Extending the Causal Consistency Condition to any Object Defined by a Sequential Specification
Matthieu Perrin, Achour Mostefaoui, Michel Raynal

To cite this version:

HAL Id: hal-02053316
https://hal.archives-ouvertes.fr/hal-02053316
Submitted on 1 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Extending the Causal Consistency Condition to any Object Defined by a Sequential Specification

Achour Mostéfaoui†, Matthieu Perrin†, Michel Raynal‡,⋆

†LINA, Université de Nantes, 44322 Nantes, France
‡Univ Rennes, IRISA, 35042 Rennes, France
⋆Department of Computing, Polytechnic University, Hong Kong

Abstract

This paper presents a generalization of causal consistency suited to the family of objects defined by a sequential specification. As causality is captured by a partial order on the set of operations issued by the processes on shared objects (concurrent operations are not ordered), it follows that causal consistency allows different processes to have different views of each object history.

Keywords: Causality, Causal order, Concurrent object, Consistency condition.

1 Processes and Concurrent Objects

Let us consider a set of $n$ sequential asynchronous processes $p_1, ..., p_n$, which cooperate by accessing shared objects. These objects are called concurrent objects. A main issue consists in defining the correct behavior of concurrent objects. Two classes of objects can be distinguished according to way they are specified.

- The objects which can be defined by a sequential specification. Roughly speaking, this class of objects includes all the objects encountered in sequential computing (e.g., queue, stack, set, dictionary, graph). Different
tools can be used to define their correct behavior (e.g., transition function, list of all the correct traces -histories-, pre and post-conditions, etc.).

It is usually assumed that the operations accessing these objects are total, which means that, whatever the current state of the object, an operation always returns a result.

As an example, let us consider a bounded stack. A pop() operation returns a value if the stack is not empty, and returns the value ⊥ if it is empty. A push(v) operation returns the value ⊤ if the stack is full, and returns ok otherwise (v was then added to the stack). A simpler example is a read/write register, where a read operation always returns a value, and a write operation always returns ok.

- The objects which cannot be defined by a sequential specification. Example of such objects are Rendezvous objects or Non-blocking atomic commit objects [10]. These objects require processes to wait each other, and their correct behavior cannot be captured by sequences of operations applied to them.

In the following we consider objects defined by a sequential specification.

2 Strong Consistency Conditions

Strong consistency conditions are natural (and consequently easy to understand and use) in the sense that they require each object to appear as if it has been accessed sequentially. In a failure-free context, this can be easily obtained by using mutual exclusion locks bracketing the invocation of each operation.

Atomicity/Linearizability The most known and used consistency condition is atomicity, also called linearizability. It requires that each object appears as if it was accessed sequentially, this sequence of operations belonging to the specification of the object, and complying with the real-time order of their occurrences.

Sequential consistency This consistency condition, introduced in [15], is similar to, but weaker than, linearizability, namely, it does not require the sequence of operations to comply with real-time order.

\footnote{Atomicity was formally defined in [16] [17] for basic read/write objects. It was then generalized to any object defined by a sequential specification in [13]. We consider these terms as synonyms in the following.}
Figure 1 presents an example of a sequentially consistent computation (which is not atomic) involving two read/write registers $R_1$ and $R_2$, accessed by two processes $p_1$ and $p_2$. The dashed arrows define the causality relation linking the read and write operations on each object (also called read-from relation when the object is a read/write register). It is easy to see that the sequence of operations made up of all the operations issued by $p_2$, followed by all the operations issued by $p_1$, satisfies the definition of sequential consistency.

Implementing a strong consistency condition in an asynchronous message-passing system  

Shared memories usually provide processes with objects built on top of basic atomic read/write objects or more sophisticated objects accessed by atomic operations such as Test&Set or Compare&Swap [11, 12, 22, 25]. This is no longer the case in message-passing systems where all the objects (except communication channels) have to be built from scratch [5, 21].

Implementations of sequentially consistent objects and atomic objects in failure-free message-passing systems can be found in [4, 5, 7, 20, 21]. These implementations rest on a mechanism which allows a total order on all operations to be built. This can be done by a central server, or a broadcast operation delivering messages in the same order at all the processes. Such an operation is usually called total order broadcast (TO-broadcast) or atomic broadcast. It is shown in [20] that, from an implementation point of view, sequential consistency can be seen as a form of lazy linearizability. The “compositional” power of sequential consistency is addressed in [8, 18].

