$. The message from $P_2$ carries a piggyback data of 0 so it is flagged as a late message with respect to the $\text{irecv}$ and is detected as a potential match (per Theorem 4.3). Figure 6 shows the clock values of this execution.

LLCP can miss potential matches. Consider the example in Figure 7 for which manual inspection easily finds that the send from $P_0$ is a potential match of the wildcard receive from $P_2$ (assuming the $P_2$’s $\text{recv}$ matches with $P_1$’s $\text{send}$). However, LLCP would not detect this potential match since at the time of receiving $P_0$’s $\text{send}$, the clock of $P_0$’s $\text{send}$ is 1, which is the same as the clock of $P_2$’s $\text{recv}(0)$, so it does not satisfy the condition of Theorem 4.3.

This issue again reflects the disadvantage of Lamport clocks given multiple concurrent sources of nondeterminism. In general, omissions may happen with multiple sources of nondeterminism and communication between processes with clocks that are out of synchronization. Fortunately, this situation rarely happens in practice because most MPI programs have well-established communication patterns that do not have cross communications between groups of processes that generate
relevant events before clock synchronization occurs. We will present our extension of LLCP to vector clocks that addresses MPI programs with subtle communication patterns for which LLCP may not determine all potential matches.

Handling synchronous sends: The MPI standard supports synchronous point-to-point communication with MPI_Ssend and MPI_Issend. In order to handle these synchronous calls, we extend the definitions of $mb$ and LLCP. Specifically, we modify condition C9 to condition C9'

- C9'. $e_1$ is a synchronous send in match-set $X$, $e_2$ is the corresponding receive in $X$, and $e_3$ is some event. $X \xrightarrow{mb} e_3$ if and only if $e_1 \xrightarrow{mb} e_\land e_2 \xrightarrow{mb} e_3$.

Typically when a receive operation and a synchronous send operation match, the MPI implementation sends an acknowledgment from the receiver to the sender. LLCP handles synchronous sends by modifying the MPI runtime to add piggyback data to this acknowledgment. Since synchronous calls are rarely used, we omit the implementation details.

B. The Lazy Vector Clocks Protocol

We extend LLCP to LVCP, which uses vector clocks. LVCP has similar clock updating rules that we modify to apply to vector clocks (e.g., instead of incrementing a single clock, $P_i$ now increments $VC[i]$). We now prove the updated lemmas and theorems that are based on LVCP.

Lemma 4.4: If $r$ is a blocking receive or a nonblocking receive not pending in $P_i$, then $r \xrightarrow{mb} e \iff r.VC[i] < e.VC[i]$.

Proof: We omit the proof for $r \xrightarrow{mb} e \Rightarrow r.VC[i] < e.VC[i]$, which is similar to the LLCP case. We now prove the converse. Observe that in LVCP as in all vector clock protocols, the only process that increments $VC[i]$ is $P_i$. Thus, $e$ is an event that occurs in $P_i$ after $r$ completes (which is the point at which $VC[i]$ becomes greater than $r.VC[i]$) or it is an event in another process that receives the piggybacked $VC[i]$ from $P_i$ (directly or indirectly). If $e$ is an event that occurs in $P_i$ after $r$ completes and $r$ is a blocking receive then $r \xrightarrow{mb} e$ due to the definition of $mb$. Alternatively, if $r$ is nonblocking receive, $P_i$ only increases its clock (i.e., $VC[i]$) under one of the following conditions:

- A wait for a nonblocking receive $r'$ that satisfies $r \xrightarrow{mb} r'$ is posted and $r$ is still pending before $r'$.

- A wait for a blocking receive $r'$ that satisfies $r \xrightarrow{mb} r'$ is posted and $r$ is still pending before $r'$.

- A wait for a nonblocking receive $r'$ that satisfies $r \xrightarrow{mb} r'$ is posted and $r$ is still pending before $r'$.

In all of these scenarios, we need a blocking operation $b$ such that $r \xrightarrow{mb} b$ in order to increase the clock of $P_i$. If $e.VC[i] \geq r.VC[i]$, $e$ must occur after $b$. Thus, $b \xrightarrow{mb} e$ and by the transitive order rule, $r \xrightarrow{mb} e$.

