Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 20, 2011

K. Ogawa
NTT Corporation
W. M. Wang
Zhejiang Gongshang University
E. Haleplidis
University of Patras
J. Hadi Salim
Mojatatu Networks
Oct 17, 2010

ForCES Intra-NE High Availability
draft-ietf-forces-ceha-00

Abstract

This document discusses CE High Availability within a ForCES NE.

Status of this Memo

Copyright Notice

Table of Contents

1. Definitions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

The following definitions are taken from [RFC3654]and [RFC3746]:

Logical Functional Block (LFB) -- A template that represents a fine-grained, logically separate aspects of FE processing.

ForCES Protocol -- The protocol used at the Fp reference point in the ForCES Framework in [RFC3746].

ForCES Protocol Layer (ForCES PL) -- A layer in the ForCES architecture that embodies the ForCES protocol and the state transfer mechanisms as defined in [RFC5810].

ForCES Protocol Transport Mapping Layer (ForCES TML) -- A layer in ForCES protocol architecture that specifically addresses the protocol message transportation issues, such as how the protocol messages are mapped to different transport media (like SCTP, IP, TCP, UDP, ATM, Ethernet, etc), and how to achieve and implement reliability, security, etc.

2. Introduction

Figure 1 illustrates a ForCES NE controlled by a set of redundant CEs with CE1 being active and CE2 and CEn-1 being a backup.

```
                           -----------------------------------------
                           | ForCES Network Element                |
                           |                        +-----------+  |
                           |                        |  CEn-1    |  |
                           |                        |  (Backup) |  |
     --------------   Fc   | +------------+      +------------+ |  |
     | CE Manager |--------+-|     CE1    |------|    CE2     |-+  |
     --------------        | |  (Active)  |  Fr  |  (Backup)  |    |
           |               | +-------+--+-+      +---+---+----+    |
           | Fl            |         |  |    Fp      /   |         |
           |               |         |  +---------+ /    |         |
           |               |       Fp|            |/     |Fp       |
           |               |         |            |      |         |
           |               |         |      Fp   /+--+   |         |
           |               |         |  +-------+    |   |         |
           |               |         |  |            |   |         |
     --------------    Ff  | --------+--+--      ----+---+----+    |
     | FE Manager |--------+-|     FE1    |  Fi  |     FE2    |    |
     --------------        | |            |------|            |    |
                           | --------------      --------------    |
                           |   |  |  |  |          |  |  |  |      |
                           ----+--+--+--+----------+--+--+--+-------
                               |  |  |  |          |  |  |  |
                               |  |  |  |          |  |  |  |
                                 Fi/f                   Fi/f
```

Fp: CE-FE interface
Fi: FE-FE interface
Fr: CE-CE interface
Fc: Interface between the CE Manager and a CE
Ff: Interface between the FE Manager and an FE
Fl: Interface between the CE Manager and the FE Manager
Fi/f: FE external interface

Figure 1: ForCES Architecture

The ForCES architecture allows FEs to be aware of multiple CEs but enforces that only one CE be the master controller. This is known in the industry as 1+N redundancy [refxxxx]. The master CE controls the FEs via the ForCES protocol operating in the Fp interface. If the master CE becomes faulty, a backup CE takes over and NE operation continues. By definition, the current documented setup is known as cold-standby [refxxxx]. The CE set is static and is passed to the FE

by the FE Manager (FEM) via the Ff interface and to each CE by the CE Manager (CEM) in the Fc interface during the pre-association phase.

From an FE perspective, the knobs of control for a CE set are defined by the FEPO LFB in [RFC5810], Appendix B. Section 3.1 of this document details these knobs further.

2.1. Document Scope

By current definition, the Fr interface is out of scope for the ForCES architecture. However, it is expected that organizations implementing a set of CEs will need to have the CEs communicate to each other via the Fr interface in order to achieve the synchronization necessary for controlling the FEs.

The problem scope addressed by this document falls into 2 areas:

1. To describe with more clarity (than [RFC5810]) how current cold-standby approach operates within the NE cluster.

2. To describe how to evolve the cold-standby setup to a hot-standby redundancy setup so as to improve the failover time and NE availability.

2.2. Quantifying Problem Scope

The NE recovery and availability is dependent on several time-sensitive metrics:

1. How fast the CE plane failure is detected the FE.

2. How fast a backup CE becomes operational.

3. How fast the FEs associate with the new master CE.

4. How fast the FEs recover their state and become operational.

The design goals of the current [RFC5810] choices to meet the above goals are driven by desire for simplicity.

To quantify the above criteria with the current prescribed ForCES CE setup in [RFC5810]:

1. How fast the CE side detects a CE failure is left undefined. To illustrate an extreme scenario, we could have a human operator acting as the monitoring entity to detect faulty CEs. How fast such detection happens could be in the range of seconds to days. A more active monitor on the Fr interface could improve this

detection.

2. How fast the backup CE becomes operational is also currently out of scope. In the current setup, a backup CE need not be operational at all (for example, to save power) and therefore it is feasible for a monitoring entity to boot up a backup CE after it detects the failure of the master CE. In this document Section 4 we suggest that at least one backup CE be online so as to improve this metric.

