Unobtrusive Deferred Update Stabilization for Efficient Geo-Replication

Chathuri Gunawardhana, Manuel Bravo, and Luis Rodrigues, ***University of Lisbon***

https://www.usenix.org/conference/atc17/technical-sessions/presentation/gunawardhana

This paper is included in the Proceedings of the 2017 USENIX Annual Technical Conference (USENIX ATC '17).

July 12–14, 2017 • Santa Clara, CA, USA

ISBN 978-1-931971-38-6

Open access to the Proceedings of the 2017 USENIX Annual Technical Conference is sponsored by USENIX.

Unobtrusive Deferred Update Stabilization for Efficient Geo-Replication

Chathuri Gunawardhana[1], *Manuel Bravo*[1,2] *and Luís Rodrigues*[1]
[1]INESC-ID, Instituto Superior Técnico, Universidade de Lisboa [2]Université Catholique de Louvain, Belgium

Abstract

In this paper, we propose a novel approach to manage the throughput vs visibility latency tradeoff that emerges when enforcing causal consistency in geo-replicated systems. Our approach consists in allowing full concurrency when processing local updates and using a deferred *local* serialisation procedure before shipping updates to remote datacenters. This strategy allows to implement inexpensive mechanisms to ensure system consistency requirements while avoiding intrusive effects on update operations, a major performance limitation of previous systems. We have implemented our approach as a variant of Riak KV. Our evaluation shows that we outperform sequencer-based approaches by almost an order of magnitude in the maximum achievable throughput. Furthermore, unlike previous sequencer-free solutions, our approach reaches nearly optimal remote update visibility latencies without limiting throughput.

1 Introduction

Geo-replication is a requirement for modern internet-based services in order to improve user-perceived latency. Unfortunately, due to the long network delays among sites, synchronous replication is prohibitively slow for most practical purposes. Therefore, many systems resort to weaker consistency semantics that permit some form of asynchronous replication strategy.

Among the many consistency guarantees that allow for asynchronous replication [15], causal consistency [9] has been identified as the strongest consistency model that an always-available system can implement [14, 37], becoming of practical relevance in geo-replicated settings. In fact, causal consistency is key in many geo-replicated storage systems offering from weak [38, 35, 12, 44] to strong consistency guarantees [41, 34, 17].

Unfortunately, implementing causal consistency is costly due to the computation, communication, and storage overhead caused by metadata management [19, 27, 16]. A common solution to reduce this cost consists in compressing metadata by serializing sources of concurrency, which unavoidably creates *false dependencies* among concurrent events, increasing *visibility latencies* (time interval between the instant in which an update is installed in its origin datacenter and when it becomes visible in remote datacenters).

To safely compress metadata, designers of causally consistent systems rely either on: (i) centralized *sequencers* (commonly one per datacenter) [44, 12]; or (ii) *global stabilization* procedures [24, 10] (executed across datacenters). The former has the advantage of making trivial—and therefore inexpensive—the dependency checking procedures at the cost of severely limiting concurrency, as sequencers operate in the critical path of clients. On the contrary, the latter avoids centralized synchronization points at the cost of periodically running a global stabilization procedure in the background. The cost of this procedure has pushed some systems to over-compress metadata to avoid impairing throughput, with a significant penalty on the visibility latencies [24].

In this paper, we propose, implement, and evaluate a novel approach to address the metadata size versus visibility latency tradeoff. Our approach has some similarities with systems that rely on global stabilization but also significant differences. As with [24, 10], we let local updates proceed without any a priori synchronization. However, unlike previous systems, we totally order all updates, in a manner consistent with causality, before shipping them to remote datacenters. As a result, expensive global stabilization is avoided, as it is trivial for a datacenter to check whether all updates subsumed in the timestamps piggybacked by remote updates have been locally applied (similarly to sequencer-based solutions).

We have implemented our approach as a variant of the open source version of Riak [6]. We have augmented Riak with Eunomia[1], a service that totally orders all lo-

[1]Greek goddess of law, her name can be translated as "good order".

cal updates, before shipping them. Our results show that Riak+Eunomia outperforms sequencer-based systems by almost an order of magnitude while serving significantly better quality-of-service to clients compared with systems based on global stabilization procedures.

In summary, the contributions of this paper are: i) The introduction of Eunomia, a new service for unobtrusively ordering updates (§3); ii) A fault tolerant version of Eunomia (§3.3); iii) An experimental comparison of the maximum load that traditional sequencers and Eunomia can handle, and their potential bottlenecks (§7.1); iv) The Integration of Eunomia into an always-available geo-replicated data store (§4) and its performance comparison to state-of-the-art solutions (§7.2).

2 Motivation and Goals

We start by motivating our work with a simple experiment, showing that: (i) the major throughput impairment of sequencer-based solutions is the fact that they operate in the critical path of clients; and (ii) global stabilization procedures are expensive in practice, forcing designers to favour either throughput or visibility latencies.

Figure 1 plots the throughput penalty and visibility latency overhead introduced by state-of-the-art causally consistent solutions. Results are normalized against an eventually consistent system, which adds no overhead due to consistency management. We vary from 1*ms* to 100*ms* the interval between global stabilization computations to better understand the cost and the consequences of such mechanism. Our deployment consists of 3 datacenters. The round-trip-times across datacenters are 80*ms* between datacenter 1 (dc_1) and both dc_2 and dc_3; and 160*ms* between dc_2 and dc_3. In the figure (left plot), latencies refer to the (90^{th} percentile) delays incurred by each system at dc_2 for updates originating at dc_1. We compare the performance of 4 systems, namely *S-Seq*, *A-Seq*, *GentleRain* and *Cure*. For each solution, we deploy as many clients as possible (not necessarily the same amount for each experiment) without saturating the system.

S-Seq is a system that relies on a sequencer per datacenter to compress metadata; it uses a vector with an entry per datacenter to track causality, as in [12, 44]. *A-Seq* is an asynchronous (bogus) variant of *S-Seq*, that contacts the sequencer in parallel with applying the update. *A-Seq* does the same total amount of work as *S-Seq* and, although it fails to capture causality, it serves to reason about the potential benefits of removing sequencers from client's critical operational path. *GentleRain* [24] and *Cure* [10] are well known solutions that rely on global stabilization. The former favours throughput, over-compressing metadata into a single scalar; the latter favours visibility latencies, compressing metadata into a vector with an entry per datacenter.

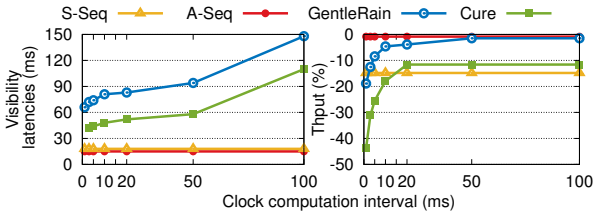

Figure 1: Update visibility latency vs throughput tradeoff.

The results confirm that the costs inherent to global stabilization force designers to choose between optimizing throughput and visibility latencies. As Figure 1 shows, Cure offers lower visibility latencies than GentleRain (as causality is more precisely tracked) at the cost of penalizing throughput. GentleRain does the opposite tradeoff favouring throughput. Cure can tune this tradeoff by choosing longer intervals among global stabilization occurrences. Nevertheless, even with long intervals (100*ms*), Cure still significantly degrades system throughput by 11.6%. Interestingly, results also show that by removing the sequencer from client's critical operational path, sequencer-based approaches could potentially pick a better spot in the tradeoff space, by providing throughput and visibility latencies comparable to GentleRain and Cure respectively, with almost no performance overhead when compared to the baseline. Note that in the above experiment, sequencers are not saturated; therefore, the throughout penalty (14.8%) is exclusively caused by the synchronous communication between the sequencer and the partitions at every client update operation. Later, in §7.1, we experimentally measure the maximum load that sequencers can handle before getting saturated.