Implementations of a strong consistency condition (such as atomicity) in failure-prone message-passing systems is more difficult. More precisely, except for a few objects including read/write registers (which can be built only in systems where, in each execution, a majority of processes do not crash [3]), it is impossible to implement an atomic object in the presence of asynchrony and process crashes [9]. Systems have to be enriched with additional computing power (such as randomization or failure detectors) to be able to implement objects defined by a strong consistency condition.
3 Causal Consistency on Read/Write Objects
(Causal Memory)

Causality-based consistency condition  A causal memory is a set of read/write objects satisfying a consistency property weaker that atomicity or sequential consistency. This notion was introduced in [1]. It relies on a notion of causality similar to the one introduced in [14] for message-passing systems.

The main difference between causal memory and the previous strong consistency conditions lies in the fact that causality is captured by a partial order, which is trivially weaker than a total order. A total order-based consistency condition forces all the processes to see the same order on the object operations. Causality-based consistency does not. Each process can have its own view of the execution, their ”greatest common view” being the causality partial order produced by the execution. Said differently, an object defined by a strong consistency condition is a single-view object, while an object defined by a causality-based consistency condition is a multi-view object (one view per process).

Another difference between a causality-based consistency condition and a strong consistency condition lies in the fact that a causality-based consistency condition copes naturally with process crashes and system partitioning.

Preliminary definitions  As previously indicated, a causal memory is a set of read/write registers. Its semantics is based on the following preliminary definitions (from [1,13]). To simplify the presentation and without loss of generality, we assume that (a) all the values written in a register are different, and (b) each register has an initial value written by a fictitious write operation.

- A local (execution) history $L_i$ of a process $p_i$ is the sequence of read and write operations issued by this process. If the operations $op_1$ and $op_2$ belong to $L_i$ and $op_1$ appears before $op_2$, we say “$op_1$ precedes $op_2$ in $p_i$’s process order”. This is denoted $op_1 \rightarrow_i op_2$.

- The write-into relation (denoted $\rightarrow_{wi}$) captures the effect of write operations on the read operations. Denoted $\rightarrow_{wi}$, it is defined as follows: $op_1 \rightarrow_{wi} op_2$ if $op_1$ is the write of a value $v$ into a register $R$ and $op_2$ is a read operation of the register $R$ which returns the value $v$.

- An execution history $H$ is a partial order composed of one local history per process, and a partial order, denoted $\rightarrow_{po}$, defined as follows: $op_1 \rightarrow_{po} op_2$ if
  - $op_1, op_2 \in L_i$ and $op_1 \rightarrow_i op_2$ (process order), or
  - $op_1 \rightarrow_{wi} op_2$ (write-into order), or
∃ op3 such that op1 \xrightarrow{po} \ op3 and \ op3 \xrightarrow{po} \ op2 (transitivity).

- Two operations not related by \xrightarrow{po} are said to be independent or concurrent.
- The projection of \ H \ on a register \ R \ (denoted \ H|R) \ is \ the \ partial \ order \ \ H \ from \ which \ are \ suppressed \ all \ the \ operations \ which \ are \ not \ on \ \ R.
- A serialization \ S \ of \ an \ execution \ history \ \ H \ (whose \ partial \ order \ is \ \xrightarrow{po}) \ is \ a \ total \ order \ such \ that, \ if \ op1 \xrightarrow{po} \ op2, \ then \ op1 \ precedes \ op2 \ in \ \ S.

A remark on the partial order relation

As we can see, the read-from relation mimics the causal send/receive relation associated with message-passing [14]. The difference is that zero, one, or several reads can be associated with the same write. In both cases, the (write-into or message-passing) causality relation is a global property (shared by all processes) on which is built the consistency condition. It captures the effect of the environment on the computation (inter-process asynchrony), while process orders capture the execution of the algorithms locally executed by each process.

Causal memory

Let \ H_{i+w} \ be \ the \ partial \ order \ \xrightarrow{po}, \ from \ which \ all \ the \ read \ operations \ not \ issued \ by \ p_i \ are \ suppressed (the \ subscript \ i + w \ means \ that \ only \ all \ the \ operations \ issued \ by \ p_i \ plus \ all \ write \ operations \ are \ considered).