3. How fast an FE associates with new master CE is also currently undefined. The cost of an FE connecting and associating adds to the recovery overhead. As mentioned above we suggest having at least one backup CE online. In Section 4 we propose to zero out the connection and association cost on failover by having each FE associate with all online backup CEs after associating to the active CE. Note that if an FE pre-associates with backup CEs, then the system will be technically operating in hot-standby mode.

4. And last: How fast an FE recovers its state depends on how much NE state exists. By ForCES current definition, the new master CE assumes zero state on the FE and starts from scratch to update the FE. So the larger the state, the longer the recovery.

3. RFC5810 CE HA Framework

To achieve CE High Availabilty, FEs and CEs MUST inter-operate per [RFC5810] definition which is repeated for contextual reasons in Section 3.1. It should be noted that in this default setup, which MUST be implemented by CEs and FEs needing HA, the Fr plane is out of scope (and if available is proprietary to an implementation).

3.1. Current CE High Availability Support

As mentioned earlier, although there can be multiple redundant CEs, only one CE actively controls FEs in a ForCES NE. In practice there may be only one backup CE. At any moment in time only one master CE can control the FEs. In addition, the FE connects and associates to only the master CE. The FE and the CE PL are aware of the primary and one or more secondary CEs. This information (primary, secondary CEs) is configured on the FE and the CE PLs during pre-association by the FEM and the CEM respectively.

Figure 2 below illustrates the Forces message sequences that the FE uses to recover the connection in current defined cold-standby scheme.

```
         FE                   CE Primary        CE Secondary
         |                       |                    |
         |  Asso Estb,Caps exchg |                    |
       1 |<--------------------->|                    |
         |                       |                    |
         |       state update    |                    |
       2 |<--------------------->|                    |
         |                       |                    |
         |                       |                    |
         |                   FAILURE                  |
         |                                            |
         |         Asso Estb,Caps exchange            |
       3 |<------------------------------------------>|
         |                                            |
         |              Event Report (pri CE down)    |
       4 |------------------------------------------->|
         |                                            |
         |         state update from scratch          |
       5 |<------------------------------------------>|
```

Figure 2: CE Failover for Cold Standby

3.1.1. Cold Standby Interaction with ForCES Protocol

High Availability parameterization in an FE is driven by configuring the FE Protocol Object (FEPO) LFB.

The FEPO CEID component identifies the current master CE and the component table BackupCEs identifies the backup CEs. The FEPO FE Heartbeat Interval, CE Heartbeat Dead Interval, and CE Heartbeat policy help in detecting connectivity problems between an FE and CE. The CE Failover policy defines how the FE should react on a detected failure.

Figure 3 illustrates the defined state machine that facilitates connection recovery.

The FE connects to the CE specified on FEPO CEID component. If it fails to connect to the defined CE, it moves it to the bottom of table BackupCEs and sets its CEID component to be the first CE retrieved from table BackupCEs. The FE then attempts to associate with the CE designated as the new primary CE. The FE continues through this procedure until it successfully connects to one of the CEs.

```
       (CE issues Teardown ||    +-----------------+
          Lost association) &&   | Pre-Association |
         CE failover policy = 0  | (Association    |
             +------------>-->-->|   in            +<----+
             |                   | progress)       |     |
             |     CE Issues     +--------+--------+     |
             |     Association        |                  | CEFTI
             |       Response         V                  | timer
             |     ___________________+                  | expires
             |     |                                     |
             |     V                                     ^
           +-+-----------+                          +-------+-----+
           |             |                          |  Not        |
           |             |  (CE issues Teardown ||  |  Associated |
           |             |    Lost association) &&  |             |
           | Associated  |  CE Failover Policy = 1  | (May        |
           |             |                          | Continue    |
           |             |---------->------->------>|  Forwarding)|
           |             |                          |             |
           +-------------+                          +-------------+
                ^                                         V
                |                                         |
                |            CE Issues                    |
                |            Association                  |
                |            Setup                        |
                +_________________________________________+
```

Figure 3: FE State Machine considering HA

When communication fails between the FE and CE (which can be caused by either the CE or link failure but not FE related), either the TML on the FE will trigger the FE PL regarding this failure or it will be detected using the HB messages between FEs and CEs. The communication failure, regardless of how it is detected, MUST be considered as a loss of association between the CE and corresponding FE.

If the FE's FEPO CE Failover Policy is configured to mode 0 (the default), it will immediately transition to the pre-association phase. This means that if association is again established, all FE state will need to be re-established.

If the FE's FEPO CE Failover Policy is configured to mode 1, it indicates that the FE is capable of HA restart recovery. In such a case, the FE transitions to the not associated state and the CEFTI timer[RFC 5810] is started. The FE MAY continue to forward packets during this state. It MAY also recycle through any configured backup

CEs in a round-robin fashion. It first adds its primary CE to the bottom of table BackupCEs and sets its CEID component to be the first secondary retrieved from table BackupCEs. The FE then attempts to associate with the CE designated as the new primary CE. If it fails to re-associate with any CE and the CEFTI expires, the FE then transitions to the pre-association state.

