Subscription Normalization for Effective Content-Based Messaging

K.R. Jayaram, Weihang Wang, and Patrick Eugster

Abstract—Efficient subscription summarization and event matching is key to the scalability of content-based publish/subscribe networks (CPSNs). Current summarization and event matching mechanisms based on subscription subsumption induce heavy event processing load on brokers degrading the performance of CPSNs especially under high rates of churn, i.e., addition, deletion, or modification of subscriptions. Yet, many modern CPS applications such as location-based services or algorithmic trading inherently rely on high frequency subscription changes. This paper describes Beretta, a dynamic CPSN which sustains high throughput and low event-propagation latencies even under a high frequency of subscription changes. Beretta leverages strong event typing and represents all subscriptions in a normalized form as combinations of value intervals and set inclusions without compromising on expressiveness. Beretta’s “split and subsume” broker algorithm reduces the complexity of matching an event from $O(KN)$ to $O(K\log N + |\text{result}|)$, with $N$ being the number of subscriptions for the event type and $K$ the number of its attributes. Event types and normalization are exploited to split subscriptions into predicates on individual event types and attributes and to efficiently regroup these in segment trees and hash maps which yield excellent subscription properties and support attribute-wise split filtering during event matching. Normalization enables the systematic introduction of parameters into subscriptions to support both parametric and structural updates. This paper also empirically demonstrates the performance improvements due to our techniques through realistic algorithmic trading and highway traffic monitoring benchmarks.

Index Terms—Subscription, summarization, subsumption, normalization, content-based, messaging

1 INTRODUCTION

By focusing on the exchanges among interacting parties rather than the parties themselves, the publish/subscribe paradigm [1] is an appealing candidate for building scalable networked distributed systems. This dynamic interaction culminates in content-based publish/subscribe (CPS), where subscriptions are based on event content rather than on channels or topics. Content-centric communication abstractions have been more recently investigated in the context of future Internet design [2], [3]; the advent of software-defined networking (SDN) supports the implementation of such models.

effectively and efficiently route published events to subscribers with corresponding interests, existing application-level CPS systems typically construct overlay networks called “content-based publish/subscribe networks” (CPSNs). A CPSN consists of several event routers—commonly referred to as brokers—that interconnect publishers and subscribers often making use of advertisements of publishers in addition to subscriptions to transmit events to subscribers. That is, advertisements and subscriptions are propagated down- and up-stream respectively to set up connections such as to ensure that there is a path for the propagation of an event from any publisher to any subscriber with a potentially matching subscription [1]. Current CPSNs however present a number of limitations in their core constituents:

- **Matching algorithms.** Matching events to subscriptions, which occurs at each broker in a CPSN, is key to the efficient routing of events. Many matching algorithms, however, have a time complexity of $O(N)$, $N$ being the number of subscriptions matched.

- **Subscription summaries.** Efficient subscription summarization is a key requirement for efficient routing of subscriptions and to avoid storing every subscription at every broker. Existing summarization algorithms, however, choose accuracy over efficiency thereby producing complex subscription summaries.

- **Subscription updates.** Several emerging CPS applications involve updates to their subscriptions at a high frequency [4], [5]. Examples include algorithmic trading and location-based services; examples for corresponding updates are changes to threshold values on prices for purchasing stock and updates to absolute geographical ranges of interest based on the change of vantage point induced by mobility respectively. In existing CPSNs, the typical way to update a subscription is through a re-subscription, which involves issuing a new subscription and canceling the old (stale) subscription. Re-subscriptions, however, typically involve $O(N)$ operations at least at one broker and potentially at many more brokers depending on the effectiveness of subscription summarization algorithm.

---

*K.R. Jayaram is with IBM T.J. Watson Research Center, New York, NY.*

*W. Wang and P. Eugster are with the Department of Computer Science, Purdue University, West Lafayette, IN.*

*E-mail: wang1315@purdue.edu, p@cs.purdue.edu.*

*Manuscript received 16 July 2013; revised 10 June 2014; accepted 18 June 2014. Date of publication 7 Sept. 2014; date of current version 7 Oct. 2015. Recommended for acceptance by D. A. Bader.*

*For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.*

*Digital Object Identifier no. 10.1109/TPDS.2014.2355823*
In this paper, we present Beretta, a novel CPSN which tackles the above-mentioned three problems through the following technical contributions:

1) **Subscription normalization** based on strong event typing without loss of expressivity of subscriptions. The normalized form is a combination of value intervals and set inclusions. Our use of event types is in lieu of a structural approach allowing for aggressive performance optimizations without hampering interoperability or the addition of new event types.

2) An efficient algorithm for matching events to subscriptions, which reduces the number of elementary predicates evaluated for matching an event from \(O(KN)\) to \(O(K \log N)\), with \(N\) being the number of subscriptions for the event type and \(K\) the number of its attributes. Our algorithm relies on a divide and conquer strategy which we call “split and subsume”—event types and normalization are exploited to split subscriptions twice, first into predicates on individual event types and further based on attributes; and to efficiently regroup these in augmented segment trees and hash maps.

3) **Inherent parameterization** of normalized subscriptions to efficiently implement parameter-based [4], [5] as well as structural subscription updates.

4) **Subscription summarization approximation** to efficiently handle subscription updates as well as joining/leaving of subscribers in decentralized CPSNs.

While normalization has long been used in the management of data “at rest”, we believe our work is the first to apply such concepts to queries on live data, and to leverage them for query updates. This paper extends a previous publication at ICDCS 2011 [6], yet includes a novel means of intersecting partial match sets efficiently based on augmented segment trees, and evaluates our approach through a new real-life benchmark.

The rest of this paper is organized as follows. Section 2 presents background information. Section 3 presents the state-of-the-art and its limitations. Section 4 introduces our subscription model. Section 5 describes our algorithm, and Section 6 presents an overview of our empirical evaluation. Section 7 draws conclusions. Due to space constraints, this paper also includes supplementary material consisting of three appendices. Evaluation results are presented in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPDS.2014.2355823. Appendix B, available in the online supplemental material, analyzes the complexity of our algorithm. A detailed analysis of related work is available in Appendix C, available in the online supplemental material.

