General Usage Profile for LDAPv3 Replication
draft-ietf-ldup-usage-profile-01.txt

Status of This Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/lid-abstracts.txt.

The list of Internet-Drafts Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

Copyright Notice

Copyright (C) The Internet Society (2001). All Rights Reserved.

Abstract

Support for replication in LDAP directory systems is often one of the key factors in the decision to deploy them. But replication brings design constraints along with its benefits.

We discuss some of the factors that should be taken into consideration when designing a replicated directory system. Both programming and architectural/operational concerns are addressed.
1 Introduction

As LDAP directories become part of critical infrastructure in many applications, the need for extremely high reliability and availability becomes significant. And as directories are more widely used, the load on individual directory servers gets higher.

Distributed, replicated directories can reduce the problems of reliability and capacity. But applications that work well with a single, standalone directory will develop problems in a distributed environment unless both the application and the environment are carefully designed.

The particular areas of concern depend partly on whether the distributed directory is a single- or multi-master system, but there are many concerns that are common to both. In the remainder of the document, we have flagged some sections as being specific to single- or multi-master directories. Unflagged sections pertain to both.

This document addresses general issues. There may be additional drafts in the future that address specific applications.

2 Meta-data Considerations

Any LDAP directory contains meta-data as well as the user data in the directory. Examples of this meta-data include descriptions of the data in the directory (e.g. schema), policies for use of the data (e.g. access controls), and configuration/status information (e.g. replication agreements); this is not an exhaustive list.

Many types of meta-data are stored in the directory itself, accessible as regular data or as operational attributes. Since meta-data can appear in the directory, issues arise when it is replicated. Different issues arise if it is not replicated.

2.1 Schema Considerations

If the schema of one or more of the copies of a replica differs from the schema of the other replicas, then there exists the possibility of schema mismatch when data is exchanged between them. The extensibility feature of LDAP schema nearly guarantees that replica rings comprised of a heterogeneous mix of systems will not contain homogeneous schema because of directory vendors' built-in extensions. Because a given directory may not utilize its complete schema, not all of the potential
schema differences in a ring will result in a schema mismatch under replication.

Schema mismatch issues are further complicated by the possibility of replicating the "subschemaSubentry" itself. Some directories distribute schema changes through that mechanism. Currently there is no standard for LDAP schema representation. In the absence of such a standard, schema interoperability is not possible in the IETF sense. Designers should establish common schema on all servers holding a replica.

The following is a partial example list of schema mismatches:

1. Base syntax of an attribute type
2. Structure Rule of an object class
3. Optional vs. Mandatory attribute in an object class
4. Structural vs. Auxiliary in an object class
5. Object class not defined
6. Attribute type not defined
7. Object identifiers differs on an attribute type or on an object class
8. Type and number of attributes defined in a class
9. Multi-valued vs. Single valued attribute types
10. ACL format (and consequently, ACL calculus)
11. Matching rule of an attribute type
12. Naming collisions of attribute type names
13. Attribute name aliasing ("street" vs. "streetAddress" vs. "Strasse")

Schema mismatches that can cause data corruption in one or more of the replicas must result in meta-data (e.g. log entries) to comply with Requirement MM5 of [ReplicaReq]. However, not all schema differences produce corruption in all circumstances. Some schema differences may have little or no impact on the proper storage of replicated data. However, any time data is added to the directory replication may result in data corruption due to a schema mismatch. It is therefore recommended that all schema mismatches be removed from, or otherwise rationalized among, the replica-group, if possible. The tool described by requirement AM8 of [ReplicaReq] would help designers detect schema conflicts as early as possible. The other option that can be considered in this situation is the use of fractional replication to replicate only those attributes that do not have differences.

2.2 Replication Agreements

Replication Agreements are central to replication, as they allow configuration of most of the aspects of the replication process, including the triggers for replica cycles (from Requirement M1 in Huber, et al... [Page 4]
[ReplicaReq]), critical OID information (from Requirement M6 in [ReplicaReq]), and replication process parameters (Requirement M7 in [ReplicaReq]). Through the use of a standard schema (Requirement SC2) it is possible to replicate the replication agreement.

If a replication agreement includes replication credentials, the agreement should be read protected in the directory and transport of the replication agreement should be encrypted. Replication agreements are themselves distributed through the LDUP. They are therefore subject to same "loose consistency" problems caused by replication delay and deferred conflict resolution as other data. Even a temporary inconsistency among replication agreements may cause unbalanced replication and further inconsistency. As "multi-mastering" complicates "loose consistency" issues, avoidance of these issues by making all replication agreement changes through the same master (see Sections 4 and 5) is strongly advised.