From these results, it is possible to get the following insight: in order to alleviate the tension between throughput and visibility latencies, one has to (i) avoid global stabilization, and (ii) rely on an abstraction similar to sequencers that allows for trivial—therefore inexpensive—dependency checking procedures, while removing its operation from the client's critical path. Our goal was then to design Eunomia, a system with such characteristics.

3 Eunomia: Unobtrusive Ordering

In this section, we present the design and rationale underlying Eunomia, a new service conceived to replace sequencers as building blocks in weakly consistent geo-replicated storage systems. Unlike traditional sequencers, Eunomia lets local client operations to execute without synchronous coordination, an essential characteristic to avoid limiting concurrency and increasing operation latencies. Then, in the background, Eunomia establishes a serialization of all updates occurring in the local datacenter in an order consistent with causality, based

N	Number of partitions
$Clock_c$	Client c clock
p_n	Partition n
$Clock_n$	Current physical time at p_n
Ops	Set of unstable operations at Eunomia
$PartitionTime$	Vector with an entry per partition at Eunomia
$u_j.ts$	Timestamp assigned to update u_j

Table 1: Notation used in the protocol description.

Algorithm 1 Operations at client c

```
1:  function READ(Key)
2:      send READ(Key) to RESPONSIBLE(Key)
3:      receive ⟨Value, Ts⟩ from RESPONSIBLE(Key)
4:      Clock_c ← MAX(Clock_c, Ts)
5:      return Value
6:  function UPDATE(Key, Value)
7:      send UPDATE(Key, Value, Clock_c) to RESPONSIBLE(Key)
8:      receive Ts from RESPONSIBLE(Key)
9:      Clock_c ← Ts
10:     return ok
```

on timestamps generated locally by the individual servers that compose the datacenter. We refer to this process as *site stabilization procedure*. Thus, Eunomia is capable of abstracting the internal complexity of a multi-server datacenter without limiting the concurrency. Eunomia can be used to improve any existing sequencer-based solution to enforce causal consistency across geo-locations [38, 44, 12], as shown in §4.

3.1 Eunomia Into Play

In order to convey how Eunomia works, we start by presenting the protocol used to support the interaction between Eunomia and the machines that constitute a datacenter. In the exposition, we assume that the object-space is divided into *N partitions* distributed among datacenter machines. Updates to objects belonging to the same partition are serialized by the native update protocol. To simplify the presentation, our pseudocode assumes FIFO links among partitions and Eunomia. Later, in §3.3, we eliminate this assumption, making its implementation explicit. Table 1 provides a summary of the notation used in the protocols.

Eunomia assumes that each individual partition can assign a timestamp to each update without engaging in synchronous coordination with other partitions, or with Eunomia. We will explain below how this can be easily achieved. These timestamps must satisfy two properties.

Property 1. If an update u_j causally depends on a second update u_i, then the timestamp assigned to u_j ($u_j.ts$) is strictly greater than $u_i.ts$.

Property 2. For two updates u_i and u_j received by Eunomia from partition p_n, if u_i is received before u_j then $u_j.ts$ is strictly greater than $u_i.ts$.

These two properties imply that updates are causally ordered across all partitions and that once Eunomia receives an update coming from a partition p_n, no update with smaller timestamp will be ever received from p_n. In order to ensure these properties, clients play a fundamental role. A client c maintains a local variable, $Clock_c$, that stores the largest timestamp seen during its session. This clock value captures the client's causal dependencies and it is included in every update request. As described below, partitions compute update timestamps taking into account the value of client clocks.

Algorithm 2 Operations at partition p_n

```
1:  function READ(Key)
2:      ⟨Value, Ts⟩ ← KV_GET(Key)
3:      return ⟨Value, Ts⟩
4:  function UPDATE(Key, Value, Clock_c)
5:      MaxTs_n ← MAX(Clock_n, Clock_c + 1, MaxTs_n + 1)
6:      KV_PUT(Key, ⟨Value, MaxTs_n⟩)
7:      u_j ← ⟨Key, Value, MaxTs_n, p_n⟩
8:      send ADD_OP(u_j) to Eunomia
9:      return MaxTs_n
10: function HEARTBEAT                          ▷ Every Δ time
11:     if Clock_n ≥ MaxTs_n + Δ then
12:         send HEARTBEAT(p_n, Clock_n) to Eunomia
```

The protocol assumes that each partition p_n is equipped with a physical clock. Clocks are loosely synchronized by a time synchronization protocol such as NTP [5]. The correctness of the protocol does not depend on the clock synchronization precision and can tolerate clock drifts. However, as discussed later, large clock drifts could have a negative impact on the protocol performance (in particular, on how fast the datacenter can ship updates to remote datacenters). To avoid this limitation, our protocol uses hybrid clocks [30], which have been shown to overcome some of the limitations of simply using physical time.

We now describe how events are handled by clients, partitions and Eunomia (Algs. 1, 2, and 3 respectively).

Read. A client c sends a read request on item Key to the responsible partition p_n (Alg. 1, line 2). When p_n receives the request, it fetches the $Value$ and the timestamp Ts that is locally stored for Key and returns both to the client. Ts is the timestamp assigned by p_n to the update operation that generated the current version. After receiving the pair $\langle Value, Ts\rangle$, the client computes the maximum between $Clock_c$ and Ts (Alg. 1, line 4) to include the read operation in its causal history.

Update. A client c sends an update request operation to the responsible partition p_n of the object being updated. Apart from the Key and $Value$, the request includes the client's clock $Clock_c$ (Alg. 1, line 7). When p_n receives the request, it first computes the timestamp of the new update (Alg. 2, line 5). This is computed by taking the maximum between $Clock_n$ (physical time), the maximum timestamp ever used by p_n ($MaxTs_n$) plus one and $Clock_c$ (client's clock) plus one. This ensures that

Algorithm 3 Operations at Eunomia

```
1:  function ADD_OP(u_j)
2:      Ops ← Ops ∪ u_j
3:      ⟨Key, Value, Ts, p_n⟩ ← u_j
4:      PartitionTime[p_n] ← Ts
5:  function HEARTBEAT(p_n, Ts)
6:      PartitionTime[p_n] ← Ts
7:  function PROCESS_STABLE                          ▷ Every θ time
8:      StableTime ← MIN(PartitionTime)
9:      StableOps ← FIND_STABLE(Ops, StableTime)
10:     PROCESS(StableOps)
11:     Ops ← Ops \ StableOps
```

the timestamp is greater than both $Clock_c$ and any other update timestamped by p_n. Then, p_n stores the *Value* and the recently computed timestamp in the local key-value store and asynchronously sends the operation to the Eunomia service. Finally, p_n returns update's timestamp to the client who updates $Clock_c$ with it, since it is guaranteed to be greater than its current one.

Timestamp Stability. When Eunomia receives an operation from a given partition, it adds it to the set of non-stable operations Ops and updates the p_n entry in the *PartitionTime* vector with operation's timestamp (Alg. 3, lines 2–4). A timestamp Ts is *stable* at Eunomia when one is sure that no update with lower timestamp will be received from any partition (i.e., when Eunomia is aware of all updates with timestamp Ts or smaller). Periodically, Eunomia computes the value of the maximum stable timestamp (*StableTime*), which is computed as the minimum of the *PartitionTime* vector (Alg. 3, line 8). Property 2 implies that no partition will ever timestamp an update with an equal or smaller timestamp than *StableTime*. Thus, Eunomia can confidently serialize all operations tagged with a timestamp smaller than or equal to *StableTime* (Alg. 3, line 9). Eunomia can serialize them in timestamp order, which is consistent to causality (Property 1), and then send them to other geo-locations (Alg. 3, line 10). Note that non-causally related updates coming from different partitions may have been timestamped with the same value. In this case, operations are concurrent and Eunomia can process them in any order.

