The Minimal Dependency Relation for Causal Event Ordering in Distributed Computing

Saúl Eduardo Pomares Hernández

The Minimal Dependency Relation for Causal Event Ordering in Distributed Computing

S.E. Pomares Hernandez[a,b,c]

[a] *Computer Science Department, Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro 1, C.P. 72840, Tonantzintla, Puebla, Mexico*
[b] *CNRS, LAAS, 7 avenue du Colonel Roche, F-31400 Toulouse, France*
[c] *Univ de Toulouse, LAAS, F-31400 Toulouse, France*

Abstract

Several algorithms of different domains in distributed systems are designed over the principle of the *Happened-Before Relation* (HBR). One common aspect among them is that they intend to be efficient in their implementation by identifying and ensuring the necessary and sufficient dependency constraints. In this pursuit, some previous works talk about the use of a transitive reduction of the causality. However, none of these works formally prove in a broad manner that such transitive reduction is the minimal expression of the HBR. In this paper, a formal study of the minimal binary relation (transitive reduction) of the HBR is presented, which is called the *Immediate Dependency Relation* (IDR). The study shows that since the transitive closure of the HBR is antisymmetric and finite, it implies that the IDR is unique. This is important because it means that all of the works that deal with a minimal expression of the HBR discuss the same minimal binary relation. In addition, an extension to the IDR to identify causal immediate dependencies only among a subset of relevant events is presented. Finally, as case of study, the extension of the IDR is applied to the causal delivery of messages.

Keywords: Happened-Before Relation, Event Ordering, Partial Ordering, Distributed Systems

1. Introduction

The *Happened-Before Relation* (HBR) introduced by Lamport [5], denoted by "$\rightarrow$", without using global references establishes the conditions to determine for any pair of single events a, b in a system if the event a causally occurs before the event b (denoted by $a \rightarrow b$). Several solutions in different domains are designed over this principle. For example, the HBR was applied to ensure temporal and causal dependencies among heterogeneous data in multimedia distributed

Email address: spomares@inaoep.mx (S.E. Pomares Hernandez)

systems such as telehealth systems [9]. One common aspect among works based on the HBR is that most of them intend to be efficient in their implementation by identifying and ensuring the necessary and sufficient dependency constraints among events. In this pursuit, the present paper analyzes the *minimal binary relation* (transitive reduction) of the HBR that is called the *Immediate Dependency Relation* (IDR). The IDR, denoted in this paper by "$\downarrow$", identifies the smallest set of causally related pair of events in a given distributed computation $\hat{E} = (E, \rightarrow)$, such that for every causal path between a pair of events established with the HBR, there exists a causal path between those events established by the IDR. This property means in graph theory that the $(E, \rightarrow)$ and the $(E, \downarrow) \subset \hat{E}$ have the same reachability.

Some previous works for a particular domain deal with a transitive reduction of the HBR; nevertheless, none of these works formally prove in a general way that such transitive reduction is the minimal expression of the HBR. Some of the most important works are: [2, 4] in causality tracking for relevant events, [7] for context graphs, [11] and [8] for multicast and group communication, respectively, and [10] for a consistent and compact representation of a distributed system. As far as I know, the first work that indirectly talked about the transitive reduction of the causality by considering only immediate causal predecessors was the work presented by Peterson in [7].

In this paper, an abstract and general study of the IDR with the objective of being independent of a particular domain is presented. The IDR is proven to be the transitive reduction of the HBR. In particular, it is proven that the IDR has the same transitive closure as does the HBR. Moreover, it is shown that since the transitive closure of the HBR is antisymmetric and finite, it implies that the IDR is unique. This property is important because it means that all present, past or future works that deal with or will deal with a minimal expression or a transitive reduction of the HBR discuss the same binary relation.

In addition, an extension to the IDR in order to identify causal immediate dependencies only among a subset of *relevant* events[1] is presented. This is important since for a given distributed computation, usually only a subset of events is taken into account according to the problem to be solved. For example, for snapshot algorithms, a relevant event corresponds to the modification of a local variable involved in a global predicate; and for checkpointing algorithms, a relevant event is the definition of a local checkpoint [2].

The extension of the IDR applied to the set of relevant events for the causal delivery of messages is presented as case of study. It is shown that ensuring the IDR dependencies among the set of relevant events is *necessary* and *sufficient* in order to ensure the causal delivery of messages in the system.