2.3 Access Control

The following considerations complicate replication of Access Control Information:

- Access Control Information (ACI) is treated as though it were stored as attributes in the directory [ACModel]
- LDAP [RFC2251] declares that changes resulting from a single LDAP MODIFY are atomic (but see caveats for multi-master replication in Sections 4 and 5)
- The ACI affecting a given entry may not be part of that entry (it could be part of a group entry or part of an ancestor of the entry in question)
- The ACI cannot always be changed atomically with associated data changes
- The interaction of replication and partitioning is still unclear (i.e. what happens when access control policy is inherited from a superior partition [Inherit]).
- Thus, if you aren’t careful, you can leave windows where data is unprotected

To reduce risk:

- In all environments, access control changes should be made before adds and after deletes
- In multi-master environments, access control changes and the associated data changes should be made on same system.

Even when ACI is faithfully replicated (with the same transfer format) among heterogeneous members of a replica ring, there is no guarantee that an ACI change is expressed similarly everywhere in the ring. This
caveat is partly due to the open issues with respect to partitioning mentioned above, and partly due to vendor differences with regard to the expression of security policy.

2.4 Change Logs

Requirement G4 of [ReplicaReq] states that meta-data must not grow without bound. Since it is unrealistic to assume that meta-data won’t be needed during replication, designers must consider how and when meta-data can be purged.

Replicas that use connections with intermittent quality should use explicit replica cycle scheduling. Since the systems know when replication should have occurred, delayed replication can be detected and manual intervention initiated before the meta-data grows without bound. In extreme cases, it may be necessary to remove a replica from the replication ring and re-add them once better connectivity is available.

In a multi-master system, it is possible for a consumer to receive changes that cannot be applied. For example, a modify request for an entry may arrive before the add request that creates that entry. The replication system will typically queue this change and wait for additional changes.

3 Naming Considerations

A number of naming models have been proposed for directories ([RFC1255], [RFC2377], [CIMNames]), and many others have been implemented on an ad hoc basis. Each of these models specifies the naming attributes to be used and provides rules for using them. The naming scheme may also specify containment rules.

The naming plan applies to the directory as a whole, not the individual servers holding replicas. In a heterogeneous replicated environment, all of the replicating servers must be capable of supporting all of the rules for the naming plan in use for that directory.

Some directory implementations have naming constraints (e.g. containment rules, restrictions on attributes that can be used for naming). If such an implementation is part of a replicated directory, those constraints will have to be observed by all participating directories. If the environment contains implementations with incompatible constraints there is a major problem. This should be checked as early in the design phase as possible to avoid problems. Applications often have their own requirements on naming. If two independent applications start sharing previously separate directory...
information, care should be taken that the naming is consistent across
the applications. A difference in name form may be accepted through
LDUP without constraint violation, but nevertheless result in
unexpected behavior from a cross-application perspective. Consistent
naming is not only important to the directory, but to the applications
that consume directory information as well. Shared application access
often mandates shared naming in the directory.

4 Conflict Resolution Considerations

4.1 Consistent Access after Changes

Many operations on a directory are done as a set of steps. For
example, a new object may be created by one operation, and its values
may be filled in as part of a separate LDAP operation. An
administrator may add a user to a directory, and that user may then try
to log in using the new entry.

Replicated LDAP directories provide loose consistency [ReplicaReq]. A
new entry or a change to an existing entry will not reach all replicas
immediately; there will be some delay before changes are available on
all replicas. Changes made (e.g. adding a new user) on one physical
system may appear to be "lost" if checked on another physical system
before replication is complete.

In general, LDAP applications should be prepared to operate correctly
in the face of replication delays. In some cases, this means designing
to allow for delay. In the case of the newly created user, it should
be standard practice to ask the user to wait a while before trying to
use the entry. In the case where the new object must be filled in, the
application should make appropriate use of LDAP sessions to make sure
that the same server is reached for both operations.

As a general rule, an LDAP application should bind once and not unbind
until a complete set of related operations have been performed. To
achieve load balancing of write operations in a multi-master
environment, balancing the write-enabled connections is recommended
over balancing LDAP write operations.

In the single-master case, all write requests go to one server. If a
set of related reads and writes are done, they should all be done on
the master if possible. Ideally, only sets of related operations that
cannot include a write should go to one of the slave servers. But load
balancing concerns may make this impractical.

In some cases, related requests will deal with data in different
partitions that are not all available on a single server. In this
case, it is safer to keep sessions open to all servers rather than
closing the session with one server and opening one with another server.