2 BACKGROUND, MODEL, AND DEFINITIONS

This section presents notation employed, assumptions made, and definitions of concepts used in this paper.

2.1 System Model

We consider CPSNs using decentralized, dedicated, interconnected, broker processes \(b_i\) to convey events between client nodes \(c_i\), i.e., publishers and subscribers. Brokers which serve client processes are called edge brokers. We use the term *client* to refer to either publishers or subscribers. For presentation simplicity we assume CPSNs forming directed graphs.

We focus on CPS systems whose routing algorithms follow the principles of (a) **downstream replication** where an event is routed in a single copy as far as possible from the publisher and only cloned downstream as close as possible to the subscribers interested in receiving it, (b) **upstream evaluation** where unwanted events are filtered away as close as possible to the publisher to avoid wasting bandwidth, and (c) **subscription-based reverse-path forwarding**, where subscriptions are routed from the subscriber to the edge broker to which publishers are connected, thereby forming a spanning tree rooted at each subscriber, and events follow the reverse path along this tree to the subscriber. In this context, the CPSN can either (c.1) perform subscription flooding, i.e., route every subscription to every publisher in the CPSN, or (c.2) use advertisements. An advertisement defines the types of events produced by a publisher, and a subscription on an event type \(T\) has to be routed *only* to the edge broker with at least one publisher which has advertised \(T\).

2.2 Events and Subscriptions

An event \(e\) is a set of attribute/value pairs \(\{a_1 : v_1, \ldots, a_k : v_k\}\) which are typically of primitive types \(\tau_i\) including scalar types \(\text{int}, \text{float}, \text{etc.}\) and character strings (cf. [1]). A subscription is usually represented as a predicate \(\Phi\) based on a grammar like the following:

\[
\text{Predicate } \Phi ::= \Phi \wedge p | p \\
\text{Condition } p ::= a \circ \text{op} v \\
\text{Operator } \text{op} ::= \leq | < | = | > | \geq.
\]

We focus on equality in the context of strings. Evaluation of a subscription \(\Phi\) on event \(e\), written \(\Phi(e)\) involves substituting the value \(v_i\) for \(a_i\), for all \(a_i \in e\). Obviously, satisfying a subscription \(\Phi = p_1 \wedge \cdots \wedge p_n\) requires satisfying each of its conditions \(\Phi(e) = \bigwedge_{i=1}^{n} p_i(e)\). Without loss of validity but for simplifying presentation and comparison with predating work, our subscriptions do not support disjunctions. Subscriptions are usually viewed to be in disjunctive normal form (DNF), where conjunctions are handled individually as subscriptions. Disjunctions can then be dealt with by matching events sequentially and memorizing the last event sent to a given process, avoiding re-sends in case multiple conjunctions in a DNF subscription match a same event.

2.3 Content-Based Matching

The content-based matching problem (cf. [7]) is formally defined as: given an event \(e\) and a set of subscriptions \(\Theta\), compute \(\theta = \{\Phi | \Phi \in \Theta \land \Phi(e)\}\), where \(\Phi(e)\) is short for \(\Phi = \text{true}\). Clearly \(\theta \subseteq \Theta\). A naive algorithm for content-based matching works as follows: given an event \(e\), for each \(\Phi \in \Theta\), if \(\Phi(e)\), add \(\Phi\) to \(\theta\) (which is initially set to \(\emptyset\)). The matching complexity, i.e., the number of evaluated constraints, of this algorithm is \(O(KN)\), if \(|\Theta| = N\) and there are \(K\) attributes in the system. The time complexity of this algorithm is also \(O(KN)\). Note that a broker in a CPSN, which matches events to subscriptions has to perform \(O(|\theta|)\) to multicast an event to the subscribers/downstream brokers.
2.4 Subscription Summarization and Routing

To avoid storing and evaluating all subscriptions at each broker, CPSNs perform subscription summarization based on covering relationships (subsumptions) among subscriptions. When a broker receives a subscription \( \Phi \) from a subscriber or a downstream router, it forwards the subscription to an upstream router unless \( \Phi \) is covered by a previous subscription \( \Phi' \), that is, unless the set of events \( \mathcal{E} \) satisfying \( \Phi \) (\( \mathcal{E} = \{ \mathcal{E} | \Phi(\mathcal{E}) \} \)) is a subset of the set of events \( \mathcal{E}' \) satisfying \( \Phi' \) (\( \mathcal{E}' = \{ \mathcal{E}' | \Phi'(\mathcal{E}') \} \)). The condition is expressed simply as \( \mathcal{E} \subseteq \mathcal{E}' \), or put differently, \( \forall \mathcal{E} \in \Phi(e) \Rightarrow \Phi'(e) \). We say that predicate \( \Phi' \) covers (subsumes) \( \Phi \), denoted by \( \Phi \preceq \Phi' \), iff \( \forall e, \Phi(e) \Rightarrow \Phi'(e) \). To facilitate the checking of subsumption relationships between predicates, these are typically stored in a partially ordered set (poset), ordered by \( \preceq \).

