Delegated Distributed Mappings
draft-watson-dinrg-delmap-02

Abstract

Delegated namespaces underpin almost every Internet-scale system - domain name management, IP address allocation, Public Key Infrastructure, etc. - but are centrally managed by entities with unilateral revocation authority and no common interface. This draft specifies a generalized scheme for delegation that supports explicit time-bound guarantees and limits misuse. Mappings may be secured by any general purpose distributed consensus protocol that supports voting; clients can query the local state of any number of participants and receive the correct result, barring a compromise at the consensus layer.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on October 27, 2019.

Copyright Notice

Copyright (c) 2019 IETF Trust and the persons identified as the document authors. All rights reserved.
1. Introduction

Internet entities rely heavily on delegated namespaces to function properly. Typical web services have been delegated a domain name (after negotiation with an appropriate registrar) under which they host the entirety of their public-facing content, or obtain a public IP range from their ISP, which itself has been delegated to intermediary registries by the Internet Numbers Registry [RFC7249]. An enormous amount of economic value is therefore placed in these assignments (in this draft, _mappings_) yet they are dangerously ephemeral. Delegating authorities, either maliciously or accidentally, may unilaterally revoke or replace mappings they’ve made, compromising infrastructure security. Presented in this draft is a generalized mechanism for securely managing such mappings and their delegations. Known entities identified by public key are assigned namespaces (e.g. domain prefixes) under which they are
authorized to create mapping records, or _cells_. Cells in a namespace are grouped into logical units we term _tables_.

Alone, this structure does not ensure security, given that any hosting server could arbitrarily modify cells or serve bogus entries to unwitting clients. We maintain security and consistency by relying on a distributed consensus algorithm. While detailed descriptions of varying consensus protocols are out of scope for this draft, we provide for a general-purpose interface between the delegation structure and a consensus layer. At a minimum, the consensus layer must apply mapping updates in a consistent order, prevent equivocation, disallow unauthorized modification, grant consensus nodes the ability to enforce high-level rules associated with the tables, and perform voting among nodes to decide top-level governance. We find that federated protocols such as the Stellar Consensus Protocol [I-D.mazieres-dinrg-scp] are promising given their capability for open participation, broad diversity of interests among consensus participants, and providing accountability for malicious behavior. Clients may query any number of trusted servers to retrieve a correct result, barring widespread collusion.

The ability to impose consistency yields several useful properties. The foremost is enforcing delegation semantics: a table’s authority may choose to recursively delegate a portion of its own namespace, but must document the specific range and delegate in one of its table’s cells. Since delegation forms a new table, for which a delegate is the sole authority, assigned namespace ranges must be unique. Consensus is also used to enforce that the delegating authority not make modifications to any delegated table and thus need not be trusted by the delegatee.

In addition, we provide explicit support for commitments that enforce an explicit lower-bound on the duration of delegations. Otherwise valid changes to cells that have a valid commitment are disallowed, including revoking delegations. Upon expiration, however, the same namespace may be delegated to another party.

Finally, decentralized infrastructure is highly visible and commonly misused. As mappings are replicated among consensus nodes, of primary concern is resource exhaustion. We limit undesired abuse of the structure by embedding recursive scale restrictions inside mappings, verified and ratified at consensus time. Combined with time-bound delegations, this ensures that the system is resistant to spam in the short-term and can remove misbehaving hierarchies in the long-term.
The remainder of this draft specifies the structure for authenticated mapping management as well as its interfaces to consensus protocol implementations and users.