It may not always be obvious to clients that they are using different servers. If a load distribution system is used between the client and the server, the client may find that a change request and a subsequent lookup are directed to different physical servers even though the original requests were sent to the same server name and/or address.

Since LDAP is session oriented, any load distribution system used should take sessions into account. Thus, keeping all related read and write requests within a single bind/unbind session should be the goal in this situation as well.

4.2 Conflict Resolution in Single-Master Systems

It is possible that resolution conflicts could occur in a single master replication system. Because requirement SM2 of [ReplicaReq] is a "SHOULD" and not a "MUST", it is possible for implementers to reorder changes. If changes are reordered, it is quite possible for a conflict to occur. Consider a case where schema changes are declared critical and must be moved to the front of the replication queue. Then the consumer servers might have to delete an attribute that still has values, and later process requests to delete the values of that attribute.

However, directory administrators may have scenarios where re-ordering of replication information is desirable. On a case-by-case basis, the directory administrator should make such decisions.

Many vendors may not implement conflict resolution for single-master replication. If such a system receives out-of-order changes from a system that does support them, replication errors will almost certainly occur. Designers should be aware that mismatches in the capabilities of replicating single-master directories could cause problems. Designs should not permit the master to re-order changes unless all slave copies are known to handle the situation correctly.

4.3 Problem Cases

4.3.1 Atomicity

The fact that replication does not guarantee the time order arrival of changes at a consumer allows situations where changes that were applied successfully at the supplier may fail in part when an attempt is made to apply the same change at the consumer. Examples appear below.
4.3.1.1 Locking

There is an entry with distinguished name DN that contains attributes X, Y, and Z. The value of X is 1. On replica A, a ModifyRequest is processed which includes modifications to change that value of X from 1 to 0 and to set the value of Y to "USER1". At the same time, replica B process a ModifyRequest which includes modifications to change the value of X from 1 to 0 and to set the value of Y to "USER2" and the value of Z to 42. The application in this case is using X as a lock and is depending on the atomic nature of ModifyRequests to provide mutual exclusion for lock access.

In the single-server case, the two operations would have occurred sequentially. Since a ModifyRequest is atomic, the entire first operation would succeed. The second ModifyRequest would fail, since the value of X would be 0 when it was attempted, and the modification changing X from 1 to 0 would thus fail. The atomicity rule would cause all other modifications in the ModifyRequest to fail as well.

In the multi-master case, it is inevitable that at least some of the changes will be reversed despite the use of the lock. Assuming the changes from A have priority per the conflict resolution algorithm, the value of X should be 0 and the value of Y should be "USER1". The interesting question is the value of Z at the end of the replication cycle. If it is 42, the atomicity constraint on the change from B has been violated. But for it to revert to its previous value, grouping information must be retained and it is not clear when that information can be safely discarded. Thus, requirement G6 in [ReplicaReq] may be violated.

The utility of locking mechanisms cannot be guaranteed with multi-master replication, and therefore results are likely to be misleading. As discussed further in section 6.1 below, its use in multi-master environments should be deprecated.

4.3.1.2 Partitioning

Partitioning also adds complexity. For example, suppose two servers, A and B, are members of Replica-ring X while servers B and C are members of replica-ring Y. It is possible to issue a ModifyRDN operation on server B that moves an entry from ring X to ring Y. Replication in ring X would delete the entry on server A while replication in ring Y would add the entry to server C. However, if another change on server C prevented the add operation from working (e.g. an RDN of a different GUID exists there already), then the change on server A is inconsistent and will need to be reversed. Other examples of cases of this class include group membership modification and access control scoping.
4.4 General Principles

The examples above discuss some of the most difficult problems that can arise in multi-master replication. Dealing with them is difficult and can lead to situations that are quite confusing to the application and to users.

The common characteristics of the examples are:

- Several directory users/applications are changing the same data
- They are changing the data at the same time
- They are using different directory servers to make these changes
- They are changing data that are parts of a distinguished name or they are using ModifyRequest to both read and write a given attribute value in a single atomic request

If any one of these conditions is reversed, the types of problems described above will not occur. There are many useful applications of multi-master directories where at least one of the above conditions does not occur. For cases where all four do occur, application designers should be aware of the possible consequences.

5 Failover Considerations

One of the major reasons to use directory replication is to improve reliability of the directory system as a whole. Replication permits hot- and warm-standby configurations to be built easily.

But there are some issues that must be considered during design. In this situation, single-master systems actually raise more concerns than multi-master. Both are addressed below.