Using the updated definition of late message (Definition 4.1) where $m.LC \leq r.LC$ is replaced by $m.VC[i] \leq r.VC[i]$, we now prove the LVCP matching theorem, which states that all potential matches are recognized.

Theorem 4.5: An incoming send $s$ carrying message $m$ with tag $\tau$ that is received by event $r'$ in process $P_i$ is a potential match to a wildcard receive event $r$ with tag $\tau$ issued before $r'$ if and only if $m.VC[i] < r'.V.C[i] \land \text{late}(m, r)$.

Proof: We omit the proof for the if part, which is similar to the proof of Theorem 4.3. We now prove the converse, which can be stated as: if $r \xrightarrow{mb} s \land s \xrightarrow{mb} r$ then $m.VC[i] < r'.V.C[i] \land \text{late}(m, r)$. First we notice that $r$ due to the nonovertaking rule, $r \xrightarrow{mb} r'$, which gives us $r.VC[i] < r'.V.C[i]$ according to Lemma 4.4. We apply Lemma 4.4 to $r \xrightarrow{mb} s$, to obtain $m.VC[i] < r.VC[i]$, which means $m.VC[i] < r'.V.C[i] \land \text{late}(m, r)$ (note that $m.VC[i]$ is the same as $s.VC[i]$ since $m.VC[i]$ is the piggybacked value attached due to $s$).

V. Experimental Results

We implement LLCP and LVCP as separate P$^N$MPI [17] modules and integrate them in DAMPI, which is our distributed analyzer of MPI programs. DAMPI analyzes the program for freedom of deadlocks, resource leaks, and assertion errors. DAMPI provides the unique ability to guarantee coverage over the space of nondeterminism through replay. When either protocol (LLCP or LVCP) detects alternative matches to nondeterministic receives, DAMPI records the matches so that a scheduler can analyze the matches offline to generate a decision database that the processes parse during replay. The decision database contains the information that the processes need to replay themselves correctly up until the branching point (i.e., the wildcard receive that should match a different source during the particular replay) and rewrite the MPI_ANY_SOURCE field into the specified value. The replay mechanism of the scheduler is beyond the scope of this paper. More details on DAMPI in general and its scheduler in particular are available in a related dissertation [22].

Experimental Setup: We run our benchmarks on the Atlas cluster at Lawrence Livermore National Laboratory. this Linux-based cluster has 1152 compute nodes, each with 8 cores and 16GB of memory. All experiments were compiled and run under MVAPICH2-1.5 [1].

We first report on the latency and bandwidth impact of the two protocols. For latency testing, we use the OSU multipair latency benchmark [1] and report the latency testing result for 4-byte messages as the number of processes in the job increases. These small messages represent the class of messages for which latency impact is most significant. On
the other hand, for bandwidth testing, we use a simple ping-pong test between two processes in the system, while others sit idle. While typical bandwidth tests report the bandwidth as the message size grows, such tests do not consider the number of processes in the system and thus do not capture the scalability of the protocols. Thus, we report the message sizes at which the system achieves half of the peak bandwidth for the bandwidth testing (the R/2 value).

Figures 8 and 9 summarize our latency and bandwidth testing results. Higher bars indicate worse performance since they correspond to longer latencies and to larger message sizes to achieve the same bandwidth. The results show that both protocols have manageable latency at low process counts but the impact of vector clocks becomes more significant as the system scales up while LLCP maintains stable latency penalties across the entire range. At 1024 processes, the latency of messages under LLCP increases by only 10% compared to the original, uninstrumented messaging while it increases 240% under LVCP. Both protocols achieve essentially the same peak bandwidth as uninstrumented messaging but reduce the bandwidth achieved at intermediate message sizes. Importantly, this impact is more pronounced with LVCP and increases with increasing process count, while the impact of LLCP is independent of the process count.