Heartbeats. If a partition p_n does not receive an update for a fixed period of time, it will send a heartbeat including its current time to Eunomia (Alg. 2, lines 10–12). Thus, even if a partition p_n receives updates at a slower pace than others, it will not slow down the processing of other partitions updates at Eunomia. When Eunomia receives a heartbeat from p_n, it simply updates its entry in the *PartitionTime* vector (Alg. 3, line 6).

Hybrid Clocks. Our protocol combines logical and physical time. Although Eunomia could simply use logical clocks and still be correct, the rate at which clocks from different partitions progress would depend on the rate in which partitions receive update requests. This may cause Eunomia to process local updates in a slower pace and thus increase remote visibility latencies, as the stable time is set to the smallest timestamp received among all partitions. Differently, physical clocks naturally progress at similar rates independently of the workload characterization. This fact—previously exploited by [24, 10]—makes stabilization procedures resilient to skewed load distribution. Unfortunately, physical clocks do not progress exactly at the same rate, forcing protocols to wait for clocks to catch up in some situations in order to ensure correctness [23, 24, 10, 25]. The logical part of the hybrid clock makes the protocol resilient to clock skew by avoiding artificial delays due to clock synchronization uncertainties [30]. Briefly, if a partition p_n receives an update request with $Clock_c > Clock_n$, instead of waiting until $Clock_n > Clock_c$ to ensure correctness, the logical part of the hybrid clock ($MaxTs_n$) is moved forward. Then, when a partition p_n receives an update from any client, if the physical part $Clock_n$ is still behind the logical ($MaxTs_n$), the update is tagged with $MaxTs_n + 1$ in order to ensure clock monotonicity and thus guarantee Property 2. The interested reader can find the correctness proof of the algorithm in [29].

3.2 Resilience to Stragglers

A straggler is a partition that, due to a transient lack of network or processing resources, experiences delays in contacting other system components. Naturally, stragglers do not affect only Eunomia, but affect any system that attempts to provide the same guaranties. Here, we discuss how Eunomia differs from other solutions when coping with stragglers (later in §7.2.3, we report on experiments with stragglers). We distinguish delays that affect the communication between distinct datacenters (*inter-dc stragglers*) and delays that affect the interaction of components inside the same datacenter (*intra-dc stragglers*). We expect the former to be more frequent than the latter [11, 26].

Inter-dc stragglers have a similar impact on every system, no matter it is sequencer-based or stabilization-based (Eunomia, GentleRain [24], Cure [10]). The reason is that inter-dc disturbances affect the transmission of the data and, therefore, delays the visibility of updates in a way that is orthogonal to the metadata scheme used.

Intra-dc stragglers are more interesting, because they affect different approaches in different ways. In a sequencer-based approach, the straggler experiences delays when contacting the sequencer, which happens before the update takes place. Therefore, intra-dc stragglers affect local clients (because sequencer operation is in client's critical path) but have no effect on the remote visibility of updates from healthy partitions. Conversely, in stabilization-based approaches, local clients

are shielded from the instability (because stabilization is performed in the background) but the remote visibility of updates from healthy partitions of the straggler's datacenter is affected (because only stable updates are propagated/applied and the contribution of all partitions is required to achieve stability). Although there is a trade-off, given that there is evidence that an increase in the user-perceived latency may translate into concrete revenue loss [40], we argue that stragglers may affect more sequencer-based approaches.

3.3 Fault-Tolerance

In the description above, for simplicity, we have described the Eunomia service as if implemented by a single non-replicated server. Naturally, as any other service in a datacenter, Eunomia must be made fault-tolerant. In fact, if Eunomia fails, the site stabilization procedure stops, and thus, local updates can no longer be propagated to other geo-locations. In order to avoid such limitation, we now propose a fault-tolerant version of Eunomia. Note that we disregard failures in datacenters, as the problem of making data services fault-tolerant has been widely studied and is orthogonal to our work.

In this new version, Eunomia is composed by a set of *Replicas*. Algorithm 4 shows the behaviour of a replica e_f of the fault-tolerant Eunomia service. We assume the initial set of Eunomia replicas is common knowledge: every replica knows every other replica and every partition knows the full set of replicas. Partitions send operations and heartbeats (Alg. 2, lines 8 and 12 respectively) to the whole set of Eunomia replicas. The correctness of the algorithm requires the communication between partitions and Eunomia replicas to satisfy the *prefix-property* [38]: an Eunomia replica r_f that holds an update u_j originating at p_n also holds any other update u_i originating at p_n such that $u_i.ts < u_j.ts$. This property can be ensured with inexpensive protocols that offer only *at-least-once delivery*. Stronger properties, such as inter-partition order or exactly-once delivery are not required to enforce the prefix-property. Our implementation achieves the prefix-property by having each partition to keep track of the latest timestamp acknowledged by each of the Eunomia replicas in a vector denoted as Ack_n. Thus, to each Eunomia replica e_f, a partition p_n sends not only the lastest update but the set of updates including all updates u_j such that $u_j.ts > Ack_n[f]$. Upon receiving a new batch of updates *Batch* (Alg. 4, lines 1–5), e_f process it—in timestamp order—filtering out those updates already seen, and updating both Ops_f and $PartitionTime_f$ accordingly with the timestamps of the unseen updates. After processing *Batch*, e_f acknowledges p_n including the greatest timestamp observed from updates originating at p_n ($PartitionTime_f[p_n]$). This algorithm is

Algorithm 4 Operations at Eunomia replica e_f

```
1:  function NEW_BATCH(Batch, p_n)
2:      for all u_j ∈ Batch, PartitionTime_f[p_n] < u_j.ts do
3:          PartitionTime_f[p_n] ← u_j.ts
4:          Ops_f ← Ops_f ∪ u_j
5:      send ACK(PartitionTime_f[p_n]) to p_n
6:  function PROCESS_STABLE                          ▷ Every θ time
7:      if Leader_f == e_f then
8:          StableTime ← MIN(PartitionTime_f)
9:          StableOps ← FIND_STABLE(Ops_f, StableTime)
10:         PROCESS(StableOps)
11:         Ops_f ← Ops_f \ StableOps
12:         send STABLE(StableTime) to Replicas_f \ {e_f}
13: function STABLE(StableTime)
14:     StableOps ← FIND_STABLE(Ops_f, StableTime)
15:     Ops_f ← Ops_f \ StableOps
16:     for all p_n ∈ PartitionTime_f do
17:         PartitionTime_f[p_n] ← MAX(PartitionTime_f[p_n], StableTime)
18: function NEW_LEADER(e_g)
19:     Leader_f ← e_g
```

resilient to message lost and unordered delivery. Nevertheless, it adds redundancy, as replicas may receive the same update multiple times. §5 proposes a set of optimizations that aim to reduce this overhead.

In addition, to avoid unnecessary redundancy when exchanging metadata among datacenters, a leader replica is elected to propagate this information. The existence of a unique leader is not required for the correctness of the algorithm; it is simply a mechanism to save network resources. Thus, any leader election protocol designed for asynchronous systems (such as Ω [20]) can be plugged into our implementation. A change in the leadership is notified to a replica e_f through the NEW_LEADER function (Alg. 4, line 19). The notion of a leader is used to optimize the service's operation as follows. When the PROCESS_STABLE event is triggered, only the leader replica computes the new stable time and processes stable operations (Alg. 4, lines 7–10). Then, after operations have been processed, the leader sends the recently computed *StableTime* to the remaining replicas (Alg. 4, line 12). When replica e_f receives the new stable time, it removes the operations already known to be stable from its pending set of operations, since it is certain that those operations have been already processed (Alg. 4, lines 14–15).