5.1 Common Issues

In both the single- and multi-master cases, clients must be able to find an alternate quickly when a server fails. Some possible ways to do this are detailed in [FindingLDAP] and [LDAPinDNS]. If all else fails, a list of possible servers can be built into client applications. Designers should consider how clients are notified that the server is again available.

When the failed server comes back up, it is brought back into synchronization with the other servers and is ready to take requests. It is always possible that the failed server, if it was acting as a supplier, was unable to completely distribute its pending changes before removal from service, leaving its consumers in an inconsistent state. During the period between its removal from service and its eventual return, the inconsistency may have been compounded by further
application activity. As there is no current automatic mechanism to rectify the problem, the administrator should use whatever mechanism is available to compare the replicas for consistency as soon after the event as is reasonable.

Note that the process used to bring a failed server back into replication can also be used to add a server to a set of replicating servers. In this case, the new server might be initialized from a backed-up copy of the directory or it may acquire the entire DIB via replication. The former method is usually preferable when the directory is large.

5.2 Single Master Issues

In a single-master system, the master is a single point of failure, as all modification has to originate at the master server. When high availability is a requirement, a quick, automated failover process for converting a slave replica to a new master is desirable, as the failover time becomes a major factor in determining system availability. The considerations in section 5.1 apply here; clients must know how to find the new master or a new slave in case of failure.

To aid in promotion of a slave replica, the master could replicate control information and meta-data (including replication credentials) so that this information is available during failover promotion. This data may either be replicated on a single "failover designate" slave (which would become the master in during failover) or it could be replicated to all slaves. The first possibility has the advantage of minimizing the amount of extra replication while the second more robustly handles multiple failovers (i.e. failover of the newly promoted master to another slave before the original master has been restored). If this method is followed, data privacy mechanisms should be used to protect the replication session.

If data privacy mechanisms (e.g. encryption) are used to protect the replication session, the new master must have the necessary key information. Further this key information should be independent of the master that is using it (i.e. not tied to the IP address of the master server). If it is not independent, slave replicas could be pre-configured with the keys for all possible masters to reduce failover time.

Restoration of the failed or broken master can be handled in one of two ways:

- It could join the ring and function as a slave.
- It could join the ring and negotiate with the new master to synchronize and then take over as master.
In either case, clients need a way to know that a new server is available. If the broken master is returned to service as a slave, then the administrator must, external to LDUP, distribute and resolve whatever pending changes remained undistributed and unresolved from the time immediately before it was removed from service. If the broken master is returned as a new master, then care must be taken with its replacement master to ensure that all of its pending changes are distributed and resolved before it is returned to duty as a slave.

The slave replicas may also use the replication agreement to filter which master is allowed to submit changes. Such a model allows the slave servers to function correctly when the master server is "broken" and sending out incorrect updates. However, then it is necessary to update the replication agreement during the fail over process so that the slaves will accept updates from the new master. This is the case for both the original failure and the restoration of the restored master if that is how the restored master rejoins the replica ring.

5.3 Multi-Master Issues

Typically, a multi-master configuration is used when high availability is required for writes as well as reads in the directory. Because there are multiple active servers prepared to take write requests, there is no "switchover" time in this case. But clients still need to be able to find an alternate server, so the considerations of Section 5.1 apply here.

6 Other Issues

6.1 Locking

Section 4.3.1.1 discussed the problems that can arise when the "modify" command in LDAP is used for locking in a multi-master environment. There are more general principles at work there. LDAP is a distributed and replicated directory service that is typically described as "loosely consistent".

In loose consistency, the data should eventually converge among the replicas, but at any given instant, replicas may be in disagreement. This stipulation is the general result of:

1. the delay due to replication or extended replication intervals
2. the out of natural time order arrival of data at a replica
3. the temporary isolation of distributed systems from one another
4. failure to accept a change due to conflict resolution failure on a replica

Because of loose consistency, data preconditions to an LDAP operation may differ among replicas. Multi-mastering may exacerbate this situation, but single-mastering will not totally eliminate it if out-of-order replication is allowed (see Section 4.2). One must carefully assess the effect of loose consistency when evaluating operations that place specific preconditions on data to work correctly. Applications which depend on such operations may be better suited for transactional models and/or non-distributed data.

Distributed locking is one operation that depends on strict data preconditions. When data preconditions cannot be guaranteed, the lock is moot. The same principles hold for "atomic operations", defined here as any mix of allowable operations contained within the same LDAP PDU. RFC2251 requires that they either all fail or are applied as a unit. If strict data preconditions cannot be guaranteed, then the atomic operation may itself result in a further inconsistency requiring human intervention at one of the consumers.