Latency and bandwidth do not always translate to overhead since programs typically do not exchange messages 100% of the time. Therefore, we evaluate the performance of three scientific MPI applications: ParMETIS, a parallel hypergraph partitioning library [9]; AMG2006, an algebraic multi-grid solver for unstructured mesh available from the ASC Sequoia benchmark [2]; and SMG2000, a semicoarsening multi-grid solver from the ASCII Purple benchmark [3]. These applications are designed to run at large scales so we vary the
process count from 64 to 1024 (8 processes per node). We run ParMETIS using the supplied testing parameters. We run both AMG2006 and SMG2000 with a 6x6x6 grid as input. We report the average time for five runs. ParMETIS and SMG2000 do not have any nondeterministic receives and thus the results reflect the overhead of checking for common MPI errors (e.g., resource leaks). AMG2006 has wildcard receives and wildcard probes (the protocols can handle errors (e.g., first-in-first-out channels [20]). Nonetheless, the efficiency of these schemes are highly dependent on the communication pattern and all require more local memory storage for bookkeeping than traditional vector clocks. As supercomputers employ more and more cores, the amount of physical memory per core is decreasing so this additional overhead could prevent the tools from running. Further, all compressed vector clocks schemes send a variable amount of piggyback information, including at collective calls. Since most MPI implementations do not yet support native piggybacking, implementing a variable piggybacking layer forces the developer to use explicit buffer packing, which greatly increases performance overhead (DAMPI currently uses datatype piggybacking, which has lower overhead compared to explicit buffer packing). Also, one would need to break collective calls into pair-wise calls to support variable message sizes and thus forfeit any MPI collective optimization or to send the entire vector clock for collectives.

### VI. Related Work

**Schedule control** techniques are routinely used to analyze thread-based concurrent programs. These techniques range from sophisticated dynamic partial order methods [6], [25] to schedule bounding techniques [4], [14] to randomized search that relies on bug depth [13]. Unfortunately, none of these methods can apply directly to MPI programs [21], [24] because message passing (and MPI in particular) programs have different methods of interactions compared to shared memory programs as indicated in our discussion of MPI's matches-before model.

While some works have attempted to use logical clocks to track nondeterminism such as MPIRace-Check [15] and NOPE [10], most rely on the traditional vector clock algorithm and, thus, do not capture the full semantics of MPI. We have run MPIRace-Check on the example in Figure 1 and it failed to report the match (the source code for NOPE was not available at the time of our experiments).

**ISP** [21], [24] is a dynamic scheduling control algorithm that provides formal coverage guarantees. However, we showed that ISP's centralized schedule generation and control algorithms do not scale well [23].

**DAMPI** [23] is our dynamic verifier for MPI programs that uses an *eager* update mechanism to maintain causality between events. We earlier showed that eager clock updates can lead to high omission rates (in particular, the 'crooked barrier' example was shown to be outside its scope). This paper introduces the notion of *matches-before* which provides the basis for lazy protocols. We formally show that lazy protocols can correctly capture the full semantics of MPI with modest performance overhead.
MPI-SPIN is a model checker for MPI that can help to verify the concurrency and synchronization skeleton of MPI program models formally [18], [19]. However, this approach requires the user to build the model for the MPI program manually, which is not practical for large scale MPI programs.

VII. CONCLUSIONS

The MPI standard offers a rich set of features such as non-blocking primitives and nondeterministic constructs that help developers write high performance applications. These features, however, complicate the task of large-scale debugging, especially over the space of nondeterminism, which requires causality tracking. Traditional causality tracking algorithms, such as Lamport clocks and vector clocks, are usually not sufficient to handle its complex semantics. In this paper we propose the lazy Lamport clocks protocol and its vector clock variant that can correctly track nondeterminism in MPI programs. The lazy update scheme correctly models nonblocking MPI calls and thus can track nondeterminism more precisely. A scheduler can then use the traced information to perform replays to explore different executions of the program. We have integrated our proposed approach into DAMPI, thus enabling the ability to track and to replay nondeterministic operations to detect common MPI usage errors. We have tested DAMPI with large and practical MPI programs on up to a thousand processes. Our experiments show that the lazy Lamport clocks protocol is scalable and provides complete coverage in practice.

REFERENCES