Fig. 1 shows an example of a CPSN with six clients—four subscribers \( (c_1, c_2, c_3, c_4) \), two publishers \( (c_5, c_6) \) and five brokers \( (b_1, b_2, b_3, b_4, b_5) \). We focus on a single event type \( \text{StockQuote} \) with two attributes \( a_1 = \text{firm} \) and \( a_2 = \text{price} \) of the \( \text{string} \) and \( \text{float} \) types respectively. We assume that all the clients subscribe to \( \text{StockQuotes} \) of the same \( \text{firm} \), e.g., \( \text{firm} = \text{"IBM"} \). \( c_1 \) subscribes to \( \text{StockQuote} \) with \( \Phi_1 = (\text{price} < 10) \). \( b_1 \) gets the subscription, stores it and propagates it to \( b_3 \). Then \( c_2 \) subscribes with \( \Phi_2 = (\text{price} < 5) \). \( b_1 \) gets this subscription, but does not forward it to \( b_3 \), as \( \text{price} < 10 \) covers \( \text{price} < 5 \), i.e., \( \Phi_2 \preceq \Phi_1 \). Fig. 2 illustrates subscription summarization throughout the overlay. Brokers \( b_1 \) and \( b_2 \) summarize subscriptions from \( \{c_1, c_2\} \) and \( \{c_3, c_4\} \) respectively, and \( b_3 \) further summarizes the summaries from \( b_1 \) and \( b_2 \).

3 Problem Description and State-of-the-Art

Beretta tackles several inefficiencies in content-based publish/subscribe systems. In this section, we describe each of them and why they arise in detail, while also placing them in their historical context, where relevant. Related work is further elaborated on in Appendix C, available in the online supplemental material.

3.1 Inefficient Event Matching

Early systems advocated mostly for structural conformance between events and subscriptions [8]. This leads to collapsing all subscriptions for all types of events into one single data-structure (e.g., poset) thus exacerbating bottlenecks. Some early CPS systems [1] reused the poset for matching events to subscriptions to avoid separate data-structures for subscription storage/summation and event matching. This places a high load on brokers as all subscriptions end up being stored in the same data-structure with a worst-case depth equalling the number of nodes \( N \) in the poset. In turn, \( N \) is in the worst case equal to the total number of subscribers in the system though in practice it may be better. Assuming that there are \( K \) attributes in an event, and therefore up to \( O(K) \) constraints in a subscription, the complexity of event matching is typically \( O(KN) \), leading to low throughput and high end-to-end latency.

Most systems, even such described without event types, use types in practice, but only much later have the benefits of typing of events started to be exploited [9], [10]. Similarly, only little work exists on taking splitting a step further by systematically handling subscriptions attribute-wise (e.g., [11], [12]). Two prominent categories of smarter matching algorithms are those based on binary decision diagrams (BDDs—e.g. [13], [14]), and on Bloom filters (e.g. [15]). But, both these solutions have still \( O(KN) \) matching complexity, unless assumptions are made, e.g., on the amount of common constraints in subscriptions. Bloom filters also require \( O(2^B) \) space [15], where \( B \) is the number of bits used to store a subscription. Ineffective algorithms explain why, in spite of its more generic nature than the predating less dynamic topic-based multicast model, CPSNs
like Siena [11, 16], HERMES [17], REBECa [18], Gryphon [7], PADRES [14, 19] and JEDI [20] have not dethroned topic-based multicast systems like ActiveMQ [21], ZeroMQ [22] and FioranoMQ [23]. Topic-based routing is easy to implement—routing a message only involves hashing a string (its topic) to forward the message to interested downstream nodes.

### 3.2 Inefficient Subscription Updates

Several emerging CPS applications such as high frequency trading (HFT) or mobile location-based services need to update their subscriptions at a high frequency [4, 5]. HFT hinges on rapid subscription updates, using several mathematical techniques to determine and change price thresholds during the “trading day”. In location-aware applications (location-based advertising and social networks like loopt [24], etc.), a subscription is a function of the subscriber location such as a perimeter surrounding the location. Whenever the device moves, the subscription needs to change.

Re-subscriptions—the typical solution for subscription changes, where a new subscription is issued and the superseded one is canceled—have several limitations:

- **High computational cost**, due to operations at each affected broker on the poset storing subscriptions. Deletion and insertion of new subscriptions into the poset has a time complexity of $O(KN)$. Thus two $O(KN)$ operations have to be performed at a broker. As shown empirically [5], under high rates of subscription updates, the bulk of the computational resources of event brokers in a CPSN is spent on processing re-subscriptions rather than filtering and forwarding events. This leads to drastic drops in throughput and increased latency overall.

- **Cascading re-subscriptions** at upstream brokers: when the poset’s least upper bound (LUB) changes as a result of un- or re-subscription, then the posets at upstream brokers also have to be updated. Hence, re-subscriptions can lead to two update messages being propagated per update the path from a subscriber to a publisher, as illustrated by Fig. 2.

- In the absence of synchronization of **unsubscribe** and **subscribe** messages during re-subscriptions, and of guarantees by the CPSN on the time taken for re-subscriptions to reach relevant brokers, the application must cater for duplicates if the old and new subscriptions overlap—the common case.

*Parametric subscriptions* [4, 5] are subscriptions referring to variables of subscribers. A typical subscription to IBM stock quotes with values below a specific threshold expressed through a CPS API as `CPS.subscribe("StockQuote", "firm == ‘IBM’ and price < 100.0")`. Updating the price threshold to $150 would typically require a re-subscription. A correspond parametric subscriptions can be expressed as `CPS.subscribe("StockQuote", "firm == ‘IBM’ and price < ” + ref threshold)`, where threshold is an internal subscriber variable. Instead of re-subscribing with the new price of $150, the subscriber just sends the new value of `threshold`. Note that solutions for specific applications, such as context-aware publish/subscribe [25], [26] do not apply to HFT.

The decentralized implementation of parametric subscriptions described in [5], [27] still exhibits considerable limitations in application scenarios requiring fast, frequent subscription updates. First off, being “tacked on to” existing (poset-based) matching algorithms, it still incurs a high reaction time to updates. In addition parametric subscriptions do not handle structural updates, e.g., they can not handle updates from “price < 100” to “price > 100 & price < 150”. For scenarios like HFT which are governed by non-trivial algorithms additional flexibility is needed for structural subscription changes.

