Brent 11-1

RPC Protocol Spec

Sun Remote Procedure Call Protocol Specification
Version 2
(a.k.a. Sun of Courier)

Bob Lyon

ABSTRACT

Herein lies a message protocol specification used in implementing Sun Microsystems' remote procedure call package. The protocol is specified using the xdr data specification language.

1. Introduction

This file is pumpkinseed:~blyon/rpc/memos/rpc_prot.txt.

This document assume that the reader is familiar with Sun's remote procedure call (rpc) and external data representation (xdr) packages. It does not attempt to justify rpc or its uses. Also, the casual user of rpc does not need to be familiar with the information in this document.

RPC Model

The remote procedure call model is similar to the local procedure call model. In the local case, the caller places arguments to a procedure in some well specified location; it then transfers control to the procedure and eventually receives control back from the procedure. At that point, the results of the procedure are extracted from some well specified location (like a result's register) and the caller continues execution.

The remote procedure call is very similar except that the one thread of control winds through two processes – one is the caller's process, the other is a server's process. That is, the caller process sends a *call message* to the server process and waits (blocks) for a *reply message*. The call message contains (among other things) the procedure's parameters. The reply message contains (among other things) the procedure's results. Once the reply message is received, the results of the procedure are extracted and caller's execution is resumed.

On the server side, a process is dormant awaiting the arrival of a call message. When one arrives the server process extracts the procedure's parameters, computes the results, sends a reply message, and then awaits the next call message.

Note that in this model, only one of the two processes are active at any given time. That is, the rpc protocol does not explicitly support multi-threading of caller or server processes.

Transports and semantics

The rpc protocol is independent from transport protocols. That is, rpc does not care how a message is past from one process to another. The protocol only deals with the specification and interpretation of messages.

Because of transport independence, the rpc protocol does not attach specific semantics to the remote procedures or their execution. Some semantics can be inferred (but should be explicitly specified) from the underlying transport protocol. For example, rpc message passing using UDP/IP is unreliable. Thus, if the caller retransmits call messages after short time-outs, the only thing he can infer from no reply message is that the remote procedure was executed zero or more times (and from a reply message, one or more times). On the other hand, rpc message passing

using TCP/IP is reliable. No reply message means that the remote procedure was executed at most once, whereas a reply message means that the remote procedure was exactly once.

(Note: At Sun, rpc is currently implemented on top of three transport protocols, two of which are TCP/IP and UDP/IP.)

Binding and Rendezvous Independence

The act of binding a client to a service is NOT part of the remote procedure call specification. This important and necessary function is left up to some higher level software. (The software may use rpc itself; see Appendix 3.) –

Implementors should think of the rpc protocol as the JSR of a network; the loader (binder) makes JSR useful and the loader itself uses JSR to accomplish its task.

Message Authentication

The rpc protocol provides the fields necessary for a client to identify himself to a service and vise versa. Security and access control mechanisms can be built on top of the message authentication.

2. Requirements

The rpc protocol must provide for the following functionally:

1. Unique specification of a procedure to be called.
2. Provisions for matching response messages to request messages.
3. Provisions for authenticating the caller to server and vise versa.

Besides these requirement, features which detect the following are worth supporting because of protocol roll-overs errors, implementation bugs, end user error, and general network administration:

1. Actual RPC protocol mismatches.
2. Remote program protocol version mismatches.
3. Protocol errors (like mis-specification of a procedure's parameters).
4. Reasons why remote authentication failed.

Remote Programs and Procedures

The rpc call message has three unsigned fields for remote program number, remote program version number, and remote procedure number. The three fields uniquely identify the procedure to be called. Program numbers are administered via some some central authority (like Sun). Once an implementor has program number, he can implement his remote program; the first implementation would most likely have the version number of 1. Because most new protocols evolve into better, stable and mature protocols, a version field of the call message identifies which version of the protocol the caller is using. Version numbers make speaking old and new protocols through the same server possible.

The procedure number identifies the procedure to be called. These numbers are documented in the specific program's protocol specification. For example, a file server's protocol specification may state that its procedure number 5 is "read" and procedure number 12 is "write".

Just as remote program protocols may change over several versions, the actual rpc message protocol could also change. Therefore, the call message also has the rpc version number in it (this field must be two (2)).

The reply message to a request message has enough information to describe the following error condition:

1) The remote implementation of rpc does speak protocol version 2. The lowest and highest supported rpc version numbers are returned.

2) The remote program is not available on the remote system.

3) The remote program does not support the requested version number. The lowest and highest supported remote program version numbers are returned.

4) The requested procedure number does not exist (this is usually a caller side protocol or programming error).

5) The parameters to the remote procedure appear to be garbage from the server's point of view. (Again, this caused by a disagreement about the protocol between client and server.)