2. Structure

Trust within the delegation structure is based on public key signatures. Namespace authorities must sign mapping additions, modifications, delegations, and revocations to their table as proof to the consensus participants that such changes are legitimate. For the sake of completeness, the public key and signature types are detailed below. All types in this draft are described in XDR [RFC4506].

typedef publickey opaque<>; /* Typically a 256 byte RSA signature */

struct signature {
    publickey pk;
    opaque data<>;
};

2.1. Cells

Cells are the basic unit of the delegation structure. In general, they compose an authenticated record of a mapping that may be queried by clients. We describe two types of cells:

    enum celltype {
        VALUE = 0,
        DELEGATE = 1
    };

Value cells store individual mapping values. They resolve a lookup key to an arbitrary value, for example, an encryption key associated with an email address or the zone files associated with a particular domain. The public key of the cell’s owner (e.g. the email account holder, the domain owner) is also included, as well as a signature authenticating the current version of the cell. Since the cell’s contents are controlled by the owner, its "value_sig" must be made by the "owner_key". The cell owner may rotate their public key at any time by signing the update with the old key.

    struct valuecell {
        opaque value<>;
        publickey owner_key;
        /* Owner signs contents */
        signature value_sig;
    };
Delegate cells have a similar structure but different semantics. Rather than resolving to an individual mapping, they authorize the **delegee** to create arbitrary value cells within a table mapped to the assigned namespace. This namespace must be a subset of the **delegator**’s own namespace range. Like the table authority, the delegee is uniquely identified by their public key. Each delegate cell and subsequent updates to the cell are signed by the delegator - this ensures that the delegee cannot unilaterally modify its namespace, which limits the range of mappings they can create to those legitimately assigned to them.

```c
struct delegatecell {
    opaque namespace<>
    publickey delegee;
    /* Table authority controls delegations, not delegee */
    signature delegation_sig;
    unsigned int allowance;
};
```

Both cell types share a set of common data members, namely a set of UNIX timestamps recording the creation time and, if applicable, the time of last modification. An additional commitment timestamp must be present in every cell. Each commitment is an explicit guarantee on behalf of the table’s authority that the mapping will remain valid until at least the time specified. Therefore, while value cell owners may modify their cell’s contents at any time (e.g. this scheme supports key rotation), the authority cannot change or remove the cell until its commitment expires, as enforced by the consensus nodes. Similarly, delegated namespaces are guaranteed to be valid until the commitment timestamp expiration, although after expiration, they can be reassigned to other parties. Likely, most long-term delegations will be renewed (with a new commitment timestamp) before the expiration of the current period. The tradeoff between protecting delegees from arbitrary authority action and allowing quick delegation reconfiguration is customizable to the use case. Likely, widely-used services will see larger delegation periods for stability whereas small namespaces with relatively unknown delegees will experience shorter delegations.

An **allowance** must be provided, which limits the upper-bound size of a delegated table. For value cells, the allowance value is ignored and set to 0. Importantly, for delegate cells, an allowance with value 0 indicates no limit is placed on the size of the delegated table. Given that the delegee has complete control over the contents of their table, it is emphatically not recommended to grant a delegatecell an unlimited allowance, in order to limit the storage burden on consensus nodes. This limit is recursive along delegations - the total number of cells in a table plus the sum of allowances.
among its "delegatecells" must be less than or equal to the table’s allowance, if non-zero. Further, a table with a non-zero allowance may not grant a delegate an unlimited allowance. These properties must be validated during consensus before adding new cells to a table, which can be performed at every consensus node because table entry counts are visible publicly.

Finally, a valid table cell’s timestamps and allowance is signed by the table authority and placed in "authority_sig".

```c
union innercell switch (celltype type) {
    case VALUE:
        valuecell vcell;
    case DELEGATE:
        delegatecell dcell;
};

struct cell {
    /* 64-bit UNIX timestamps */
    unsigned hyper create_time;
    unsigned hyper *revision_time;
    unsigned hyper commitment_time;
    /* Ignored by value cells */
    unsigned int allowance;
    signature authority_sig;
    innercell c;
}
```

### 2.2. Tables

Every cell is stored in a table, which groups all the mappings created by a single authority public key for a specific namespace. Individual cells are referenced by an application-specific label in a lookup table. The combination of a lookup key and a referenced cell value forms a mapping.

```c
struct tableentry {
    opaque lookup_key<>
    cell c;
};
```

Delegating the whole or part of a namespace requires adding a new lookup key for the namespace and a matching delegate cell. Each delegation must be validated in the context of the other table entries and the table itself. For example, the owner of a table delegated an /8 IPv4 block must not to delegate the same /16 block to two different tables.
2.3. Prefix-based Delegation Correctness

To generalize correctness, each table must conform with a prefix-based rule: for every cell with value or delegation subset "c" in a table controlling namespace "n", "n" must (1) be a prefix of "c" and (2) there cannot exist another cell with value or delegation subset "c2" such that "c" is a prefix of "c2".