As defined in [1], an execution history \ H \ is causal if, for each process \ p_i, \ there is a serialization \ S_i \ of \ \ H_{i+w} \ in \ which \ each \ read \ from \ a \ register \ R \ returns \ the \ value \ written \ in \ \ R \ by \ the \ most \ recent \ preceding \ write \ in \ \ R.

This means that, from the point of view of each process \ p_i, \ taken independently from the other processes, each register behaves as defined by its sequential specification. It is important to see, that different processes can have different views of a same register, each corresponding to a particular serialization of the partial order \ \xrightarrow{po} \ from \ which \ the \ read \ operations \ by \ the \ other \ processes \ have \ been \ eliminated.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example.png}
\caption{Example of an execution of a causal read/write memory}
\end{figure}

An example of a causal memory execution is depicted in Figure 2. Only one write-into pair is indicated (dashed arrow). As \ R1.write(1) \ and \ R1.write(2) \ are
independent, each of the operations $R1.read()$ by $p_2$ and $p_3$ can return any value, i.e., $u, v \in \{1, 2\}$. For the same reason, and despite the write-into pair on the register $R2$ involving $p_1$ and $p_3$, the operation $R1.read()$ issued by $p_3$ can return $w \in \{1, 2\}$. This shows that different processes can obtain different “views” of the same causal memory execution. Once a read returned a value, a new write-into pair is established.

Implementations of a causal read/write memory (e.g., [2]) rest on an underlying communication algorithm providing causal message delivery [6, 24]. It is shown in [1, 23] that, in executions that are data race-free or concurrent write-free, a causal memory behaves as a sequentially consistent read/write memory.

4 Causal Consistency for any Object

The problem Albeit it was introduced more than 20 years ago, it appears that, when looking at the literature, causal consistency has been defined and investigated only for read/write objects (the only exception we are aware of is [19]). This seems to be due to the strong resemblance between read/write operations and send/receive operations. Hence, the question: Is it possible to generalize causal consistency to any object defined by a sequential specification? This section answers positively this question.

Preliminary definitions The notations and terminology are the same as in the previous section, but now the operations are operations on any object $O$ of a set of objects $O$, each defined by a sequential specification.

Considering a set of local histories and a partial order $\rightarrow_{po}$ on their operations, let $Assignment_{i}(\rightarrow_{po})$ denote the partial order $\rightarrow_{po}$, in which, for each operation $op()$ not issued by $p_i$, the returned value $v$ is replaced by a value $v'$, possibly different from $v$, the only constraint being that $v$ and $v'$ belong to the same domain (as defined by the corresponding operation $op()$). Let us notice that $Assignment_{i}(\rightarrow_{po})$ is not allowed to modify the values returned by the operations issued by $p_i$. Moreover, according to the domain of values returned by the operations, a lot of different assignments can be associated with each process $p_i$.

Given a partial order $\rightarrow_{po}$, and an operation $op$, the causal past of $op$ with respect to $\rightarrow_{po}$ is the set of operations $\{op' \mid op' \rightarrow_{po} op\}$. A serialization $S_i$ of a partial order $\rightarrow_{po}$ is said to be causal past-constrained if it is such that, for any operation $op$ issued by $p_i$, only the operations of the causal past of $op$ appear before $op$. 
Causal consistency for any object

Let $H = \langle L_1, \ldots, L_n \rangle$ be a set of $n$ local histories (one per process) which access a set $O$ of concurrent objects, each defined by a sequential specification. $H$ is causally consistent if there is a partial order $\rightarrow_{po}$ on the operations of $H$ such that for any process $p_i$: 

- $(\text{op}1 \rightarrow_{po} \text{op}2) \Rightarrow (\text{op}1 \rightarrow \text{op}2)$, and
- $\exists$ an assignment $\text{Assignment}_i$ and a causal past-constrained serialization $S_i$ of $\text{Assignment}_i(\rightarrow_{po})$ such that, $\forall O \in O$, $S_i|O$ belongs to the sequential specification of $O$.

The first requirement states that the partial order $\rightarrow_{po}$ must respect all process orders. The second requirement states that, as far as each process $p_i$ is concerned, the local view (of $\rightarrow_{po}$) it obtains is a total order (serialization $S_i$) that, according to some value assignment, satisfies the sequential specification of each object $O$.  