Authentication

Provisions for authentication of caller to server and vise versa are provided as a wart on the side of the rpc protocol. The call message has two authentication fields, the *credentials* and *verifier*. The reply message has one authentication field, the *response verifier*. The rpc protocol specification defines all three fields to be the following opaque type:

```
enum auth_flavor {
        AUTH_NULL   = 0,
        AUTH_UNIX   = 1,
        AUTH_SHORT  = 2
        /* and more to be defined */
};

struct opaque_auth {
        union switch (enum auth_flavor) {
                default: string auth_body<400>;
        };
};
```

In simple English, any *struct opaque_auth* is an *enum auth_flavor* followed by a counted string, whose bytes are opaque to the rpc protocol implementation.

The interpretation and semantics of the data contained within the authentication fields is specified by individual, independent authentication protocol specifications. Appendix 1 defines three authentication protocols.

If authentication parameters were rejected, the response message contains information which states why they were rejected.

2.1. Program Number Assignment

Program numbers are given out in groups of 0x20000000 (536870912) according to the following chart:

0	-	1fffffff	defined by Sun
20000000	-	3fffffff	defined by user
40000000	-	5fffffff	reserved
60000000	-	7fffffff	reserved
80000000	-	9fffffff	reserved
a0000000	-	bfffffff	reserved
c0000000	-	dfffffff	reserved
e0000000	-	ffffffff	transient

The first group is a range of numbers administered by Sun Microsystems, and should be identical for all Sun customers. The second range is for applications peculiar to a particular customer. This range is intended primarily for debugging new programs. When a customer develops an application that might be of general interest, that application should be given an assigned number in the first range. The third group is for applications that generate program numbers dynamically. The final groups are reservered for future use, and should not be used.

The exact registration process for Sun defined numbers is yet to be established.

3. Other Uses (Abuses) of the RPC Message Protocol

The intended use of this protocol is for remote procedure calling. That is, each call message is matched with a response message. However, the protocol itself is a message passing protocol with which other (non-rpc) protocols can be implemented. Sun currently uses (abuses) the rpc message protocol for the following two (non-rpc) protocols: "Batching" (or pipelining) and "Broadcast Rpc". These two protocols are discussed (but not defined) below.

Batching

Batching allows a user to send an arbitrarily large sequence of call messages to a server; batching uses reliable bytes stream protocols (like TCP/IP) for their transport. In the case of batching, the client never waits for a reply from the server and (similarly) the server does not send replies to batch requests. A sequence of batch calls is usually terminate by a legitimate rpc call in order to flush the pipeline (with positive acknowledgement).

Broadcast Rpc

In broadcast rpc based protocols, the client sends an a broadcast packet to the network and waits for numerous replies. Broadcast rpc uses unreliable, packet based protocols (like UDP/IP) as their transports. Servers that support broadcast protocols only respond when the request is successfully processed, and are silent in the face of errors.

4. The RPC Message Protocol

This section defines the rpc message protocol in the xdr data description language. The message is defined in a top down style.