While there exist many more hierarchical naming schemes, many can be simply represented in a prefix scheme. For example, suffix-based delegations, including domain name hierarchies, can use reversed keys internally and perform a swap in the application layer before displaying any results to clients. Likewise, 'flat' delegation schemes where there is no explicit restriction can use an empty prefix.

2.4. Root Key Listing

Each linked group of delegation tables for a particular namespace is rooted by a public key stored in a flat root key listing, which is the entry point for lookup operations. Well-known application identifier strings denote the namespace they control. We describe below how lookups can be accomplished on the mappings.

```c
struct rootentry {
    publickey namespace_root_key;
    string application_identifier<>;
    signature listing_sig;
    unsigned int allowance;
}

struct rootlisting {
    rootentry roots<>;
}
```

A significant question is how to properly administer entries in this listing, since a strong authority, such as a single root key, can easily protect the listing from spam and malicious changes, but raises important concerns about misuse. Concurrent work on IP address allocation [IP-blockchain] explores using a Decentralized Autonomous Organization built on the Ethereum blockchain to manage all delegations where proper behavior is economically motivated. We identify similar challenges: controlling spam and misuse, while operating in a decentralized manner.
In this draft, however, we focus on enabling governance through consensus operations. For that reason, potential root entries are nominated with a proposed allowance, which will restrict the total number of cells currently supported by an application. For large systems such as IP delegation or well-known entities like the IETF, the limit can be disabled as discussed earlier in this draft. It is important that decisions regarding root listing membership be made by the consensus nodes themselves, since they bear the largest burden to store tables, communicate with other nodes, and service client queries. This structure further allows table authorities to focus on content-specific administration of their own namespaces, which is not provided for in the generic delmap semantics. If an application begins to run out of allowance (too many cells or large delegations), it can sign and nominate a new "rootentry" for the same application identifier with a larger value, at which point the other nodes can (given global knowledge of table sizes and growth rates, along with additional real-world information, if applicable) determine whether or not to accept the change.

Thus, this draft explicitly requires prospective consensus algorithms to provide a mechanism for inter-node voting on governance issues. This is already common in protocols that provide for periodic updates: quorums in slice infrastructures like Stellar, Bitcoin-style percentage- and time-based agreement windows, or hard forks. Finally, although the possibility of fundamental disagreement in governance between different sets of consensus nodes is real, in realistic settings, in the worst case such groups should not necessarily continue to trust each other in consensus relationships and may indicate a (valid) need to split.

3. Interacting with a Consensus Node

3.1. Storage Format

Delegation tables are stored in a Merkle hash tree, described in detail in [RFC6962]. In particular, it enables efficient lookups and logarithmic proofs of existence in the tree, and prevents equivocation between different participants. Among others, we can leverage Google’s [Trillian] Merkle tree implementation which generalizes the datastructures used in Certificate Transparency. In map mode, the tree can manage arbitrary key-value pairs at scale, but critically, this requires flattening the delegation links such that each table may be queried, while ensuring that a full lookup from the application root is made for each mapping.