4 Supporting Geo-replication

In our previous protocol, we have shown how to unobtrusively timestamp local updates in a partial order consistent with causality. In this section, we complete the protocol with the necessary mechanisms to ensure that remote updates—coming from other datacenters—are made visible locally without violating causality. Our solution resembles protocols implemented by other causally consistent geo-replicated storage systems [12,

M	Number of datacenters
$VClock_c$	Client c vector (M entries)
p_n^m	Partition n at datacenter m
r_m	Receiver at datacenter m
$SiteTime_m$	Applied updates vector at r_m
$Queue_m$	Queues of pending updates at r_m
$u_j.vts$	Update u_j timestamp vector (M entries)

Table 2: Notation used in the geo-replicated protocol extension.

44]. We assume a total of M datacenters, each of them replicating the full set of objects. Each datacenter uses the Eunomia service and thus propagates local updates in a total order consistent to causal consistency.

Apart from the Eunomia service, each datacenter is extended with a *receiver*. This component coordinates the execution of remote updates. Thus, it receives remote updates coming from remote Eunomia services (as a result of PROCESS_STABLE), and forwards them to the local datacenter partitions when its causal dependencies are satisfied. Standard replication techniques [43, 33, 13, 39] can be employed to make receivers robust to failures, as otherwise they represent a single point of failure.

In order to simplify the presentation, our pseudocode assumes FIFO links between each Eunomia service and the receivers. Nevertheless, this assumption can be easily dropped if the Eunomia service includes on every message send to a receiver, no only the latest update but all previous updates that have not been acknowledge (by the receiver) yet. This mechanism, which is similar to the one described in §3.3, preserves the prefix-property, and therefore tolerates message lost and unordered delivery.

We now explain how the metadata is enriched and the changes we need to apply to our previous algorithms. Table 2 summarizes the notation used in this section.

Updates are now tagged with a vector with an entry per datacenter, capturing inter-datacenter dependencies. The client clock is consequently also extended to a vector ($VClock_c$). We could easily adapt our protocols to use a single scalar, as in [24]. Nevertheless, vector clocks make a more efficient tracking of causal dependencies introducing no false dependencies across datacenters, which reduces the update visibility latency, at the cost of slightly increasing the storage and computation overhead. This overhead, unlike in [10], is negligible in our protocol as Eunomia allows for trivial dependency checking procedures. Note that the lower-bound update visibility latency for a system relying on vector clocks is the latency between the originator of the update and the remote datacenter, while with a single scalar it is the latency to the farthest datacenter.

Update. When a client c issues an update operation, it piggybacks its $VClock_c$ summarizing both local and remote dependencies. A partition p_n computes u_j vector timestamp ($u_j.vts$) as follows. First, the local entry of the vector $u_j.vts[m]$ is computed as the maximum between $Clock_n$, $MaxTs_n + 1$ and $VClock_c[m] + 1$, similarly to Algorithm 2, line 5. This permits Eunomia to still be able to causally order local updates based on $u_j.vts[m]$. Second, the remaining entries (remote datacenter entries) are assigned to their sibling entries in $VClock_c$. When the operation is completed, p_n returns $u_j.vts$ to the client who can directly substitutes its $VClock_c$ since $u_j.vts$ is known to be strictly greater than $VClock_c$.

Algorithm 5 Operations at r_m

```
1:  function NEW_UPDATE(u_j, k)
2:      Queue_m[k] ← [Queue_m[k]|u_j]                          ▷ add to tail
3:  function CHECK_PENDING                                     ▷ Every ρ time
4:      ⟨Queue_m, SiteTime_m⟩ ← FLUSH(1, Queue_m, SiteTime_m)
5:  function FLUSH(k, Queue_m, SiteTime_m)
6:      if k > M then
7:          return ⟨Queue_m, SiteTime_m⟩
8:      else if k = m then
9:          FLUSH(k + 1, Queue_m, SiteTime_m)
10:     else
11:         u_j ←HEAD(Queue_m[k])
12:         if ∀d ∈ M \ {m,k}, SiteTime_m[d] ≥ u_j.vts[d] then
13:             p_n^m ← RESPONSIBLE(u_j.key)
14:             send APPLY(u_j) to p_n^m
15:             receive ok from p_n^m
16:             SiteTime_m[k] ← u_j.vts[k]
17:             POP(Queue_m[k])
18:             FLUSH(1, Queue_m, SiteTime_m)
19:         else
20:             FLUSH(k + 1, Queue_m, SiteTime_m)
```

Read. Read operations execute as in Algorithms 1 and 2. The only difference is that the returned timestamp is a vector instead of a scalar. Thus, in order to update $VClock_c$, a client c applies the MAX operation per entry.

Update Propagation. The site stabilization procedure proceeds as before, totally ordering local updates based on the local entry of their vector timestamp ($u.vts[m]$). Eunomia propagates local updates to remote datacenters in $u.vts[m]$ order. Each update piggybacks its $u.vts$.

Remote Update Visibility. Algorithm 5 details receivers' operation. A receiver r_m maintains two important pieces of state: a queue of pending updates per remote datacenter ($Queue_m[k]$), and a vector with an entry per remote datacenter ($SiteTime_m$) indicating the latest update operation locally applied from each of the remote datacenters. When r_m receives a remote update u_j coming from datacenter k, it simply adds it to its corresponding queue. Periodically, r_m triggers the CHECK_PENDING function (Algorithm 5 lines 4 and 18). This function ensures, by means of the tail recursive FLUSH function, that no pending operation is left unexecuted. Two conditions have to be satisfied before sending an update u_j to the local partitions: *(i)* all previously received updates coming from k have already been applied locally; and *(ii)* u_j dependencies, which are subsumed in $u_j.vts$, are visible locally. Both conditions are trivially checked by relying on the information subsumed in $Queue_m$ and $SiteTime_m$. When a pending operation

u_j originating at k is applied, both $Queue_m[k]$ and $SiteTime_m[k]$ are updated consequently.

5 Optimizations

We propose a set of optimizations that aim at enabling Eunomia to handle even heavier loads.

Communication Patterns. Eunomia constantly receives operations and heartbeats from partitions. This is an all-to-one communication schema and, if the number of partitions is large, it may not scale in practice. In order to overcome this problem and efficiently manage a large number of partitions, two simple techniques have been used: *(i)* build a propagation tree among partition servers; and *(ii)* batch operations at partitions, and propagate them to Eunomia only periodically. Both techniques are able to reduce the number of messages received by Eunomia per unit of time at the cost of a slight increase in the stabilization time.

Separation of Data and Metadata. In the protocols described before, partitions send updates (including the update value) to the Eunomia service, which is responsible for eventually propagating them to remote datacenters. This can limit the maximum load that Eunomia can handle and become a bottleneck due to the potentially large amount of data that has to be handled. In order to overcome this limitation, we decouple data from metadata.

In our prototype, for each update operation, partitions generate a unique update identifier ($u.id$), composed of the local entry of the update vector timestamp ($u.vts[m]$) and the object identifier (*Key*). We avoid sending the value of the update to Eunomia. Instead, partitions only send the unique identifier $u.id$ together with the partition id (p_n^m). Eunomia is then only responsible for handling and propagating these lightweight identifiers, while the partitions itself are responsible for propagating (with no order delivery constraints) the update values together with $u.id$ to its sibling partitions in other datacenters. A receiver r_m proceeds as before, but a partition p_n^m can only install the remote operation once it has received both the data and the metadata. This technique slightly increases the computation overhead at partitions, but it allows Eunomia to handle a significantly heavier load independently of update payloads.

6 Implementation

The Eunomia service is approximately 200 lines of `C++` code[2]. We integrated it with a version of Riak KV [6], a very popular [3] weakly consistent datastore used by many companies offering cloud-based services including Uber [2], bet365 [2] and Rovio [7]. Its integration consisted of 100 lines of Erlang code. We expect that integrating Eunomia into other popular NoSQL datastores such as Cassandra [32] would require a comparable effort as these datastores are architecturally very similar.