NB: This is an xdr specification. This is NOT C code.

```
enum msg_type {
        CALL = 0,
        REPLY = 1
};

/*
 * A reply to a call message can take on two forms: the message was either accepted or rejected.
 */
enum reply_stat {
        MSG_ACCEPTED = 0,
        MSG_DENIED = 1
};

/*
 * Given that a call message was accepted, the following is the status of
 * an attempt to call a remote procedure.
 */
enum accept_stat {
        SUCCESS = 0,                    /* The remote procedure was successfully executed */
        PROG_UNAVAIL = 1,               /* The remote machine does export the program number */
        PROG_MISMATCH = 2,              /* The remote machine does not support the version number */
      PROC_UNAVAIL = 3,                 /* The remote program does not know about the desired procedure */
        GARBAGE_ARGS = 4                /* The remote procedure could not make sense out of the parms */
};
```

```
/*
 * Reasons why a call message was rejected:
 */
enum reject_stat {
     RPC_MISMATCH = 0,    /* The rpc version number was not two (2) */
     AUTH_ERROR = 1       /* The caller was not authenticated on the remote machine */
};

/*
 * Why authentication failed:
 */
enum auth_stat {
     AUTH_BADCRED=1,              /* bogus credentials (seal broken) */
     AUTH_REJECTEDCRED=2,         /* client should begin new session */
     AUTH_BADVERF=3,              /* bogus verifier (seal broken) */
     AUTH_REJECTEDVERF=4,         /* verifier expired or was replayed */
     AUTH_TOOWEAK=5,              /* rejected due to security reasons */
};

/*
 * THE rpc message:
 * All messages start with a transaction identifier, xid. The xid is followed by
 * a two-armed discriminated union. The union's discriminant is a msg_type
 * which switches to one of the two types of the message. The xid of a REPLY
 * message always matches that of the initiating CALL message.
 * NB: The xid field is only used for clients matching reply messages with
 * call messages; the service side cannot treat this id as any type of
 * sequence number.
 */
struct rpc_msg {
     unsigned  xid;
     union switch (enum msg_type) {
          CALL:    struct call_body;
          REPLY:   struct reply_body;
     };
};
```

```
/*
 * Body of an rpc request call:
 * In version 2 of the rpc protocol specification, rpcvers must be equal to 2.
 * The fields prog, vers, and proc specify the remote program, its version,
 * and the procedure within the remote program to be called. These fields are
 * followed by two authentication parameters, cred (authentication credentials)
 * and verf (authentication verifier). The authentication parameters are followed
 * by the parameters to the remote procedure; these parameters are specified
 * by the specific program protocol.
 */
struct call_body {
      unsigned rpcvers;       /* must be equal to two (2) */
      unsigned prog;
      unsigned vers;
      unsigned proc;
      struct opaque_auth cred;
      struct opaque_auth verf;
      /* protocol specific parameters start here */
};

/*
 * Body of a reply to an rpc request.
 * The call message was either accepted or rejected.
 */
struct reply_body {
      union switch (enum reply_stat) {
            MSG_ACCEPTED:   struct accepted_reply;
            MSG_DENIED: struct rejected_reply;
      };
};
```

```
/*
 * Reply to an rpc request that was accepted by the server.
 * Note: there could be an error even though the request was accepted.
 * The first field is an authentication verifier which the server generates
 * in order to validate itself to the caller. It is followed by a union
 * whose discriminant is an enum accept_stat. The SUCCESS arm of the union is
 * protocol specific. The PROG_UNAVAIL, PROC_UNAVAIL, and GARBAGE_ARGS arms
 * of the union are void. The PROG_MISMATCH arm specifies the lowest and
 * highest version numbers of the remote program that are supported by the
 * server.
 */
struct accepted_reply {
        struct opaque_auth     verf;
        union switch (enum accept_stat) {
                SUCCESS: struct {
                        /* protocol specific results start here */
                };
                PROG_MISMATCH: struct {
                        unsigned low;
                        unsigned  high;
                };
                default: struct {
                        /*
                         * void. Cases include PROG_UNAVAIL, PROC_UNAVAIL,
                         * and GARBAGE_ARGS.
                         */
                };
        };
};

/*
 * Reply to an rpc request that was rejected by the server.
 * The request can be rejected because of two reasons - either the server is
 * not running a compatible version of the rpc protocol (RPC_MISMATCH), or
 * the server refused to authenticate the caller (AUTH_ERROR). In the case of
 * an rpc version mismatch, the lowest and highest supported rpc version numbers
 * are returned by the server. In the case of refused authentication, the
 * failure status is returned.
 */
struct rejected_reply {
        union switch (enum reject_stat) {
                RPC_MISMATCH: struct {
                        unsigned low;
                        unsigned high;
                };
                AUTH_ERROR: enum auth_stat;
        };
};
```

Appendix 1: Authentication Parameters Specification

As previously stated, authentication parameters are opaque, but open-ended to the rest of the rpc protocol. This section defines some "flavors" of authentication which have been implemented at (and supported by) Sun.

4.1. Null Authentication

Often calls must be made where the caller does not know who he is and the server does not care who the caller is. In this case, the auth_flavor value (the discriminant of the opaque_auth's union) of the rpc message's *credentials*, *verifier*, and *response verifier* is AUTH_NULL (0).

The bytes of the auth_body string are undefined. It is recommended that the string length be zero.

4.2. UNIX Authentication

(UNIX is a trademark of AT&T Bell Laboratories.)

The caller of a remote procedure may wish to identify himself as he is identified on a UNIX system. The value of the *credential's* discriminant of an rpc call message is AUTH_UNIX (1). The bytes of the *credential's* string encode the the following (xdr) structure:

```
struct auth_unix {
        unsigned   stamp;
        string     machinename<255>;
        unsigned   uid;
        unsigned   gid;
        unsigned   gids<10>;
};
```

The *stamp* is an arbitrary id which the caller machine may generate. The *machinename* is the name of the caller's machine (like "krypton"). The *uid* is the callers effective user id. The *gid* is the callers effective group id. The *gids* is a counted array of groups which contain the caller as a member.

The *verifier* accompanying the *credentials* should be of AUTH_NULL (defined above).