This paper proceeds as follows. In Section 2, the system model is presented, as well as some order theory concepts and the Happened-Before Relation. In Section 3, the immediate dependency relation is presented, along with its extension for relevant events. Next, in Section 4, the IDRs minimality proof is

[1] In general, the relevant events are also referred as *observable* events.

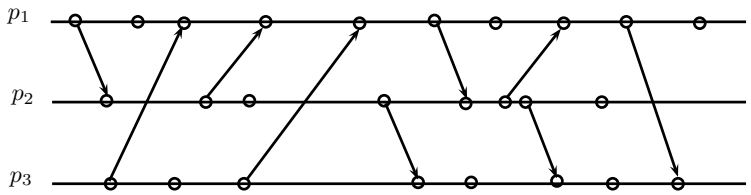

Figure 1: A Distributed Computation Scenario

given. Finally, in Section 5, some conclusions are presented.

2. Preliminaries

System Model

Processes. The system (see Figure 1) is composed of a set of processes $P = \{p_1, p_2, ..., p_n\}$. The processes present an asynchronous execution and communicate only by message passing.

Messages. There is a finite set of messages M, where each message $m \in M$ is sent considering an asynchronous reliable network that is characterized by no transmission time boundaries, no order delivery, and no loss of messages. The set of destinations of a message m is identified by $Dest(m)$.

Events. There are two types of events under consideration: internal and external events. An *internal* event is a unique action that occurs at a process p in a local manner (denoted in this paper by $internal(p)$) and which changes only the local process state. The finite set of internal events is denoted as E_i. On the other hand, while an external event is also a unique action that occurs at a process, it is seen by other processes, thus, affecting the global state of the system. The external events considered in this paper are the *send* and *delivery* events. Let m be a message. $send(m)$ denotes the emission event, while $delivery(p, m)$ represents the delivery event of m to participant $p \in P$. The set of events associated to M is the set $E_m = \{send(m) : m \in M\} \cup \{delivery(p, m) : m \in M \wedge p \in P\}$. The whole set of events in the system is the finite set $E = E_i \cup E_m$. Each event $e \in E$ is identified by a tuple $e = (p, x)$, where $p \in P$ is the producer of e, and x is the local logical clock for events of p, when e is carried out.

2.1. Order Theory Concepts

Transitive Closure. The transitive closure in our domain establishes the *reachability* between events. For a pair of events in the system, it is said that an event a is reachable from an event b if a causal path exists between them. The transitive closure is defined in general as follows [6].

Definition 1. *The transitive closure of a binary relation R on a set W is the smallest transitive relation on W that contains R.*

Property 1. *If the original relation is transitive, the transitive closure will be that same relation; otherwise, the transitive closure will be a different relation.*

Transitive Reduction. The transitive reduction of a binary relation is the minimal binary relation that expresses the same behavior (in this case, distributed computation) with the smallest set of related pair of elements. Its definition is as follows [1].

Definition 2. *A transitive reduction of a binary relation R on a set W is a minimal relation R' on W, such that the transitive closure of R' is the same as the transitive closure of R.*

Property 2. *If the transitive closure of R is antisymmetric and finite, then R' is unique.*

However, neither existence nor uniqueness of transitive reduction are generally guaranteed .

Covering Relation. In the order theory, a covering relation is a binary relation which holds between two comparable elements in a partially ordered set if they are immediate neighbors [1]. The covering relation is commonly used to graphically express the partial order by means of the Hasse diagram. Its definition is as follows:

Definition 3. *Let u and v be elements of a partially ordered set W. Then v covers u, written as $u <: v$, if $u < v$ and there is no element $w \in W$ such that $u < w < v$.*

Property 3. *If a partially ordered set (W, R) is finite, then its covering relation R' is the transitive reduction of the partial order relation R.*

Only if Property 3 is accomplished, a partially ordered set (W, R) is completely described by its Hasse diagram. On the other hand, for example in a *dense order*, such as in the case of the rational numbers, no element covers another.