Given a "rootentry", the corresponding table in the Merkle tree can be queried at the following key (where || indicates concatenation):
root_table_name = app_id || namespace_root_key

It follows that tables for delegated namespaces are found at:

table = root_table_name || delegee_key_1 || ... || delegee_key_n

And finally, individual entries are identified by the namespace lookup key:

cell = table || desired_lookup_key

Once an entry is found in the tree, a logarithmic proof can be constructed with the hashes of the siblings of each node in the tree’s path to the entry.

struct merkleproof {
    opaque sibling_hashes[32]<>
    cell entry_cell;
    signature tree_sig;
}

The entry is hashed together with each "sibling_hash" - if the total matches the known tree root hash, then the entry must have been in the tree.

### 3.2. Client Interface

The presence of a natural mapping structure motivates an external client interface similar to a key-value store.

struct MerkleRootOperation { }

struct MerkleRootReturn {
    opaque root_hash[32];
    signature tree_sig;
}

It is important to note that the client should not rely on a root hash that has been provided by a single server to verify a "merkleproof", instead querying multiple consensus nodes using this interface. Upon discovering that different servers are advertising non-matching hashes, the signed proof should be used to prove to other clients/nodes that one or more malicious trees are equivocating.
enum ReturnCode {
    CELL = 0,
    TABLE = 1,
    ERROR = 2
}

struct GetOperation {
    string application_identifier;
    opaque full_lookup_key<>;
}

union GetReturn switch (ReturnCode ret) {
    case CELL:
        cell value;
        merkleproof p;
    case TABLE:
        table t;
        merkleproof p;
    case ERROR:
        string reason;
}

Given an application identifier and the fully-qualified lookup key, the map described in the previous section can be searched recursively. At each table, we find the cell whose name matches a prefix of the desired lookup key. If the cell contains a "valuecell", it is returned if the cell’s key matches the lookup key exactly, else an "ERROR" is returned. If the cell contains a "delegatecell", it must contain the key for the next table, on which the process is repeated. If no cell is found by prefix-matching, the node should return "ERROR" if the key has not been fully found, else the table itself (containing all of the current cells) is provided to the client. As in every interaction with the delegated mapping structure, users should verify the attached proof. Verifying existence of an entry follows from the same method.
Internet-Draft            Delegated Mappings     April 2019

struct SetOperation {
    string application_identifier;
    opaque full_lookup_key<>;
    cell c;
}

struct SetRootOperation {
    rootentry e;
    bool remove;
}

union SetReturn switch (ReturnCode ret) {
    case SUCCESS:
        opaque empty;
    case ERROR:
        string reason;
}

Creating or updating a cell at a specified path requires once again
the full lookup key, as well as the new version of the cell to place.
The new cell must be well-formed under the validation checks
described in the previous section, else an "ERROR" is returned. For
example, updating a cell’s owner without a signature by the previous
owning key should not succeed. Both value cells and new/updated
delegations may be created through this method. Removing cells from
tables (after their commitment timestamps have expired) can be
accomplished by replacing the value or delegated namespace with an
empty value and setting the owner’s key to that of the table
authority. Asking the consensus layer to approve a new root entry
follows a similar process, although the application identifier and
lookup key is unnecessary (see "SetRootOperation"). Nodes can also
trigger votes to remove entries from the root key listing to redress
misbehaving applications.

4. Consensus

Safety is ensured by reaching distributed consensus on the state of
the tree. The general nature of a Merkle tree as discussed in the
previous section enables almost any consensus protocol to support
delegated mappings, with varying guarantees on the conditions under
which safety is maintained and different trust implications. For
example, a deployment on a cluster of nodes running a classic
Byzantine Fault Tolerant consensus protocol such as [PBFT] requires a
limited, static membership and can tolerate compromises in up to a
third of its nodes. In comparison, proof-of-work schemes including
many cryptocurrencies have open membership but rely on economic
incentives and distributed control of hashing power to provide
safety, and federated consensus algorithms like the Stellar Consensus

Protocol (SCP) [I-D.mazieres-dinrg-scp] combine dynamic members with real-world trust relationships but require careful configuration. Determining which scheme, if any, is the "correct" protocol to support authenticated delegation is an open question.