The value of the discriminate of the *response verifier* received in the reply message from the server may be AUTH_NULL or AUTH_SHORT (2). In the case of AUTH_SHORT, the bytes of the *response verifier's* string encode an *auth_opaque* structure. This new *auth_opaque* structure may now be passed to the server instead of the original AUTH_UNIX flavor credentials. The server keeps a cache which maps short hand *auth_opaque* structures (passed back via a AUTH_SHORT style *response verifier*) to the original credentials of the caller. The caller can save network bandwidth and server cpu cycles by using the new credentials.

The server may flush the short hand *auth_opaque* structure at any time. If this happens, the remote procedure call message will be rejected due to an authentication error. The reason for the failure will be AUTH_REJECTEDCRED. At this point, the caller may wish to try the original AUTH_UNIX style of credentials.

Appendix 2: Record Marking Standard (RM)

When rpc messages are passed on top of a byte stream protocol (like TCP/IP), it is necessary (or at least desirable) to delimit one message from another in order to detect and (possibly) recover from user protocol errors. Sun uses this RM/TCP/IP transport for passing rpc messages on TCP streams. One rpc message fits into one RM record.

A record is composed of one or more record fragments. A record fragment is a two-byte header followed by 0 to $2^{15}-1$ bytes of fragment data. (There are 8 bits in each byte or octet.) The bytes encode an unsigned binary number; the high-order byte precedes the low-order byte. The number encodes two values - a boolean which indicates whether the fragment is the last fragment of the record (bit value 1 implies the fragment is the last fragment) and a 15-bit unsigned binary value which is the length in bytes of the fragment's data. The boolean value is the high-order bit of the header; the length is the 15 low-order bits.

(Note that this record specification is *not* in xdr standard form!)

Appendix 3: Port Mapper Program Protocol

Introduction

The port mapper program maps rpc program and version numbers to UDP/IP or TCP/IP port numbers. This program makes dynamic binding of remote programs possible.

This is desirable because the range of reserved port numbers is very small and the number of potential remote programs is very large. By running only the port mapper on a reserved port, the port numbers of other remote programs can be ascertained by querying the port mapper.

The Port Mapper RPC Protocol

The protocol is specified by the xdr description language.

```
Port Mapper RPC Program Number: 100000
    Version Number: 1
    Supported Transports:
        UDP/IP on port 111
        RM/TCP/IP on port 111

/*
 * Handy tranport protocol numbers
 */
#define IPPROTO_TCP    6    /* Protocol number used for rpc/rm/tcp/ip */
#define    IPPROTO_UDP 17   /* Protocol number used for rpc/udp/ip */

/* Procedures */

    /*
     * Convention: procedure zero of any protocol takes no parameters
     * and returns no results.
     */
    0. PMAPPROC_NULL () returns ()

    /*
     * Procedure 1, setting a mapping:
     * When a program first becomes available on a
     * machine, it registers itself with the port mapper program on the
     * same machine. The program passes its program number (prog),
     * version number (vers), transport protocol number (prot),
     * and the port (port) on which it awaits service request. The
     * procedure returns success whose value is TRUE if the procedure
     * successfully established the mapping and FALSE otherwise. The
     * procedure will refuse to establish a mapping if one already exists
     * for tuple [prog, vers, prot].
     */
    1. PMAPPROC_SET (prog, vers, prot, port) returns (success)
        unsigned prog;
        unsigned vers;
        unsigned prot;
        unsigned port;
        boolean success;
```

```
/*
 * Procedure 2, Unsetting a mapping:
 * When a program becomes unavailable, it should unregister itself
 * with the port mapper program on the same machine.  The parameters
 * and results have meanings identical to those of PMAPPROC_SET.
 */
2. PMAPPROC_UNSET (prog, vers, dummy) returns (success)
     unsigned prog;
     unsigned vers;
     unsigned dummy1;  /* this value is always ignored */
     unsigned dummy2;  /* this value is always ignored */
     boolean success;

/*
 * Procedure 3, looking-up a mapping:
 * Given a program number (prog), version number (vers) and
 * transport protocol number (prot), this procedure returns the port
 * number on which the program is awaiting call requests.  A port
 * value of zeros means that the program has not been registered.
 */
3. PMAPPROC_GETPORT (prog, vers, prot, dummy) returns (port)
     unsigned prog;
     unsigned vers;
     unsigned prot;
     unsigned dummy;  /* this value is always ignored */
     unsigned port;  /* zero means the program is not registered */

/*
 * Procedure 4, dumping the mappings:
 * This procedure enumerates all entries in the port mapper's database.
 * The procedure takes no parameters and returns a "list" of
 * [program, version, prot, port] values.
 */
4. PMAPPROC_DUMP () returns (maplist)
     struct maplist {
          union switch (boolean) {
               FALSE: struct { /* void, end of list */ };
               TRUE: struct {
                    unsigned prog;
                    unsigned vers;
                    unsigned prot;
                    unsigned port;
                    struct maplist the_rest;
               };
          };
     } maplist;
```