2.2. Happened-Before Relation

The Happened-Before Relation (HBR) was defined by Lamport [5]. It establishes logical precedence dependencies over a set of events. The HBR is a strict partial order (transitive, irreflexive and antisymmetric) defined as follows:

Definition 4. *The causal relation "$\rightarrow$" is the smallest relation on a set of events E satisfying the following properties:*

1. *If a and b are events belonging to the same process, and a was originated before b, then $a \rightarrow b$.*

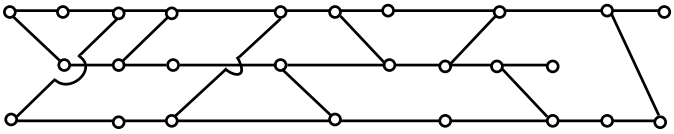

Figure 2: Hasse Diagram for the IDR of the scenario in Figure 1 (the partial order is established from left to right)

2. *If a is the sending of a message by one process, and b is the reception of the same message in another process, then $a \rightarrow b$.*
3. *If $a \rightarrow b$ and $b \rightarrow c$, then $a \rightarrow c$.*

By using Definition 4, one can say that a pair of events is concurrently related "$a||b$" only if $\neg(a \rightarrow b \vee b \rightarrow a)$.

The poset $\hat{E} = (E, \rightarrow)$ constitutes the formal model adopted in this paper for a distributed computation.

The Happened-Before Relation for Relevant Events (HBR-R). Usually for a given distributed computation $\hat{E}$, only a subset of events $R \subseteq E$ is relevant. The HBR for relevant events denoted in this paper by "$\stackrel{R}{\rightarrow}$" has been defined in [3] in the following way:

$$\forall(a, b) \in R \times R : (a \stackrel{R}{\rightarrow} b) \Leftrightarrow (a \rightarrow b)$$

The poset $\hat{R} = (R, \stackrel{R}{\rightarrow}) \subseteq (E, \rightarrow)$ constitutes the abstraction considered in this paper of the distributed computation for the relevant events.

3. Immediate Dependency Relation

The Immediate Dependency Relation (IDR) is known in order theory as a *covering* relation (see Definition 3). According to Property 3, if a partially ordered set is finite, its covering relation is the *transitive reduction* of the partial order relation. In this context, the IDR is then the covering relation of the HBR. Moreover, according to Property 2, since the poset $(E, \rightarrow)$ is finite and the HBR is a strict partial order, the IDR is unique. In this paper, the IDR is denoted by "$\downarrow$", and its formal definition is as follows:

Definition 5. *Two events $a, b \in E$ have an immediate dependency relation "$a \downarrow b$" if the following restriction is satisfied.*

$$a \downarrow b \text{ if } a \rightarrow b \text{ and } \forall c \in E, \neg(a \rightarrow c \rightarrow b)$$

Thus, an event a causal immediately precedes an event b, if and only if no other event c belonging to E exists (E is the set of events of the system), such that c belongs to the causal future of a and to the causal past of b. In Section 4 it is proved that the IDR is the transitive reduction of the HBR.

Based on the IDR, the following property is presented.

Property 4. *For all* $a, b \in E$, $a \neq b$

$$\textit{if } \exists c \in E \textit{ such that } (a \downarrow c \textit{ and } b \downarrow c) \textit{ or } (c \downarrow a \textit{ and } c \downarrow b) \textit{ then } a||b$$

This means that for every pair of events $a, b \in E$ with common IDR dependencies, the events are concurrently related. This property is leveraged in [10] in order to achieve a compact and consistent representation of a distributed system.

Finally, it is noted that $(E, \downarrow) \subset (E, \rightarrow)$.

The Hasse diagram for the IDR of the scenario in Figure 1 is shown in Figure 2.

Immediate Dependency Relation for Relevant Events (IDR-R). As for the HBR-R, the IDR must only reflect the IDR among the relevant events that belong to $R \subseteq E$. For this case, the IDR is referred as IDR-R, and it is denoted by "$\downarrow_R$". It is defined over $\hat{R}$ as follows:

$$a \downarrow_R b \text{ if } a \xrightarrow{R} b \text{ and } \forall c \in R, \neg(a \xrightarrow{R} c \xrightarrow{R} b)$$

Remark 1. $(R, \downarrow_R) \subset (R, \xrightarrow{R}) \subseteq (E, \rightarrow)$, *but* $(R, \downarrow_R) \not\subset (E, \downarrow)$

This means that the IDR-R is no longer a transitive reduction of the HBR. Instead, the IDR-R is the transitive reduction of the HBR-R (the proof is similar as for the IDR).