Let us remark that the assignments $\text{Assignment}_i()$ and $\text{Assignment}_j()$ associated with $p_i$ and $p_j$, respectively, may provide different returned values in $S_i$ and $S_j$ for the same operation. Each of them represents the local view of the corresponding process, which is causally consistent with respect to the global computation as captured by the relation $\rightarrow_{po}$.

When the objects are read/write registers

The definition of a causal memory stated in Section 3 is a particular instance of the previous definition. More precisely, given a process $p_i$, the assignment $\text{Assignment}_i$ allows an appropriate value to be associated with every read not issued by $p_i$. Hence, there is a (local to $p_i$) assignment of values such that, in $S_i$, any read operation returns the last written value. In a different, but equivalent way, the definition of a causal read/write memory given in [1] eliminates from $S_i$ the read operations not issued by $p_i$. While such operation eliminations are possible for read/write objects, they are no longer possible when one wants to extend causal consistency to any object defined by a sequential specification. This come from the observation that, while a write operation resets “entirely” the value of the object, “update” operations on more sophisticated objects defined by a sequential specification (such as the operations push() and pop() on a stack for example), do not reset “entirely” the value of the object. The memory of such objects has a richer structure than the one of a basic read/write object.

\footnote{This definition is slightly stronger than the definition proposed in [19]. Namely, in addition to the introduction of the assignment notion, the definition introduced above adds the constraint that, if an operation $\text{op}$ precedes an operation $\text{op}'$ in the process order, then the serialization required for $\text{op}$ must be a prefix of the serialization required for $\text{op}'$. On the other hand, it describes precisely the level of consistency achieved by Algorithm 1 presented below.}
An example  As an example illustrating the previous general definition of a causally consistent object, let us consider three processes $p_1$, $p_2$ and $p_3$, whose accesses to a shared unbounded stack are captured by the following local histories $L_1$, $L_2$, and $L_3$. In these histories, the notation $\text{op}_i(a)r$ denotes the operation $\text{op}()$ issued by $p_i$, with the input parameter $a$, and whose returned value is $r$.

- $L_1 = \text{push}_1(a)\text{ok}, \text{push}_1(c)\text{ok}, \text{pop}_1()c$.
- $L_2 = \text{pop}_2()a, \text{push}_2(b)\text{ok}, \text{pop}_2()b$.
- $L_3 = \text{pop}_3()a, \text{pop}_3()b$.

Hence, the question: Is $H = \langle L_1, L_2, L_3 \rangle$ causally consistent? We show that the answer is “yes”. To this end we need first to build a partial order $\rightarrow$ respecting the three local process orders. Such a partial order is depicted in Figure 3, where process orders are implicit, and the inter-process causal relation is indicated with dashed arrows (let us remind that this relation captures the effect of the environment –asynchrony– on the computation).

![Figure 3: Example of a partial order on the operations issued on a stack](image)

The second step consists in building three serializations respecting $\rightarrow$, $S_1$ for $p_1$, $S_2$ for $p_2$, and $S_3$ for $p_3$, such that, for each process $p_i$, there is an assignment of values returned by the operations $\text{pop}()$ (Assignment($()$)), from which it is possible to obtain a serialization $S_i$ belonging to the specification of the stack. Such assignments/serializations are given below.

- $S_1 = \text{push}_1(a)\text{ok}, \text{pop}_3()a, \text{push}_1(c)\text{ok}, \text{pop}_2()\perp, \text{push}_2(b)\text{ok}, \text{pop}_1()c, \text{pop}_2()b, \text{pop}_3()\perp$.
- $S_2 = \text{push}_1(a)\text{ok}, \text{pop}_2()a, \text{push}_2(b)\text{ok}, \text{pop}_2()b, \text{pop}_3()\perp, \text{pop}_3()\perp, \text{push}_1(c)\text{ok}, \text{pop}_1()c$.
- $S_3 = \text{push}_1(a)\text{ok}, \text{pop}_3()a, \text{pop}_2()\perp, \text{push}_3(b)\text{ok}, \text{pop}_3()b, \text{pop}_2()\perp, \text{push}_1(c)\text{ok}, \text{pop}_1()c$. 
The local view of the stack of each process $p_i$ is constrained only by the causal order depicted in Figure 3 and also depends on the way it orders concurrent operations. As far as $p_2$ is concerned we have the following, captured by its serialization/assignment $S_2$. (The serializations $S_1$ and $S_3$ are built similarly.) We have considered short local histories, which could be prolonged by adding other operations. As depicted in the figure, due to the last causality (dashed) arrows, those operations would have all the operations in $L_1 \cup L_2 \cup L_3$ in their causal past.