Since Riak KV is implemented in Erlang, we first attempted to build Eunomia using the `Erlang/OTP` framework, but unfortunately we reached a bottleneck in our early experiments. Note that for Eunomia to work, we need to store a potentially large number of updates, coming from all logical partitions composing a datacenter, and periodically traverse them in timestamp order when a new stable time is computed. Inserting and traversing this (ordered) set of updates was limiting the maximum load that Eunomia could handle, as accessing an item in a list using the built-in `Erlang` data type requires linear time with the number of elements in the list. The `C++` version does not suffer from these limitations.

At its core, Eunomia is uses a *red-black tree* [28], a self-balancing binary search tree optimized for insertions and deletions, which guarantees logarithmic search, insert and delete cost, and linear in-order traversal cost, a critical operation for Eunomia. In our case, the *red-black tree* turned out to be more efficient than other self-balancing binary search trees such as AVL trees [8].

Furthermore, in order to fully explore the capacities of Eunomia, we have integrated Eunomia with a causally consistent geo-replicated datastore implementing the protocol presented in §3 and §4. Our prototype, namely EunomiaKV[3], is built as a variant of Riak KV [6], and includes the optimizations discussed in §5. Since the open source version of Riak KV does not support replication across Riak KV clusters, we have also augmented it with geo-replication support.

7 Evaluation

Our main goal with the evaluation is to show that Eunomia does not suffer from the limitations of the competing approaches. Therefore, we compare Eunomia both with approaches based on sequencers and based on global stabilization. We recall that the main disadvantage of sequencers is to throttle throughput, because they operate in the critical path of local clients. Therefore, we aim at showing that Eunomia does not compromise the intra-datacenter concurrency and can reach higher throughput that sequencer-based approaches. Conversely, the expensiveness of the global stabilization approach forces designers to favour either throughput or remote update visibility latencies. Thus, we also aim at showing that

[2] Available at https://github.com/chathurilanchana/C-Stabilizer/tree/master/src

[3] Available at https://github.com/chathurilanchana/riak_kv/tree/causal-dev-multidc-nostat-nostraggler

Eunomia optimizes both.

Experimental Setup. The experimental test-bed used is a private cloud composed by a set of virtual machines deployed over 20 physical machines (8 cores and 40 GB of RAM) connected via a Gigabit switch. Each VM, which runs Ubuntu 14.04, and is equipped with 2 (virtual) cores, 10GB disk and 9GB of RAM memory; is allocated in a different physical machine. Before running each experiment, physical clocks are synchronized using the NTP protocol [5] through a near NTP server.

Workload Generator. Each client VM runs its own instance of a custom version of Basho Bench [1], a benchmarking tool. For each experiment, we deploy as many client instances as possible without overloading the system. Latencies across datacenters are emulated using `netem` [4], a Linux network emulator tool. The values used in operations are a fixed binary of 100 bytes. Our key-space is composed by 100k keys. The ratio of reads and updates is varied depending on the experiment. Before running the experiments, we populate the database. Each experiment runs for more than 6 minutes. In our results, the first and the last minute of each experiment is ignored to avoid experimental artifacts.

7.1 Eunomia Throughput

We report on a number of experiments that aim at: *(i)* measuring the maximum load that our efficient implementation of Eunomia can handle, varying the number of partitions connected to it; and *(ii)* assessing how replication and failures affect Eunomia's performance.

For comparison, these experiments also compute the throughput upper-bound of a traditional sequencer. Our implementation of a sequencer mimics traditional implementations [44, 12]. In every update operation, datacenter partitions synchronously request a monotonically increasing number to the sequencer before returning to the client. We have also implemented a fault-tolerant version of the sequencer based on chain replication [43]: Replicas are organized in a chain. Partitions send requests to the head of the chain. Requests traverse the chain up to the tail. When the tail receives a request, it replies back to the partition, which returns to the client.

In order to stretch as much as possible the implementation, circumventing potential bottlenecks in the system, we directly connect clients to Eunomia, bypassing the data store. Thus, each client acts as a partition in a multi-server datacenter. This allowed us to emulate very large datacenters, with much more servers than the ones that were at our disposal for these experiments, and overload Eunomia in a way that would be otherwise impossible with our testbed.

Throughput Upper-Bound. We first compare the non fault-tolerant version of the Eunomia against a non fault-tolerant implementation of a sequencer. In these experiments, partitions batch updates and only send them to Eunomia after 1ms.

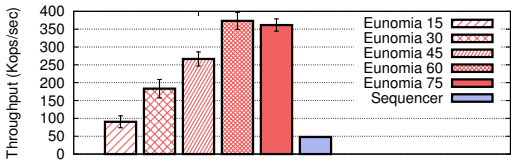

Figure 2: Maximum throughput achieved by Eunomia and an implementation of a sequencer. We vary the number of partitions that propagate operations to Eunomia.

Figure 2 plots the maximum throughput achieved by both services. As results show, Eunomia maximum throughput is reached when having 60 partitions issuing operations eagerly (with zero waiting time between operations). We observe that Eunomia is able to handle almost an order of magnitude more operations per second than a sequencer (more precisely, 7.7 times more operations, exceeding 370kops while the sequencer is saturated at 48kops). Considering that according to our experiments, a single machine in a Riak cluster is able to handle approximately 3kops per second, results confirm that sequencers limit intra-datacenter concurrency and can easily become a bottleneck for medium size clusters (i.e, for clusters above 150 machines, the sequencer would be the limiting factor of system performance), even assuming a read dominant (9:1) workload, a common workload for internet-based services. Nevertheless, under the same workload assumptions, more than a thousand machines could be used before saturating Eunomia.

Another advantage of Eunomia in comparison to sequencers is that batching is not in client's critical path. Thus, Eunomia's throughput can be further stretched by increasing the batching time (while slightly increasing the remote update visibility latency). Such stretching cannot be easily achieved with sequencers, as any attempt to batch requests at the sequencer blocks clients.

A final conclusion can be drawn from this experiment: Eunomia maximum capacity does not significantly varies with the number of partitions. Although we hit the maximum load with 60 partitions, we run an extra experiment increasing the number to 75 to see if this negatively impacts Eunomia's performance and we observed a very similar throughput. The reason is that the bottleneck of our Eunomia implementation is the propagation to other geo-locations rather than the handling of operations. This confirms that the use of a red-black self-balancing search tree was an appropriate design choice.

Fault-Tolerance Overhead. In the following experiments we measure the overhead introduced by the fault-tolerant version of Eunomia. Figure 3 compares the maximum throughput achievable by Eunomia when increas-

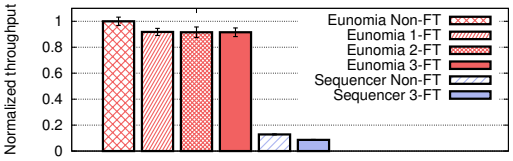

Figure 3: Maximum throughput achieved by a fault-tolerant version of Eunomia and sequencers. Non-FT denotes non fault-tolerant versions while 1-, 2-, and 3-FT denote fault-tolerant versions with 1, 2, and 3 replicas

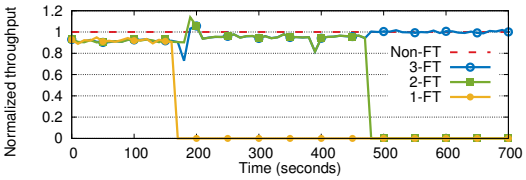

Figure 4: Impact of failures in Eunomia.

ing the number of replicas up to three. For completeness, the plot also includes the throughput for a non fault-tolerant sequencer and its fault-tolerant version with a chain of three replicas. We normalized the throughput against the non fault-tolerant version of Eunomia. As results show, the fault-tolerant version of Eunomia only adds a small overhead (roughly 9% penalty) independently on the number of replicas. We expect this overhead to increase as the number of replicas increases, but we consider three replicas to be a realistic number. On the other hand, adding fault-tolerance to the sequencer version adds a penalty of almost 33%, thus being more expensive proportionally. The reason for this difference is that Eunomia replicas do not need to coordinate as their results are independent of relative order of inputs, while sequencer replicas need to coordinate to avoid providing inconsistent sequence numbers.