3.1. Case of Study: Message Causal Delivery

The selection of the set of relevant events must be determined according to the problem to be solved. For message causal delivery, there are two possible types of relevant events which are the *send* and the *delivery* events. It has been shown in [8] and [11] for group and multicast communication, respectively, that in order to ensure the causal delivery of messages in the system, it suffices to ensure the causal delivery of immediately related send events. Therefore, in general, to ensure message causal delivery for group and multicast communication, the set of relevant events is determined to be $R = \{send(m) : m \in M\}$ (see Figure 3). Formally, the message causal delivery based on the IDR-R can be defined as follows:

Theorem 1. *If* $\forall((send(m), send(m')) \in R, send(m) \downarrow_R send(m') \Rightarrow$
$\forall p \in Dest(m) \bigcap Dest(m') : delivery(p, m) \rightarrow delivery(p, m')$
then
$\forall((send(m), send(m')) \in R, send(m) \xrightarrow{R} send(m') \Rightarrow$
$\forall p \in Dest(m) \bigcap Dest(m') : delivery(p, m) \rightarrow delivery(p, m')$

The **proof** relies on the fact that for any pair $send(m) \xrightarrow{R} send(m')$ if $send(m) \downarrow_R send(m')$ does not hold, then a message m'' exists such that

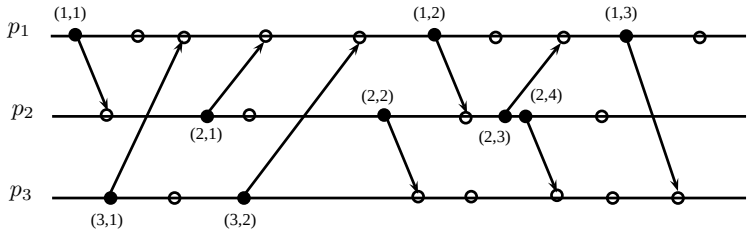

Figure 3: Relevants Events for Message Causal Delivery

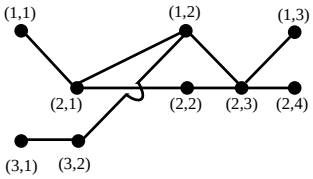

Figure 4: Hasse Diagram for the IDR-R of the Relevant Events Example in Figure 3

$send(m) \xrightarrow{R} send(m'') \xrightarrow{R} send(m')$. Using inductive reasoning and the fact that the event $send(m')$ may only have a finite number of "causes" or predecessors for the causal relation, (at least) one sequence $(send(m_i), i = 0, 1, ..., h)$ can be found, such that $m = m_0$, $m' = m_h$ and $send(m_i) \downarrow_R send(m_{i+1})$, for all $i = 0, 1, ..., h-1$. For any participant p, we have $delivery(p, m_i) \rightarrow delivery(p, m_{i+1})$, and by transitivity, the required property is obtained.

Clearly, the causal delivery of messages ensured by the IDR-R is not only a *sufficient* but also a *necessary* condition for the causal delivery of all causally related messages. From an algorithmic point of view, if the reference of some message m' IDR-R related to a message m is not piggy-backed with m, the causal delivery of m with respect to m' may fail. Theorem 1 shows that this information is sufficient. The Hasse diagram for the relevant events for message causal delivery of the scenario example is shown in Figure 4.

4. Minimality Proof of the IDR

In this section a proof to demonstrate that the IDR is the transitive reduction (minimal relation) of the HBR is given. In order to prove this, it must be demonstrated that, according to Definition 2, the IDR must have the same transitive closure as the HBR. By Property 1, which says that if the original binary relation (in this case the HBR) is transitive, then the transitive closure will be the same, one can conclude that the only property to demonstrate is that the transitive closure of the IDR is the HBR. By using the graph theory, the proof of this property is as follows.

Let $\hat{E}$ be a poset with strict partial order $\rightarrow$. Then $\hat{E}$ can be viewed as a directed graph where the vertex set is the ground set E, and the edge set is defined by $\rightarrow$.