### 3.3 Large Subscription Summaries

The nodes of a poset can get arbitrarily complex if subscriptions do not subsume each other. Consider the scenario in Fig. 3. None of the subscriptions of clients $c_1$, $c_2$, or $c_3$ subsumes any others. Hence, the LUB is the disjunction of $\Phi_1$, $\Phi_2$ and $\Phi_3$. In the worst case, if $N$ clients connect to a broker, the LUB is the disjunction of $N$ subscriptions, which propagate through the network towards publishers (e.g., $b_1$ forwards its LUB to $b_2$).

A beneficial “side-effect” observed with parametric subscriptions [5] is that many times (un-)subscriptions translate transitively in a CPSN to parameter updates. Together with the above-mentioned observations, this hints to an intriguing potential for CPSN systems that are fundamentally based on dynamic subscriptions.

**Algorithm 1. Beretta client algorithm**

```
1: init
2: $b_j$
3: to PUBLISH(e) of type $\tau$ do
4: SEND(PUB, $\tau$, e) to $b_j$
5: to SUBSCRIBE($\Phi$) of type $\tau$ do
6: $\chi \leftarrow$ NORMALIZE - TO - WELL - FORMED($\Phi$)
7: $\chi$ to $b_j$
8: to UNSUBSCRIBE from type $\tau$ do
9: SEND(UNSUB, $\tau$) to $b_j$
10: upon RECEIVE(PUB, $\tau$, e) do
11: if $\Phi(e)$ | $\Phi$ is subscription to $\tau$ then
12: DELIVER(e)
13: upon change of value in $\delta$, for subscription to type $\tau$ do
14: $\tau_r \leftarrow$ new values for $a_r$ in $\delta_r$ \{\{a_r \in [v_1, v_2] or \delta_r = (a_r \in [v_1, \ldots, v_n])\}
15: SEND(UDP, $\tau_r$, $\delta_r$) to $b_j$
```
4 Subscription Normalization

Beretta builds on interval subscriptions—subscriptions where the only boolean operator is \( \epsilon \). To avoid ambiguities, we refer to subscriptions \( \Phi \) following the grammar introduced in Section 2.2 as regular subscriptions. Beretta normalizes regular subscriptions into well-formed interval subscriptions. An interval subscription is represented as an interval predicate \( \chi \) with the following grammar:

\[
\begin{align*}
\text{Predicate } & \quad \chi ::= \chi \land \delta | \delta \\
\text{Condition } & \quad \delta ::= a \in [v, v] | a \in \{v, \ldots, v\}.
\end{align*}
\]

A constraint \([v, v']\) represents an interval of permissible values for an attribute of an ordered type such as integers. Floating point values also fit this model. A constraint \([v, \ldots, v']\) represents a set of permissible values for an enumerated (discrete) type such as strings. A well-formed interval subscription is straightforwardly one which, for any given attribute, has exactly one condition on that attribute. While interval subscriptions improve the efficiency of content-based matching as we will show, they are as expressive as the subscriptions presented in Section 2.2 (a formal account of the expressivity of interval subscriptions is given in Appendix B, available in the online supplemental material). For example, a predicate \( x > 1000 \) where \( x \) is an integer attribute, can be expressed as \( x \in [1001, \max\_\text{int}] \). An equality (e.g., \( x = 1000 \)) can be modeled as a inclusion in a set with a single element. So the first step in subscription normalization is to convert constraints involving relational operators \(<, \leq, =, \geq, \geq \) into constraints involving the inclusion operator \( \epsilon \) as we described informally above. This step yields an interval subscription which has at most one constraint on each attribute of an event type. It is important to note that a generic set inclusion with \( n \) elements corresponds to \( n \) disjoined conjunctions (i.e., subscriptions) in regular syntax.

To normalize the interval subscription further and make it well-formed, i.e., to ensure that it contains exactly one constraint for each attribute, we add wildcard constraints to the subscription depending on the types of constraint-less attributes. An (implicit) wildcard constraint for an ordered type \( \tau \) is \([\min(\tau), \max(\tau)]\), and * for an enumerated type.

5 Beretta Algorithms

This section presents a novel divide and conquer matching algorithm for CPSNs called FASTINT, which is a key component of Beretta. FASTINT consists in a client-side component and a broker component which deals with event matching and routing.

5.1 Overview

Assuming events have up to \( K \) attributes and that the total number of subscriptions is \( N \), many existing summarization and matching algorithms have \( O(KN) \) matching complexity because they store “entire” subscriptions in a single data-structure. Then, since, subscriptions contain multiple constraints on different attributes, the chances of a subscription covering entirely another can become small, especially as the number of attributes increases. The key strategy adopted by FASTINT is therefore to split well-formed interval subscriptions by event type \( \tau \), and then by attribute \( a \), thereby using one data-structure for each attribute of every event type (see Fig. 4).

Algorithm 2. FASTINT broker algorithm as executed by \( b_i \). Common processing of data-structure modifications (new subscriptions, unsubscriptions, updates) are regrouped in PROPAGATE. \( \oplus \) represents concatenation