4.1. Interface

Explicit requirement for voting HERE

At a minimum, the consensus layer is expected to provide mechanisms for nodes to

1. Submit new values (commonly cell, but also root listing, updates) for consensus

2. Receive externalized values to which the protocol has committed

3. Validate values received from other nodes for each iteration of the protocol, as specified below

4. Voting mechanism for making root listing governance decisions

Specific protocols may require additional functionality from the delegated mapping layer, which should be implemented to ensure that valid updates are eventually applied (assuming a working consensus layer).

4.2. Validation

Incorrect (potentially malicious) updates to the Merkle tree should be rejected by nodes participating in consensus. Given the known prefix-delegation scheme, each node can apply the same validation procedure without requiring table-specific or application-specific knowledge. Validation also provides a simple mechanism for rate-limiting actors attempting to perform DoS attacks, as only the most recent change to a particular cell need be retained, and the total number of updates to any particular table or overall can be capped. Upon any modification to the delegation tables, a "SetOperation" or "SetRootOperation" as defined in the previous section, the submitted change to the consensus layer should:

1. Reference an existing application identifier in the root key listing and a valid table if applicable.

2. For updates to all cells:

   * contain an unmodified "create_time" or a current timestamp if a new cell
* contain a current "revision_time" in the case of an update
* set a "commitment_time" greater than or equal to the previous commitment
* not grant unlimited allowance (value 0) to delegate cells unless the delegating table also has an unlimited allowance
* result in a total table size ("valuecell" count + "delegatecell" allowances) less than or equal to the table allowance, if not unlimited
* contain a valid signature of the overall cell data by the table authority

3. For updates to value cells:
* be signed only by the current "owner_key" if the cell commitment has not yet expired, or by a new owner upon expiration
* have a lookup key in the table that belongs to the authority’s namespace
* not conflict with other cells in its table, breaking the prefix-delegation property

4. For updates to delegate cells:
* be signed by the table authority’s public key for new delegations or updates
* retain the same "namespace" and "deleege" value unless the "commitment_time" is expired
* contain a valid namespace owned by the authority delegating the cell
* not conflict with other values or delegations in the same table, breaking the prefix-delegation property

Only after a round of the consensus protocol is successful are the changes exposed to client lookups.
4.3. SCP

While consensus can be reached with many protocols, this section describes how the interface described above can be satisfied by SCP.

Updates to the delegation tables consist of the table change itself (the new version of the cell). Since SCP does not need specific knowledge of the format of these proofs, they directly form consensus on the opaque values submitted to the consensus layer. Once a combination of proofs are agreed to as outputs for a given slot, they are externalized to the mapping layer and applied to the local node’s table states. [I-D.mazieres-dinrg-scp] requires this layer to provide a _validity_ function that is applied to each input value, allowing nodes to detect misformed cells that violate the delegation semantics as defined by the previous subsection.

SCP asks the higher-level protocol to define a _combining_ function to compose multiple candidate values. In this application, we can take the union of valid updates proposed by the consensus nodes, rejecting duplicate updates to the same cell in favor of the most up-to-date timestamp.

Finally, SCP by specification uses federated voting to confirm values, which can be used directly to propose and validate modifications to the root key listing.

5. Security Considerations

The security of the delegation tables is primarily tied to the safety properties of the underlying consensus layer. Further, incorrect use of the public key infrastructure authenticating each mapping or compromise of a namespace root key can endanger mappings delegated by the key after their commitments expire.

6. References

6.1. Normative References


[Trillian]
6.2. Informative References

[I-D.mazieres-dinrg-scp]

[IP-blockchain]

[PBFT]


Acknowledgments

We are grateful for the contributions and feedback on design and applicability by David Mazieres, as well as help and feedback from many members of the IRTF DIN research group, including Dirk Kutscher and Melinda Shore.

This work was supported by The Stanford Center For Blockchain Research.

Authors’ Addresses

Sydney Li
Electronic Frontier Foundation
815 Eddy Street
San Francisco, CA 94109
US

Email: sydney@eff.org