Proposition 1. *Suppose every interval of $\hat{E}$ has a finite height. Then $\rightarrow$ is the transitive closure of $\downarrow$.*

PROOF. This is proven by induction on height. By definition of $\downarrow$, if $a \rightarrow b$ and the height of $[a, b]$ is 1, then $a \downarrow b$.

Assume by induction that whenever $a \rightarrow b$ and the height of $[a, b]$ is at most n, then (a, b) is in the transitive closure of $\downarrow$. Suppose that $a \rightarrow b$ and that the height of $[a, b]$ is $n + 1$. Since every chain in $[a, b]$ is finite, it contains an element c which is strictly larger than a and minimal with respect to this property. Therefore $[a, c] = \{a, c\}$, from which it is concluded that $a \downarrow c$. Since the interval $[c, b]$ is a proper subinterval of $[a, b]$, it has a height of at most n, so by the induction assumption one can conclude that (c, b) is in the transitive closure of $\downarrow$. Since (a, c) and (c, b) are in the transitive closure of $\downarrow$, so is (a, b). Hence, whenever $a \rightarrow b$ and the height of $[a, b]$ is at most $n + 1$, then (a, b) is in the transitive closure of $\downarrow$.

□

5. Conclusions

A formal study of the minimal binary relation of the HBR (Happened-Before Relation) which is called the *Immediate Dependency Relation* (IDR) is presented. In this paper, it is shown that the IDR identifies the smallest set of causally related pair of events in a given distributed computation. One important aspect is that because the HBR is a strict partial ordering, it implies that the IDR is unique. In addition, the IDR-R relation to identify causal immediate dependencies only among a subset of relevant events is introduced. As case of study, the IDR-R was applied to the particular problem of causal delivery of messages. The IDR-R has shown that it suffices to ensure the causal delivery of messages with IDR-R related *send* events in order to ensure the causal delivery of all messages in the system.

References

[1] A. Aho, M. Garey, J. Ullman, The Transitive Reduction of a Directed Graph, SIAM Journal on Computing 1(2), (1992), pp.131-137.

[2] E. Anceaume, J.M. Helary, M. Raynal, A Note on the Determination of the Immediate Predecessors in a Distributed Computation, Int. Journal of Foundations of Computer Science, 13(6), (2002), pp. 865-872.

[3] E. Fromentin, C. Jard, G-V Jourdan and M. Raynal, On-the-fly Analysis of Distributed Computation, Information Processing Letters 54, (1995), pp. 267-274.

[4] J.M. Helary, G. Melideo, Minimal size of Piggybacked Information for tracking causality: a graph-based characterization, Proceedings of the 26th International Workshop on Graph-Theoretic Concepts in Computer Science, LNCS Springer-Verlag, (2000), pp. 218-229.

[5] L. Lamport, Time, Clocks and the Ordering of Events in Distributed Systems, Communications ACM 21(7) (1978), pp. 558-565. Distributed Computing (1997), pp. 190-204.

[6] R. Lidl, G. Pilz, G., Applied abstract algebra, Undergraduate Texts in Mathematics, 2nd edition, Springer, (1998).

[7] L. Peterson, N. Buchholz, R. Schlichting, Preserving and Using Context Information in Interprocess Communication, ACM Transaction on Computer Systems (7), (1989), pp. 217-246.

[8] S.E. Pomares Hernandez, J. Fanchon, K. Drira, The Immediate Dependency Relation: an Optimal way to Ensure Causal Group Communication, Annu. Rev. Scal. Compt., Ser. Scal Compt, 6, (2004)pp. 61-79.

[9] S.E. Pomares Hernandez, Jorge Estudillo Ramirez, Luis A. Morales Rosales, and Gustavo Rodriguez Gomez, An Intermedia Synchronization Mechanism for Multimedia Distributed System, International Journal of Internet Protocol Technology, 4(3), 2009, pp. 207-218.

[10] S.E. Pomares Hernandez, J.R. Perez Cruz, M. Raynal, From the Happened-Before Relation to the Causal Ordered Set Abstraction, J. Parallel Distrib. Comput. 72(6), (2012), pp. 791-795.

[11] R. Prakash, M. Raynal, M. Singhal, An Adaptive Causal Ordering Algorithm Suited to Mobile Computing Environments, J. Parallel Distrib. Comput 41(2), (1997), pp. 190-204.