1: init
2: subs[] \( \leftarrow \) {subscribed clients/brokers by event types \( \tau \)}
3: pubs[] \( \leftarrow \) {publishing clients/brokers by event types \( \tau \)}
4: vals[] \( \leftarrow \) {current constraint values by event types \( \tau \), attribute, node id}
5: \( S[] \) \( \leftarrow \) {per-attribute data-structures by event types \( \tau \) and attribute}
6: upon receive(sub, \( \tau, \{n_1, n_2\} \) from \( n_j \) do
7: \( \text{updates} \leftarrow \emptyset \)
8: subs[\( \tau \) \( \leftarrow \) subs[\( \tau \)] \( \cup \) \( \{n_j\} \) \( \{ \text{add node to subscribers of } \tau \} \)
9: for all \( a \in \tau \) do
10: \( \text{vals}[\tau][r][n_j] \leftarrow \text{vals}[\tau][r][n_j] \cup \text{old LUB} \)
11: \( \text{vals}[\tau][r][n_j] \leftarrow \text{vals}[\tau][r][n_j] \cup \text{old LUB} \)
12: \( \text{insert}(\text{vals}[\tau][r][n_j], \text{vals}[\tau][r][n_j]) \)
13: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
14: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
15: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
16: upon receive(unsub, \( \tau \) from \( n_j \) do
17: \( \text{updates} \leftarrow \emptyset \)
18: subs[\( \tau \) \( \leftarrow \) subs[\( \tau \)] \( \setminus \) \( \{n_j\} \)
19: for all \( a \in \tau \) do
20: \( \text{vals}[\tau][r][n_j] \leftarrow \text{vals}[\tau][r][n_j] \cup \text{old LUB} \)
21: \( \text{vals}[\tau][r][n_j] \leftarrow \text{vals}[\tau][r][n_j] \cup \text{old LUB} \)
22: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
23: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
24: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
25: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
26: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
27: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
28: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
29: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
30: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
31: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
32: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
33: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
34: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
35: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
36: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
37: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
38: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
39: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
40: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
41: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
42: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
43: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
44: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
45: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
46: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
47: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
48: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
49: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
50: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)
51: \( \text{PROPAGATE}(\tau, \text{updates}) \text{ end upon} \)

For presentation simplicity, in the following, we assume that 1) attributes are named uniquely in event types and are ordered in events (e.g., based on names), and that 2)
subscription identifiers are contiguous. $\mathcal{T}$ represents a sequence $x_1, \ldots, x_n$.

A broker $b_i$ in a CPSN receives well-formed interval subscriptions from subscribers $c_j$ connected to it as well as downstream brokers. Subscriptions on an event type with $k$ attributes are of the form $\chi = \delta_1 \land \cdots \land \delta_k$. A client (see Algorithm 1) communicates with its edge broker $b_j$. When creating a normalized subscription $\chi$ (line 6), wildcard constraints are automatically added for unconstrained attributes. Upon receiving an event $e$, a client $c_i$ verifies if $\chi(e)$ holds, as updates might have occurred in the meantime. Updates to a predicate $\delta$, trigger the sending of an update message to $b_j$.

5.2 Subscription Indexing

Next we focus on the FASTINT broker algorithm. As illustrated in Fig. 4 data-structures for storing constraints are accessed by a double index, namely on event type $\tau$ and attribute. In Algorithms 2 and 3, $S$ represents this index. In addition the algorithm uses three associative maps $\text{subs}$, $\text{pubs}$ and $\text{vals}$. The former two store sets of subscribing clients or downstream brokers, and sets of publishing clients or upstream brokers, respectively, indexed by event type $\tau$. For presentation simplicity operations on these data-structures are represented as assignments $\leftarrow$ and operations on the sets that they store. To perform subscription updates efficiently, $\text{vals}$ is used. $\text{vals}$ reflects the normalized nature of subscriptions, and their inherent parameterization: $\text{vals}[\tau][r][n]_i$ contains a client/downstream broker’s (with id $n_i$) current bounds on the interval query or set inclusion constraint on attribute $a_i$ of event type $\tau$. Each well-formed subscription $\chi$ on $\tau$ namely contains exactly one constraint on any attribute $a_i$. The constraint either consists in upper and lower bounds (two values) or in a set of permissible values depending on the type $\text{typeof}(r, \tau)$ of the attribute $a_i$ in $\tau$. Either of these can be represented by a sequence of values $\pi$. In the absence of actual constraints, $\text{MIN}(\tau)$ and $\text{MAX}(\tau)$ are the values for an attribute of ordered type $\tau$ and $*$ the single value for an attribute of string type. Note that this model also captures structural updates: for example, an update from $[\text{MIN}(\tau), v]$ to $[v', \text{MAX}(\tau)]$ for an attribute $a$ reflects the change of a constraint $a \leq v$ to a constraint $a \geq v'$.

5.3 SV Tree and Hash Maps

While $\text{vals}$ serves for fast updates, FASTINT stores constraints for actual matching in a $\text{SV}$ tree for an $a$ of an ordered type (see $x$ in Fig. 5), and in a $\text{hash map}$ for an attribute $a$ of an enumerated type (see $z$ in Fig. 5). The potential drawback of a matching algorithm based on the divide and conquer strategy is that when an event with values for $k$ attributes $e = \{a_1 : v_1, \ldots, a_k : v_k\}$ is received, each of the corresponding $k$ data-structures has to be queried and the results combined. If querying data-structure $i$ produces $\text{result}_i$, the set of subscriptions $\text{result}_i$ matching $e$ is the intersection $\bigcap_{i=1}^{k} \text{result}_i$ whose naive computation takes $O(\max_i(\text{result}_i))$ time. This is undesirable because $\text{result}_i$ in many cases will be smaller than any of the $\text{result}_i$.

To avoid this, we precompute three pieces of information for the relevant nodes in the attribute data-structures—$S$, $C$ and $V$. $S$ is the set of identifiers of subscriptions whose respective constraints on attribute $a$ satisfy (“include” in the case of interval subscriptions) the value(s) represented by that specific node. $C$ is simply $|S|$, and $V$ is a bit vector of size $N$, indexed by the identifiers of subscriptions. Bit $V[i]$ is set to $1$ iff $i \in S$. We use $S$, $C$ and $V$ for computing intersections efficiently.

When $a$ is of the string type, an entry in the hash map for attribute constraint $a \in [v]$ contains a tuple $(S, C, V)$ as outlined above with $S$ being the set of identifiers of subscriptions with constraints on $a$ including $v$.