Impact of Failures. Finally, we experiment injecting failures into Eunomia. Figure 4 plots the results normalized against the non fault-tolerant Eunomia (Non-FT line). We compare Eunomia with one, two, and three replicas. As the figure shows, at the beginning of the experiment, all three versions produce similar throughput (confirming Figure 3 results). After 160 seconds, we crash one replica. As expected, the throughput of 1-FT drops to zero since no more replicas are available. The rest of the versions (2-FT and 3-FT), after a short period of fluctuation, slightly increase their throughput up to 95% of the Non-FT version throughput. Finally, after 210 more seconds (at 470), we crash a second replica. Again, the 2-FT as expected drops its throughput to zero. The 3-FT version, this time almost without fluctuations, is capable of achieving the maximum throughput in few seconds. These results demonstrate that failures have negligible impact in Eunomia. Note that sometimes the multi-replica version go beyond the Non-FT line because the Non-FT line is drawn by computing the average.

7.2 Experiments with Geo-Replication

We now report on a set of experiments offering evidence that a causally consistent geo-replicated datastore built using Eunomia is capable of providing higher throughput and better quality-of-service than previous solutions that avoid the use of local sequencers.

For this purpose, we have implemented GentleRain [24] and a variation of it that uses vector clocks instead of a single scalar to enforce causal consistency across geo-locations. The latter resembles the causally consistency protocol implemented by Cure [10]. Both approaches are sequencer-free that rely on a global stabilization procedure in order to apply operations in remote locations consistently with causality. For this, sibling partitions across datacenters have to periodically send heartbeats, and each partition within a datacenter has to periodically compute its local-datacenter stable time. In our experiments, we set the time interval of this events to 10ms and 5ms respectively unless otherwise specified. These values are in consonance to the ones used by the authors of these works. For a fair comparison, both approaches are implemented using the EunomiaKV's codebase and thus integrated with Riak KV.

In most of our experiments, we deploy 3 datacenters, each of them composed of 8 logical partitions balanced across 3 servers. The emulated round-trip-times across datacenters are 80ms between dc_1 and both dc_2 and dc_3, and 160ms between dc_2 and dc_3. These latencies are approximately the round-trip-times between Virginia, Oregon and Ireland regions of Amazon EC2.

7.2.1 Throughput

In the following experiments, we measure the throughput provided by EunomiaKV, GentleRain, Cure, and an eventually consistent multi-cluster version of Riak KV. Note that the latter does not enforce causality, and thus partitions install remote updates as soon as they are received. Therefore, the comparison of EunomiaKV with Riak KV allows to assess the overhead that enforcing causal consistency adds when using our approach. As discussed below, this overhead is very small.

We experiment with both uniform and power-law key distributions, denoted with U and P respectively in Figure 5. For each of them, we vary the read:write ratio (99:1, 90:10, 75:25 and 50:50). These ratios are representative of real large internet-based services workloads. As shown by Figure 5, the throughput of all solutions decreases as we increase the percentage of updates. Nevertheless, EunomiaKV always provides a comparable throughput to eventual consistency. Precisely, on average, EunomiaKV only drops 4.7% of throughput, being extremely close in read intensive workloads (1% drop). Differently, GentleRain and Cure are al-

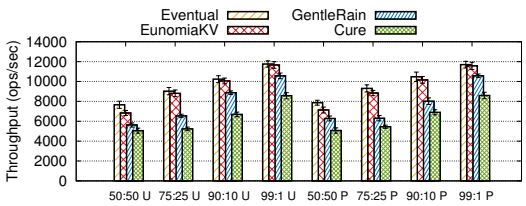

Figure 5: Throughput comparison between EunomiaKV and state-of-the-art sequencer-free solutions.

ways significantly below both eventual consistency (and EunomiaKV). This is due to the cost of the global stabilization procedure. Note that the throughput difference between GentleRain and Cure is caused by the overhead introduced by the metadata enrichment procedure of the latter (as discussed in §4). Based on our experiments, it is possible to conclude that the absolute number of updates per unit of time is the factor that has the largest impact in EunomiaKV (rather than key contention).

7.2.2 Remote Update Visibility

To compare the quality-of-service that can be provided by EunomiaKV, GentleRain, and Cure, we measure remote update visibility latencies. In EunomiaKV, we measure the time interval between the data arrival and the instant in which the update is executed at the responsible partition. Note that, for an update to be applied, a datacenter needs to have access to the metadata (in our case, provided by Eunomia) and check that all of its causal dependencies have also been previously applied locally. In our implementation, partitions ship updates immediately to remote datacenters. Therefore, we have observed that updates are always locally available to be applied by the time metadata indicates that its causal dependencies are already satisfied locally. Although other strategies could be used to ship the payload of the updates, this has a crucial advantage for the evaluation of Eunomia: under this deployment the update visibility latency is exclusively influenced by the performance of the metadata management strategy, including the stabilization delay incurred at the originating datacenter.

On the other hand, for GentleRain and Cure, we measure the time interval between the arrival of the remote operation to the partition and when the global stabilization procedure allows its visibility. Note that all values presented in the figures already factor-out the network latencies among datacenters (which are the same for all protocols); thus numbers capture only the artificial artifacts inherent to the different approaches.

Figure 6 (left plot) shows the cumulative distribution of the latency before updates originating at dc_1 become visible at dc_2. We observe that EunomiaKV offers, by far, the best remote update visibility latency. In fact, for almost 95% of remote updates, EunomiaKV only adds 15ms extra delay. On the other hand, with GentleRain

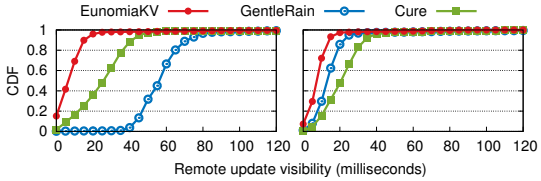

Figure 6: Left: from dc_1 to dc_2 (40ms trip-time). Right: from dc_2 to dc_3 (80ms trip-time).

and Cure the extra delay goes up to 80ms and 45ms respectively for the same amount of updates. Unsurprisingly, GentleRain extra delay is larger than Cure's because of the amount of false dependencies added when aggregating causal dependencies into a single scalar. In fact, GentleRain is not capable of making updates visible without adding 40ms of extra delay. Again, the scalar is the cause of this phenomenon since the minimum delay will not depend on the originator of the update but on the travel time to the furthest datacenter. This confirms the rationale presented in the discussion of §4.

Although both Cure and EunomiaKV rely on vector clocks for tracking causal dependencies, EunomiaKV is able to offer better remote update latencies because partitions are less overloaded since checking dependencies in EunomiaKV is trivial due to Eunomia. Note that in EunomiaKV, even 20% of remote updates are made visible without any extra delay, and thus reaching the optimal remote update visibility latency.

Finally, in order to isolate the impact of GentleRain's global stabilization procedure independently of the metadata size, we measure the remote update visibility latency at dc_3 for updates originating at dc_2. As one can observe in Figure 6 (right plot), GentleRain exhibits better remote update latencies than Cure but still worse than EunomiaKV. In this setting, vector clocks does not help reducing latencies. Thus, the gap between Cure and GentleRain is exclusively due to the storage and computational overhead caused by vector clocks. Furthermore, the fact that EunomiaKV still provides better latencies is, once again, an empirical evidence that global stabilization procedures are expensive in practice.