6.2 Backup and Restore

Backup of a directory server should have the following goals:

(1) It can be unambiguously and faithfully restored.
(2) It is an internally consistent snapshot of an entire replica during the time interval it took to make it. This can only be achieved if the server is quiescent.
(3) Replication can resume on a machine restored from that backup without information loss.

Backup and restore of a single, operating directory server (rather than the entire directory system) presents its own challenges. "Loose consistency" works against the probability of achieving a loss-free copy of all the data in the directory, except under ideal conditions. Backup and restore of distributed directories is a decidedly easier task when the constraint of continuous availability is removed. In most cases, the removal of entire directory systems from write service is impossible, even for small periods of time. It is more practical to remove a single directory server from service to achieve a condition of quiescence. Once the write load on a server is removed, albeit from replication or an application, an internally consistent copy of the data may be obtained.

Replicas that have suffered catastrophic data loss may be restored from backups of working servers temporarily removed from service.
specifically to make a copy. This scenario illustrates the benefit of having three or more replicas in the system: no single point of write failure in the event that one of the replicas must be restored from a copy of another.

The M11 requirement from [ReplicaReq] allows an empty replica to be brought up to date through replication. This feature duplicates, but does not make entirely unnecessary, backup procedures on directory servers. Backups are still needed to recover data that has been lost to all replicas, either through normal LDAP updates or through some catastrophic event.

7 Impact of Non-LDAP Changes/Constraints

7.1 Changes Outside of LDAP

LDAP directories are typically built on top of some database or file system. Thus there are ways to change the data that do not go through the normal LDAP change mechanisms (e.g. ModifyRequest). If the data is modified outside of LDAP, the changes will not be checked for schema conformance nor will access controls be checked as the changes are made. Since both integrity and security checks are omitted, security can be adversely affected.

Also, many systems use the normal LDAP modification mechanisms to trigger replication. Changes made using non-LDAP mechanisms may not be replicated at all, leading to inconsistencies between replica copies.

7.2 Application Triggers

Directory servers commonly integrate one or more specific applications. To achieve this integration the directory server may intercept updates and run application-specific "trigger" code. Such triggers enforce directory invariants that cannot be expressed by the LDAP schema.

A simple trigger example is password policy enforcement. A directory server might interpret a request to replace the current value of the userPassword attribute with some new value as a request to first check that the new value conforms to the server’s password policy (e.g. the value is sufficiently long and complex) before storing the new value. Using this trigger the directory server voids the security risk associated with passwords that are easy to attack.

A more complex trigger example is password hashing. A directory server might interpret a request to replace the current value of the userPassword attribute with some new value as a request to compute one
or more secure hashes of the new value and store these hashes in one or more attributes, storing no value in the userPassword attribute. Using this trigger the directory server avoids the security exposure of storing the plaintext password.

Replication between directory servers with different application triggers will compromise directory integrity.

### 7.3 Policy Conflicts Across Servers

In addition to the discussions of ACI in Section 2.3 and triggering in section 7.2, LDUP replication can not (by its definition) handle replication of information that makes use of policy not expressible in the LDAP protocol. A prime example of this is security encoding of attributes (e.g. userPassword). This encoding is typically implementation specific and is not easily expressible via the LDAP protocol. Therefore replication of userPassword attributes between directory servers that use different encoding schemes will impede replication in a way that is not describable as schema or syntax mismatch. This is because of the bind-time policy semantics that are the true point of conflict.

Another case arises from the use of authentication levels in the Access Control Model [ACModel]. The authentication levels represent the strength of various forms of authentication rather than specific authentication mechanisms; the association between levels and mechanisms is left for administrators though some guidance may be provided [AuthMap]. In a replicated environment, administrators must be sure that the mappings of all replicating systems are compatible. If they are not compatible, access controls may have different effects on different replicas of a partition.

In general, any attribute with semantics that are outside the scope of what is expressible by the LDAP protocol could result in strange replication errors. Therefore, distributed directory implementers should (in the absence of a way to express such semantics) either strive for a homogeneous set of servers or ensure during acceptance testing that a new server can support the existing semantics of their directory.

### 8 Security Considerations

The document discusses issues that arise in replication. Some of these issues are security related (e.g. replication of access control information) and the security implications are discussed in the relevant sections.
9 Acknowledgements

This document owes a lot to discussions on the LDUP mailing list. In particular, the authors would like to thank Ed Reed, whose email to the mailing list drove much of section 6.1, and Mark Brown for identifying and generating text on the issues discussed in section 7.

10 References


Authors’ Addresses

Richard V. Huber
Room C3-3830
Huber, et al Expires January 2002
Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.