A segment tree [28] is a data-structure for storing line segments or intervals (represented as axis-parallel line segments). In this paper, we introduce an SV tree, which is a segment tree optimized for queries on points (rather than intervals) and augmented with information such as $\langle S, C, V \rangle$ mentioned above for efficient intersection. Our SV tree is implemented as a B-tree with a fanout of $f$, i.e., each node in the B-tree has at least $f$ and up to $2f - 1$ children. Assuming that there are $n$ subscriptions to be processed, let the canonical intervals (i.e., intervals specified in the subscriptions) corresponding to an ordered attribute $a$ be $[v_1, v_2], \ldots, [v_{2n-1}, v_{2n}]$. Like traditional segment tree
construction algorithms, FASTINT partitions the real line induced by the values \( v_i \), i.e., FASTINT sorts the endpoints of the canonical intervals, obtaining \( v'_1, \ldots, v'_{2n} \), and stores the following disjoint elementary intervals \([\text{MIN}(\tau), v'_1], [v'_1, v'_2], [v'_2, v'_3], \ldots, [v'_{2n-1}, v'_{2n}], [v'_{2n}, \text{MAX}(\tau)]\) in the leaves of the segment tree. Each leaf node is also augmented with \(<s, c, v>\) just like hash map entries. An intermediate node contains simply the minimal interval covering all segments of its sub-nodes (either intermediate or leaf nodes). While matching an attribute of an event against the corresponding SV tree, matching occurs recursively one level at a time. At most one sub-node’s interval matching at any level. After reaching a leaf node, \(<s, c, v>\) is returned. \(<s, c, v>\) is similarly stored for every key in a hash map and returned upon a corresponding query.

### 5.4 Basic Operations

Assuming that the hash map implementation uses a good hashing function, load factors less than 0.5, and chaining and rehashing, all elementary hash map operations have a time complexity of \( O(1) \) [29]. In the case of a set of values \( v = v_1 \ldots v_l \) this time is multiplied by \( l \). Note that for complexity comparisons we can consider \( l = 1 \), because \( l > 1 \) would require a disjunction of \( l \) conjunctions in a regular DNF subexpression \( \bigvee_{i=1}^{l} a = v_i \). In the formal description of FASTINT in Algorithms 2 and 3, the standard operations on hash maps and SV tree s are represented in a uniform manner:

\( \text{insert}(D, \pi, id) \) creates an association between \( \pi \) and \( id \) in data-structure \( D \). In the case of a hash map, \( \pi \) represents a set of keys (values from the perspective of matching) for which \( id \) (a process identifier \( n_i \)) is added to the set of corresponding values. For an SV tree, \( \pi = v_1, v_2 \) represents the upper and lower bounds of an interval which is of interest to \( id \). In the case of a hash map, insertion of each key of \( \pi \) has a time complexity of \( O(1) \) due to hashing. But, assuming that the hashmap contains \( n \) subscriptions, i.e. \( n \) keys, the vector \( V \) associated with each key is of size \( n \). Hence updating \( n \) vectors has a time complexity of \( O(n) \), which becomes the time complexity of \( \text{insert}(D, \pi, id) \). The same is true in the case of an SV tree. Inserting an interval \([v_1, v_2]\) into the SV tree only involves traversing \( \log n \) nodes of the tree to find the spots to insert \( v_1 \) and \( v_2 \), up to \( n \) vectors have to be updated, with a time complexity of \( O(n) \).

\( \text{delete}(D, \pi, id) \) removes the association between \( \pi \) and \( id \) from data-structure \( D \). In the case of a hash map, \( \pi \) is removed for all keys \( \pi \). For an SV tree, the interval denoted by \( \pi = v_1, v_2 \) associated with \( id \) is removed. In the case of a hash map, deletion of each key of \( \pi \) has a time complexity of \( O(1) \) due to hashing. But, assuming that the hashmap contains \( n \) subscriptions, i.e. \( n \) keys, the vector \( V \) associated with each key is of size \( n \). Hence updating \( n \) vectors has a time complexity of \( O(n) \), which becomes the time complexity of \( \text{delete}(D, \pi, id) \). In the case of an SV tree, deleting an interval \([v_1, v_2]\) from the SV tree only involves traversing \( \log n \) nodes of the tree to find the tree nodes containing \( v_1 \) and \( v_2 \), up to \( n \) vectors have to be updated, with a time complexity of \( O(n) \).

\( \text{retrieve}(D, v) \) queries data-structure \( D \) for all \( id \)s associated with value \( v \). For a hash map, the return set includes all values stored for key \( v \); for an SV tree the query returns the \( id \)s of all intervals which contain \( v \). \( \text{lub}(D) \) returns the “covering constraint” for data-structure \( D \). For a hash map \( H \) this includes all keys, unless the key set includes the wildcard \( * \), in which case only that value is needed. For an SV tree \( T \), the covering interval stored in the root node is returned. 

\( \text{update}(D, \pi, id) \) changes the association between \( \pi \) and \( id \) in data-structure \( D \) to \( \pi' \) for \( id \). This can be naïvely implemented by \( \text{delete} \) and \( \text{insert} \). This operation is only used for \( S \), where both bounds for a constraint are replaced atomically for simplicity even if only one has changed; in the case of hash maps used for string attributes, updates focus on only adding new elements and removing obsolete ones. Handling of new subscriptions and unsubscriptions follow similar structures. Changes are made to all relevant data-structures. For any such data-structure \( D \), the LUB before update is stored. After all updates have been performed, the set of previous LUBs and respective new LUBs are passed to \( \text{propagate} \) which determines all actual changes. In the case of interval queries, it suffices that one of the bounds has changed. In the case of set inclusions the obsolete elements \( \pi' \) and new elements \( \pi'' \) are determined; the two sets are concatenated by inserting a special token value \( \bullet \); \( \pi'' \bullet \pi'' \). When handling a corresponding update message \( \text{upd} \), the two sets are removed from and added to the appropriate hash map respectively. After handling such an update, any transitive updates are similarly handled via \( \text{propagate} \). This illustrates the fundamental nature of updates in Beretta: except when a broker \( b_1 \) connects to another broker \( b_2 \) for the first time or disconnects from \( b_2 \), all changes in broker connections take place via update messages.