1. Process $p_2$ sees first $\text{push}_1(a)\text{ok}$, and consequently (at the implementation level) updates accordingly its local representation of the stack.
2. Then, $p_2$ sees its own invocation of $\text{pop}_2()$ which returns it the value $a$.
3. Then, $p_2$ sees its own $\text{push}_2(b)$ and $\text{pop}_2()$ operations; $\text{pop}_2()$ returns consequently $b$.
4. Finally $p_2$ becomes aware of the two operations $\text{pop}_3()$ issued by $p_3$, and the operations $\text{push}_1(c)$ and $\text{pop}_1()$ issued by $p_1$. To have a consistent view of the stack, it considers the assignment of returned values that assigns the value $\bot$ to the two operations $\text{pop}_3()$, and the value $c$ to the operations $\text{pop}_1()$. In this way, $p_2$ has a consistent view of the stack, i.e., a view which complies with the sequential specification of a stack.

A universal construction  Algorithm[1] is a universal construction which builds causally consistent objects from their sequential specification. It considers deterministic objects. This algorithm is built on top of any underlying algorithm ensuring causal broadcast message delivery [6, 24]. Let “co广播 MSG(a)” denote the causal broadcast of a message tagged MSG carrying the value a. The associated causal reception at any process is denoted “co-delivery”. “?” denotes a control value unknown by the processes at the application level.

```latex
\textbf{when } p_i \text{ invokes } O.\text{op}(\text{param}) \text{ do }\\
(1) \quad \text{result}_i \leftarrow ?; \\
(2) \quad \text{co广播 } \text{OPERATION}(i, O, \text{op}(\text{param})); \\
(3) \quad \text{wait } (\text{result}_i \neq ?); \\
(4) \quad \text{return } (\text{result}_i).

\textbf{when } \text{OPERATION}(j, O, \text{op}(\text{param})) \text{ is co-delivered do }\\
(5) \quad (r, \text{state}_i[O]) \leftarrow \delta_O(\text{state}_i[O], \text{op}(\text{param})); \\
(6) \quad \text{if } (j = i) \text{ then } \text{result}_i \leftarrow r \text{ end if.}
```

Algorithm 1: Universal construction for causally consistent objects (code for $p_i$)

Interestingly, the replacement of the underlying message causal order broadcast by a message total order broadcast, implements linearizability.
Each object $O$ is defined by a transition function $\delta_O()$, which takes as input parameter the current state of $O$ and the operation $\text{op}(\text{param})$ applied to $O$. It returns a pair $(r, \text{new\_state})$, where $r$ is the value returned by $\text{op}(\text{param})$, and $\text{new\_state}$ is the new state of $O$. Each process $p_i$ maintains a local representation of each object $O$, denoted $\text{state}_i[O]$.

When a process $p_i$ invokes an operation $\text{op}(\text{param})$ on an object $O$, it co-broadcasts the message $\text{OPERATION}(i, O, \text{op}(\text{param}))$, which is co-delivered to each process (i.e., according to causal message order). Then, $p_i$ waits until this message is locally processed. When this occurs, it returns the result of the operation.

When a process $p_j$ co-delivers a message $\text{OPERATION}(j, O, \text{op}(\text{param}))$, it updates accordingly its local representation of the object $O$. If $p_j$ is the invoking process, it additionally locally returns the result of the operation.

5 Conclusion

This short article extended the notion of causal consistency to any object defined by a sequential specification. This definition boils down to causal memory when the objects are read/write registers.

The important point in causal consistency lies in the fact that each process has its own view of the objects, and all these views agree on the partial order on the operations but not necessarily on their results. More explicitly, while each process has a view of each object, which locally satisfies its object specification, two processes may disagree on the value returned by some operations. This seems to be the "process-to-process inconsistency cost" that must be paid when weakening consistency by considering a partial order instead of a total order. On another side and differently from strong consistency conditions, causal consistency copes naturally with partitioning and process crashes.

Acknowledgments

This work has been partially supported by the Franco-German DFG-ANR Project 40300781 DISCMAT (devoted to connections between mathematics and distributed computing), and the French ANR project DESCARTES (devoted to layered and modular structures in distributed computing).
References