7.2.3 Impact of Stragglers

Finally, we assess the impact of stragglers in EunomiaKV and its competitors. Due to lack of space, and given that they provide no significant insight, we omit experimental results for inter-dc stragglers.

In these experiments, we use three datacenters (same setup of previous experiments) that run under optimal conditions during 1 minute. Then, during the second minute, we introduce a straggler. This is a partition of dc_3 that communicates abnormally with its local sequencer or Eunomia service. In Eunomia, instead of communicating every millisecond (as every other parti-

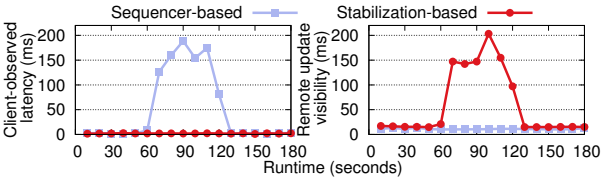

Figure 7: Client-observed latency (measured as the averaged latency observed by clients of the straggling datacenter dc_3) vs. remote update visibility latency (measured at dc_2 for updates originating at dc_3) tradeoff disclosed by intra-dc stragglers (one second straggling interval).

tion), the straggler contacts Eunomia less frequently. In the sequencer-based system, a similar delay (on average) is introduced when the straggler partition contacts the sequencer. We have experimented with three straggling intervals: 10, 100 and 1000ms, all exhibiting similar patterns. Figure 7 shows results for a 1 second straggling interval, as it is the most striking result. After the straggling period, the partition gets healed.

As expected (§3.2), intra-dc stragglers do not affect the remote visibility of updates in sequencer-based approaches but clients notice a significant increase in latency. In contrast, stabilization-based approaches are capable of shielding clients from stragglers and the cost of increasing the remote visibility of updates. Note that the stabilization-based results were obtained with Eunomia, but GentleRain and Cure exhibit a similar behaviour.

8 Related work

The support for causal consistency can already be found in early pioneer works in distributed systems, such as Bayou [38, 42], Lazy Replication [31], and the ISIS [18] toolkit. Recently, and tackling scalability challenges close to ours, multiple weakly consistent geo-replicated data stores implementing causal consistency across geo-locations have been proposed. We group them into two categories: *(i)* sequencer-based solutions [12, 44, 21]; *(ii)* and sequencer-free solutions [35, 22, 36, 24, 10].

Sequencer-based. These solutions rely on a sequencer per datacenter to enforce causal consistency. The sequencer totally orders local updates, in a causally consistent manner, and propagate them to remote locations. This design centralizes, thus simplifying, the implementation of causal consistency. Nevertheless, the use of synchronous sequencers limits the intra-datacenter concurrency, as demonstrated by our experiments. SwiftCloud [44] and ChainReaction [12] rely on a vector clock with an entry per datacenter to track causal dependencies, similarly to EunomiaKV. Practi [21], on the contrary, uses a single scalar and a sophisticated mechanism of invalidations. Similar to EunomiaKV, Practi separates the propagation of data and metadata. This and the concept of *imprecise* invalidations optimize Practi for partial replication, a setting that has not yet been explored in this work. We have shown that sequencers may get easily saturated for medium-size clusters, while Eunomia is able to handle much heavier loads (up to 7.7 times more).

Sequencer-free. There have been two major trends in this category: *(i)* solutions that rely on explicit dependency check messages [35, 22, 36]; and *(ii)* solutions based on global stabilization procedures [24, 10].

COPS [35] and Eiger [36] finely track dependencies for each individual data item allowing full concurrency within a datacenter. Updates are tagged with a list of dependencies. When a datacenter receives a remote update, it needs to explicitly check each dependency. This process is expensive and limits systems performance [24] due to the large amount of metadata managed. Orbe [22] aggregates dependencies belonging to the same logical partition into a scalar, only partially solving the problem.

Alternatives that use less metadata rely on a background global stabilization procedure [24, 10]. This procedure equips partitions with sufficient information to safely execute remote updates consistently with causality. Thus, these solutions manage to aggregate the metadata as sequencer-based solutions without relying on an actual sequencer. As our extensive evaluation has empirically demonstrated, global stabilization procedures are expensive in practice, forcing designers to favour either throughput [24] or remote visibility latency [10]. Our evaluation shows that EunomiaKV does not force designers to sacrifice any of the two, exhibiting significantly better throughput and remote visibility latencies than Cure and GentleRain respectively.

9 Conclusions

We have presented a novel approach for building causally consistent geo-replicated data stores. Our solution relies on Eunomia, a new service that abstracts the internal complexity of datacenters, a key feature to reduce the cost of causal consistency. Unlike sequencers, Eunomia does not limit the intra-datacenter concurrency by performing an unobtrusive ordering of updates. Our evaluation shows that Eunomia can handle very heavy loads without becoming a performance bottleneck (up to 7.7 times more operations per second than sequencers). Experiments also show that EunomiaKV (a causally consistent geo-replicated protocol that integrates Eunomia), unlike previous systems, permits optimizing both throughput and remote update visibility latency simultaneously. In fact, results have shown that EunomiaKV only adds a slight throughput overhead (4.7% on average) and exceptionally small artificial remote visibility delays when compared to an eventually consistent data store that makes no attempt to enforce causality.

Acknowledgments

We would like to thank our shepherd Chunqiang (CQ) Tang, Kuganesan Srijeyanthan, and anonymous reviewers for their comments and suggestions. This research has been supported in part by the Horizon 2020 project 732 505 LightKone, by the Erasmus Mundus Doctorate Programme under Grant Agreement No. 2012-0030, by the European Master in Distributed Computing (EMDC), and by FCT through projects PTDC/ EEI-SCR/ 1741/ 2014 (Abyss) and UID/ CEC/ 50021/ 2013.

References

[1] Basho Bench. http://github.com/basho/basho_bench.

[2] bet365. http://www.bet365.com/.

[3] Customers of Riak KV. http://basho.com/about/customers/.

[4] Netem. http://www.linuxfoundation.org/collaborate/workgroups/networking/netem.

[5] The network time protocol. http://www.ntp.org.

[6] Riak KV. https://github.com/basho/riak_kv.

[7] Rovio. http://www.rovio.com/.

[8] ADELSON-VELSKII, M., AND LANDIS, E. An algorithm for the organization of information. Tech. rep., DTIC Document, 1963.

[9] AHAMAD, M., NEIGER, G., BURNS, J. E., KOHLI, P., AND HUTTO, P. W. Causal memory: definitions, implementation, and programming. *Distributed Computing 9*, 1 (1995), 37–49.

[10] AKKOORATH, D., TOMSIC, A., BRAVO, M., LI, Z., CRAIN, T., BIENIUSA, A., PREGUIÇA, N., AND SHAPIRO, M. Cure: Strong semantics meets high availability and low latency. In *Proceedings of the International Conference on Distributed Computing Systems* (Osaka, Japan, 2016).

[11] AL-FARES, M., LOUKISSAS, A., AND VAHDAT, A. A scalable, commodity data center network architecture. In *Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication* (Seattle, WA, USA, 2008), pp. 63–74.

[12] ALMEIDA, S., LEITÃO, J. A., AND RODRIGUES, L. Chainreaction: A causal+ consistent datastore based on chain replication. In *Proceedings of the 8th ACM European Conference on Computer Systems* (Prague, Czech Republic, 2013).

[13] ALSBERG, P. A., AND DAY, J. D. A principle for resilient sharing of distributed resources. In *Proceedings of the 2nd International Conference on Software Engineering* (San Francisco, CA, USA, 1976).