### 5.5 Matching

When an event \( e \) with \( k \) attributes is received (Lines 52-65 in Algorithm 3), at most \( k \) corresponding data-structures are “evaluated” to determine subscriptions that match \( e \). If an attribute \( a_i \) is of an ordered type, then \( S \) is queried in \( O(1) \) to retrieve the corresponding SV tree \( T \) (Line 59). If \( e \)’s \( a_i \) value is \( v_i \), then \( T \) is searched with \( v_i \) to determine the set of intervals (and thereby the subscriptions \( S[i] \)) that match \( v_i \). If \( a_i \) is of \text{string} type, \( S \) is queried in \( O(1) \) to retrieve the corresponding hash map \( H \). The subscriptions \( S[i] \) that match \( v_i \) are obtained by querying \( H \) for \( v_i \). 

\( \text{get} \) on hash maps and SV tree s returns a tuple with three elements, \( <s, c, v> \).

In Algorithm 3, we use a variable \( \text{Vec} \) to refer to a precomputed bit matrix whose dimensions are \( k \times N \). When an event \( e = \{a_1 : v_1, \ldots, a_k : v_k\} \) is received, \( \text{Vec}[i] \) points to the vector encoding subscriptions matching \( v_i \). Since every subscription matches an unconstrained attribute, \( \text{Vec}[i] \) points to a bit vector whose bits are 1 if \( a_i \) is unconstrained. Since \( \text{result} \subseteq \min_{i=1}^{n}(\text{result}_i) \), Algorithm 3 keeps track of the index \( \text{min}(\text{result}_i) \) of the smallest \( \text{result}_i \) and the set of subscriptions contained in it. To decide whether \( e \) matches any of the subscriptions in \( \text{result}_i \), a bitwise AND operation is performed on the \( k \) bits of \( \text{Vec}[j] \) for all \( j \in \text{Vec}[\text{min}(\text{result}_i)] \). This yields a complexity of \( O(K \log N + K \min_{i=1}^{n}(|\text{result}_i|)) \), because \( k \leq K \).
5.6 Summarization Approximation

When sending a summary to a parent, a broker approximates the summarization. Remember that a broker manages a separate data-structure $S[|\tau|]$ for each attribute $a_i$ of a given event type $T$. The LUB of such a data-structure covers all conditions on the respective attribute for all subscriptions known to the broker. A broker summary for a given event type simply consists in the conjunction of these LUBs. In Algorithm 2 this conjunction arises implicitly, as update messages (line 51) contain the LUBs for the relevant types and attributes.

Algorithm 3. FASTINT-M – Matching

52: upon receive($pub, \tau, e = \{a_1 : v_1, \ldots, a_k : v_k\}$) from $n_j$
53: $Subs \leftarrow$ new 2-D array of size $k \times |\text{subs}[\tau]|$
54: $Count \leftarrow$ new 1-D array of size $k$
55: $Vec \leftarrow$ new 2-D array of size $k \times |\text{subs}[\tau]|$
56: $min\_result\_count \leftarrow |\text{subs}[\tau]|$
57: $min\_result \leftarrow \text{subs}[\tau]$
58: for $i = 1..k$ do
59: \hspace{1em} $\langle Subs[i], Count[i], Vec[i] \rangle \leftarrow \text{RETRIEVE}(S[|\tau|], a_i)$
60: if $Count[i] \leq \text{min\_result\_count}$ then
61: $\text{min\_result\_count} \leftarrow Count[i]$
62: $\text{min\_result\_index} \leftarrow i$
63: for all $i$ do $\text{index} \in \text{Subs[m\_result\_index]}$ do
64: \hspace{2em} if $\text{BITWISE-AND}(Vec[\text{index}][1..k]) = 1$ then
65: $\text{result} \leftarrow \text{result} \cup \{\text{index}\}$
66: for all $n_j \in \text{result}\{-n_j\}$ do
67: $\text{SEND}(pub, \tau, e)$ to $n_k$

This approximation can be disabled in Beretta, e.g., for sets of summarized subscriptions below a threshold size, and a disjunction (created like in traditional approaches) sent to parent brokers instead. In this scenario, a subscription sent by a broker $b_j$ to its parent broker $b_i$ consists in fact in a set of (logically disjoined) normalized subscriptions, say of size $u$, necessary to cover all subscriptions of $b_j$. Upon a change at $b_i$, to its data-structures, $e.g.$, induced by the addition, removal, or update of a subscription, there are two scenarios: (a) the set of disjoined subscriptions in this root subscription remains the same, or (b) it changes. In the former case, if the values of certain variables have changed only the updates need to be transmitted to $b_j$. In the latter case we can further distinguish three cases, based on whether the number of disjoined subscriptions in the root (b.1) grows to $v$, (b.2) remains $u$ (some subscriptions in the disjunction may have changed though), or (b.3) shrinks to $v$. In all cases, we identify $\text{min}(u, v)$ subscriptions—preferably such that were in the previous set already—with the previous ones, and send corresponding variable updates where necessary. If we have more subscriptions now ($v > u$) or fewer ones ($v < u$) then $b_i$ additionally informs $b_j$ of the new ones to be added or of those to be removed. The latter actions are handled differently from regular new subscriptions on unsubscriptions, as brokers take note of multiple subscriptions that are logically linked to a same peer or client to avoid multiple transmissions of the same event.

5.7 Initialization

Initialization on a given node with a set of subscribers (downstream nodes) is simple and thus only outlined informally here. After identifying the set of attribute/type pairs across all subscriptions an empty SV tree is created for every attribute of ordered type encountered in some subscription, and an empty hash map is likewise created for any attribute of enumerated type. These are indexed in $S$ by their respective attribute/type pairs.

Then the data-structures are populated from subscriptions, one-by-one. In contrast to the handling of a new subscription in Algorithm 2, (any) new LUBs are only determined once all subscriptions have been added.