If the FE, while in the not associated state, manages to reconnect to a new primary CE before CEFTI expires it transitions to the Associated state. Once re-associated, the FE tries to recover any state that may have been lost during the not associated state. How the FE achieves to re-synchronize its state is out of scope for the current ForCES architecture.

An explicit message (a Config message setting Primary CE component in ForCES Protocol object) from the primary CE, can also be used to change the Primary CE for an FE during normal protocol operation. In this case, the FE transitions to the Not Associated State and attempts to Associate with the new CE.

3.1.2. Responsibilities for HA

XXX: we may remove this section (not much value to overall discussion)

TML Level:

1. The TML controls logical connection availability and failover.

2. The TML also controls peer HA management.

At this level, control of all lower layers, for example transport level (such as IP addresses, MAC addresses etc) and associated links going down are the role of the TML.

PL Level:
All other functionality, including configuring the HA behavior during setup, the CE IDs used to identify primary and secondary CEs, protocol messages used to report CE failure (Event Report), Heartbeat messages used to detect association failure, messages to change the primary CE (Config), and other HA related operations described before, are the PL responsibility.

To put the two together, if a path to a primary CE is down, the TML would take care of failing over to a backup path, if one is available. If the CE is totally unreachable then the PL would be informed and it would take the appropriate actions described before.

4. CE HA Hot Standby

In this section we make some small extensions to the existing scheme to enable it to achieve hot standby HA. With these suggested changes we achieve some of the goals defined in Section 2.2, namely:

- How fast a backup CE becomes operational.

- How fast the FEs associate with the new master CE.

As described in Section 3.1, the FEM configures the FE to make it aware of all the CEs in the NE. The FEM also configures the FE to make it aware of which CE is the master and which are backup(s). The FE's FEPO LFB CEID component identifies the current master CE and table BackupCEs identifies the backup CEs. The FE only connects to the master CE and then proceeds to associate with it. The master thereafter controls the FE and receives events from it. This continues until there is communication failure between the FE and CE at which point the FE attempts to connect to a CE from the BackupCEs table until it succeeds to connect and associate with one listed CE.

It is recommended that at least one backup CE should be online. Doing so will improve how fast the backup CE will take to be operational (as opposed to bringing up a backup CE when we detect a master CE fault). If we assume that a CE implementation does state synchronization between CEs, then the cost of making the backup CE operational and ready to serve FEs; in such a case an associating FE could immediately become operational.

If we assume the presence of at least one backup CE online, we can improve how fast the FEs associate with a new master CE by making two changes:

The first change that needs to be made is to have the FE, soon after successfully connecting and associating with the master CE, to proceed and connect as well as associate with the rest of the CEs listed in the BackupCEs table.

By virtue of having multiple CE connections, the FE switchover to a new master CE will be relatively much faster. The overall effect is improving the NE recovery time in case of communication failure or faults of the master CE.

XXX: below paragraph needs more text discussion ..

The FE MUST respond to messages issued by only the master CE. This simplifies the synchronization and avoids the concept of locking FE state. The FE MUST drop any messages from backup CEs (XXX: Should we

log and increment some stat?).

XXX: below paragraph needs text discussion ..

Again for the sake of simplicity, asynchronous events and heartbeats are sent to all associated-to CEs. Packet redirects continue to be sent only to the master CE.

XXXX: We need to have an extra state for each CE (master, connected, associated, stats etc) on the FEPO - so probably another change to current FEPO components.

5. IANA Considerations

TBA

6. Security Considerations

TBA

7. References

7.1. Normative References

[RFC5810] Doria, A., Hadi Salim, J., Haas, R., Khosravi, H., Wang, W., Dong, L., Gopal, R., and J. Halpern, "Forwarding and Control Element Separation (ForCES) Protocol Specification", RFC 5810, March 2010.

7.2. Informative References

[RFC3654] Khosravi, H. and T. Anderson, "Requirements for Separation of IP Control and Forwarding", RFC 3654, November 2003.

[RFC3746] Yang, L., Dantu, R., Anderson, T., and R. Gopal, "Forwarding and Control Element Separation (ForCES) Framework", RFC 3746, April 2004.

[RFC5812] Halpern, J. and J. Hadi Salim, "Forwarding and Control Element Separation (ForCES) Forwarding Element Model", RFC 5812, March 2010.

Authors' Addresses

Kentaro Ogawa
NTT Corporation
3-9-11 Midori-cho
Musashino-shi, Tokyo 180-8585
Japan

Email: ogawa.kentaro@lab.ntt.co.jp

Weiming Wang
Zhejiang Gongshang University
149 Jiaogong Road
Hangzhou 310035
P.R.China

Phone: +86-571-88057712
Email: wmwang@mail.zjgsu.edu.cn

Evangelos Haleplidis
University of Patras
Patras
Greece

Email: ehalep@ece.upatras.gr

Jamal Hadi Salim
Mojatatu Networks
Ottawa, Ontario
Canada

Email: hadi@mojatatu.com