[14] ATTIYA, H., ELLEN, F., AND MORRISON, A. Limitations of highly-available eventually-consistent data stores. In *Proceedings of the ACM Symposium on Principles of Distributed Computing* (Donostia-San Sebastián, Spain, 2015).

[15] BAILIS, P., DAVIDSON, A., FEKETE, A., GHODSI, A., HELLERSTEIN, J. M., AND STOICA, I. Highly available transactions: Virtues and limitations. *Proc. VLDB Endow. 7*, 3 (Nov. 2013), 181–192.

[16] BAILIS, P., FEKETE, A., GHODSI, A., HELLERSTEIN, J. M., AND STOICA, I. The potential dangers of causal consistency and an explicit solution. In *Proceedings of the ACM Symposium on Cloud Computing* (San Jose, California, 2012).

[17] BALEGAS, V., DUARTE, S., FERREIRA, C., RODRIGUES, R., PREGUIÇA, N., NAJAFZADEH, M., AND SHAPIRO, M. Putting consistency back into eventual consistency. In *Proceedings of the 10th ACM European Conference on Computer Systems* (Bordeaux, France, 2015).

[18] BIRMAN, K., SCHIPER, A., AND STEPHENSON, P. Lightweight causal and atomic group multicast. *ACM Trans. Comput. Syst. 9*, 3 (Aug. 1991).

[19] BRAVO, M., DIEGUES, N., ZENG, J., ROMANO, P., AND RODRIGUES, L. On the use of clocks to enforce consistency in the cloud. *IEEE Data Eng. Bull 38*, 1 (2015), 18–31.

[20] CHANDRA, T., HADZILACOS, V., AND TOUEG, S. The weakest failure detector for solving consensus. *J. ACM 43*, 4 (July 1996), 685–722.

[21] DAHLIN, M., GAO, L., NAYATE, A., VENKATARAMANA, A., YALAGANDULA, P., AND ZHENG, J. Practi replication. In *Proceedings of the 3rd Symposium on Networked Systems Design and Implementation* (San Jose, CA, USA, 2006).

[22] DU, J., ELNIKETY, S., ROY, A., AND ZWAENEPOEL, W. Orbe: Scalable causal consistency using dependency matrices and physical clocks. In *Proceedings of the ACM Symposium on Cloud Computing* (Santa Clara, CA, USA, 2013).

[23] DU, J., ELNIKETY, S., AND ZWAENEPOEL, W. Clock-si: Snapshot isolation for partitioned data stores using loosely synchronized clocks. In *Proceedings of the 32nd IEEE Symposium on Reliable Distributed Systems* (Braga, Portugal, 2013).

[24] DU, J., IORGULESCU, C., ROY, A., AND ZWAENEPOEL, W. Gentlerain: Cheap and scalable causal consistency with physical clocks. In *Proceedings of the ACM Symposium on Cloud Computing* (Seattle, WA, USA, 2014).

[25] DU, J., SCIASCIA, D., ELNIKETY, S., ZWAENEPOEL, W., AND PEDONE, F. Clock-RSM: Low-latency inter-datacenter state machine replication using loosely synchronized physical clocks. In *Proceedings of the 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks* (Atlanta, Georgia USA, 2014).

[26] GREENBERG, A., HAMILTON, J. R., JAIN, N., KANDULA, S., KIM, C., LAHIRI, P., MALTZ, D. A., PATEL, P., AND SENGUPTA, S. Vl2: A scalable and flexible data center network. In *Proceedings of the ACM SIGCOMM 2009 Conference on Data Communication* (Barcelona, Spain, 2009), pp. 51–62.

[27] GUERRAOUI, R., PAVLOVIC, M., AND SEREDINSCHI, D.-A. Trade-offs in replicated systems. *Data Engineering* (2016), 14.

[28] GUIBAS, L. J., AND SEDGEWICK, R. A dichromatic framework for balanced trees. In *Proceedings of the 54th IEEE Annual Symposium on Foundations of Computer Science* (Ann Arbor, Michigan, USA, 1978), pp. 8–21.

[29] GUNAWARDHANA, C., BRAVO, M., AND RODRIGUES, L. Unobtrusive deferred update stabilization for efficient geo-replication. *arXiv:1702.01786 [cs.DC]* (Feb. 2017).

[30] KULKARNI, S. S., DEMIRBAS, M., MADAPPA, D., AVVA, B., AND LEONE, M. Logical physical clocks. In *Proceedings of the 18th International Conference on Principles of Distributed Systems* (Cortina d'Ampezzo, Italy, 2014).

[31] LADIN, R., LISKOV, B., SHRIRA, L., AND GHEMAWAT, S. Providing high availability using lazy replication. *ACM Trans. Comput. Syst.* (1992).

[32] LAKSHMAN, A., AND MALIK, P. Cassandra: A decentralized structured storage system. *SIGOPS Oper. Syst. Rev. 44*, 2 (Apr. 2010), 35–40.

[33] LAMPORT, L. The part-time parliament. *ACM Trans. Comput. Syst. 16*, 2 (May 1998), 133–169.

[34] LI, C., PORTO, D., CLEMENT, A., GEHRKE, J., PREGUIÇA, N., AND RODRIGUES, R. Making geo-replicated systems fast as possible, consistent when necessary. In *Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation* (Hollywood, CA, USA, 2012), pp. 265–278.

[35] LLOYD, W., FREEDMAN, M. J., KAMINSKY, M., AND ANDERSEN, D. G. Don't settle for eventual: Scalable causal consistency for wide-area storage with cops. In *Proceedings of the 23rd ACM Symposium on Operating Systems Principles* (Cascais, Portugal, 2011).

[36] LLOYD, W., FREEDMAN, M. J., KAMINSKY, M., AND ANDERSEN, D. G. Stronger semantics for low-latency geo-replicated storage. In *Proceedings of the 10th Symposium on Networked Systems Design and Implementation* (Lombard, IL, USA, 2013).

[37] MAHAJAN, P., ALVISI, L., AND DAHLIN, M. Consistency, availability, and convergence. Tech. rep., University of Texas at Austin, 2011.

[38] PETERSEN, K., SPREITZER, M. J., TERRY, D. B., THEIMER, M. M., AND DEMERS, A. J. Flexible update propagation for weakly consistent replication. In *Proceedings of the 16th ACM Symposium on Operating Systems Principles* (Saint Malo, France, 1997).

[39] SCHNEIDER, F. B. Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Comput. Surv. 22*, 4 (Dec. 1990), 299–319.

[40] SCHURMAN, E., AND BRUTLAG, J. The user and business impact of server delays, additional bytes, and HTTP chunking in web search. In *Velocity Web Performance and Operations Conference* (San Jose, CA, USA, 2009).

[41] SOVRAN, Y., POWER, R., AGUILERA, M. K., AND LI, J. Transactional storage for geo-replicated systems. In *Proceedings of the 23rd ACM Symposium on Operating Systems Principles* (Cascais, Portugal, 2011).

[42] TERRY, D. B., DEMERS, A. J., PETERSEN, K., SPREITZER, M. J., THEIMER, M. M., AND WELCH, B. B. Session guarantees for weakly consistent replicated data. In *Proceedings of the 3rdInternational Conference on Parallel and Distributed Information Systems* (Austin, TX, USA, 1994).

[43] VAN RENESSE, R., AND SCHNEIDER, F. B. Chain replication for supporting high throughput and availability. In *Proceedings of the 6th symposium on Operating systems design and implementation* (San Francisco, CA, USA, 2004).

[44] ZAWIRSKI, M., PREGUIÇA, N., DUARTE, S., BIENIUSA, A., BALEGAS, V., AND SHAPIRO, M. Write fast, read in the past: Causal consistency for client-side applications. In *Proceedings of the annual ACM/IFIP/USENIX Middleware conference* (Vancouver, Canada, 2015).