6 SUMMARY OF EVALUATION

In this section we present a summary of our empirical evaluation. Please refer to Appendix A, available in the online supplemental material, for a detailed description. We use two benchmarks for empirical evaluation, based on: (1) Marketcetera algorithmic trading, and (2) Highway Traffic Management (HTM). We compare Beretta against (1) Beretta-NoApprox, which disables summarization approximation in Beretta, (2) EV, a CPSN which uses the Rete algorithm [30] for event matching, posets for subscription summarization and supports parametric subscriptions [5], (3) EV-Resub, which is the same as EV, except for the use of re-subscriptions for subscription updates, (4) the seminal Siena [1] CPSN, and (5) Apache ActiveMQ [21]. We use four main metrics to evaluate the performance—throughput, event propagation latency (EPL), delay in propagating updates from subscribers to CPSNs, and the percentage of stale events that are delivered to a subscriber during subscription updates. In addition we measure the main memory (RAM) usage of the various CPSNs as well as bandwidth usage.

Tables 1 and 2 present an overview of our results. Details of the topology used, benchmarks, systems being evaluated, configurations and a detailed analysis of the results, along with graphical illustrations of the same are presented in

<table>
<thead>
<tr>
<th></th>
<th>Increase in Throughput</th>
<th>Decrease in EPL</th>
<th>Decrease in Delay</th>
<th>Decrease in Spurious events</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beretta vs.</td>
<td>1.14 – 1.28 ×</td>
<td>1.12 – 1.19 ×</td>
<td>1.09 – 1.21 ×</td>
<td>1.38 – 1.55 ×</td>
<td># subscribers</td>
</tr>
<tr>
<td>Beretta-NoApprox</td>
<td>1.19 – 1.46 ×</td>
<td>1.05 – 1.23 ×</td>
<td>1.13 – 1.31 ×</td>
<td>1.07 – 3.17 ×</td>
<td>upd. frequency</td>
</tr>
<tr>
<td>Beretta vs.</td>
<td>1.58 – 1.85 ×</td>
<td>2.16 – 2.44 ×</td>
<td>1.07-1.29 ×</td>
<td>2.24 – 4.48 ×</td>
<td># subscribers</td>
</tr>
<tr>
<td>EV</td>
<td>1.67 – 1.97</td>
<td>2.16 – 2.43 ×</td>
<td>1.95 – 2.44 ×</td>
<td>1.61 – 3.57 ×</td>
<td>upd. frequency</td>
</tr>
<tr>
<td>Beretta vs.</td>
<td>3.08 – 5.98 ×</td>
<td>3.12 – 4.4 ×</td>
<td>2.05 – 5.7 ×</td>
<td>3.38 – 7.43 ×</td>
<td># subscribers</td>
</tr>
<tr>
<td>ActiveMQ</td>
<td>3.2 – 7.2 ×</td>
<td>3.14 – 4.7 ×</td>
<td>2.4 – 6.5 ×</td>
<td>2.13 – 6.21 ×</td>
<td>upd. frequency</td>
</tr>
</tbody>
</table>
Appendix A, available in the online supplemental material. In the case of the Marketcetera benchmark, we observe that the use of summarization approximation increases the throughput of Beretta by up to 1.29x with an increase in the number of subscribers and by up to 1.25x with an increase in update frequency, both of which are non-trivial. EPL, delay and the percentage of spurious events are lower in Beretta by up to 1.22x, 1.58x and 1.35x respectively. With respect to Siena, the improvements in performance are very substantial, by up to more than 100x in the case of throughput. This is because of (1) the use of posets for both summarization and matching, (2) inefficiencies due to not aggressively using types, and (3) the use of re-subscriptions for handling subscription changes. Similar trends are also seen in the case of EPL and delay (more than 5x lower) as well as the percentage of spurious events (more than 20x lower). As the differences in performance are easily observed in the case of Siena, we do not quantify them in more detail in Tables 1 and 2.

In the case of Marketcetera, when compared to EV, the performance benefits are much more substantial—up to 2.47x higher throughput, and up to 3x, 4.29x and 3.47x lower EPL, delay and spurious events respectively. Since EV already supports re-subscriptions, these benefits are purely due to the efficiency of FASTINT over the Rete matching algorithm due to type-based splitting, indexing and efficient computation of intersections. In Appendix A, available in the online supplemental material, we also observe a comparison of EV versus EV-Resub as well as Beretta versus EV-Resub—EV significantly outperforms EV-Resub which characterizes the performance benefits of parametric subscriptions. We also observe that performance benefits of Beretta are much more substantial when compared to ActiveMQ—up to 9.71x higher throughput as well as up to 4.06x, 7.94x and 15.5x lower EPL, delay and spurious events respectively. This is both because of the absence of a summarization-based overlay network, as well as inefficient matching.

As hinted to by Table 1, performance benefits similar to Marketcetera are also seen in the case of HTM. The trends are very similar, even though the two benchmarks having different numbers of publishers, subscribers and event types. The causes of these performance trends between the different systems also remain the same, as contained in the discussion of Marketcetera above.

## 7 Conclusions

This paper has presented Beretta, a new content-based publish/subscribe system. Beretta introduces types, and a simplified predicate grammar which allows it to normalize all subscriptions, leading to implicit parameterization and thus fast, localized, subscription updates.

Many algorithms for CPS matching have been presented in literature, most without analysis of their complexity. FASTINT is to the best of our knowledge the first algorithm to achieve matching of events with a number of constraint matches logarithmic in the total number of subscriptions. FASTINT also lends itself well to subscription summarization [31], by allowing for the most generic constraints on individual attributes to be extracted easily from respective data structures and conjoined to form an over-approximated summary of size O(K). For precise summarization, a poset can also be used [32].

This paper, in addition to presenting FASTINT, also presents a detailed empirical evaluation of Beretta by varying several parameters—number of event types, update rate at subscribers, number of subscribers as well as selectivity. In all cases, Beretta outperforms several other state-of-the-art publish/subscribe systems.

## Acknowledgments

Research partially funded by US National Science Foundation (NSF) (grants 0644013 and 0834529) and by DARPA (grant N11AP20014).

## References


