ForCES Element Bindings and Topology Discovery Protocol

draft-ansari-forces-discovery-00.txt

Status of this Memo

By submitting this Internet-Draft, I certify that any applicable
patent or other IPR claims of which I am aware have been disclosed,
and any of which I become aware will be disclosed, in accordance
with RFC 3668.

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/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at

This Internet-Draft will expire in January, 2005.

Conventions used in this document

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].

Abstract

This document describes a protocol for dynamically determining CE/FE
element bindings discovery and inter-FE topology discovery and
maintenance. Such a mechanism is needed for all these elements in
the set to behave as a single network element, as required by the ForCES architecture. The discovery protocol operates during both the pre-association and post-association phases of ForCES protocol.

Table of Contents

1. Definitions .......................................................... 2
2. Introduction .................................................................. 3
   2.1. Motivation ......................................................... 4
3. Element Bindings and Topology Discovery Protocol ................. 5
   3.1. Minimum requirements ........................................... 5
   3.2. Protocol Details .................................................. 6
      3.2.1. Element Bindings Discovery ............................. 7
      3.2.2. Topology Discovery and Maintenance .................. 8
   3.3. Protocol and Message Headers ................................ 9
      3.3.1. TLV definitions ............................................ 10
   3.4. Inter-FE Topology Discovery Examples ....................... 10
      3.4.1. Forwarding Elements connected in a daisy chain .... 11
      3.4.2. Forwarding Elements connected in a ring ............. 12
4. Security Considerations .............................................. 13
5. References .................................................................. 13
   5.1. Normative ........................................................ 13
   5.2. Informative ....................................................... 13
6. Authors’ Addresses ..................................................... 14
7. Full Copyright Notice ................................................ 14
8. Acknowledgement ...................................................... 15

1. Definitions

Element bindings discovery: Mechanism that determines the CE/FE bindings; i.e. which FE is controlled by which CE and vice-versa. The bindings also provide information such as whether the CE will act as a primary or as a secondary controller. Once the bindings are discovered, capabilities of the respective elements are exchanged. This represents the first phase of the discovery protocol. The result of this phase causes the FEs to maintain partial topology (neighbors) information.

Inter-FE topology discovery: Topology discovery relates to how the FEs are interconnected with each other with respect to packet forwarding. This represents the second phase of the discovery protocol.
Internet Draft     ForCES Discovery     July 2004

protocol. This is the complete view of the intra-NE network as seen by the CE.

Inter-FE topology maintenance: Once the inter-FE topology has been discovered, it has to be continuously monitored to ensure that any changes to the topology are reported to the corresponding CE. This represents the steady state and final phase of the protocol.

2. Introduction

The ForCES framework document [RFC 3746] describes how a set of control elements (CEs) and forwarding elements (FEs) interact with each other to form a single network element (NE). It describes the ForCES post-association phase protocol working across the Fp reference point between CE and FE. However, it pre-supposes that the process of discovering the CE/FE bindings and determining element capabilities is already performed by some other mechanism during the pre-association phase. Further, a normal NE operation requires that the CEs have a continuous and consistent view of the inter-FE topology. This document is meant to address these requirements for proper NE operation.

We describe a protocol, called the ‘discovery protocol’, which has three distinct modes or phases. Phase one, is the CE/FE bindings discovery process, wherein the elements learn of each other’s existence and determine the respective bindings, capabilities and appropriate interface to use for communication with each other. Phase two is the inter-FE topology discovery process, wherein the CE determines the complete inter-FE/intra-NE topology of the FE network along with any additional capabilities. The third and final phase is the topology maintenance phase, wherein any changes to the inter-FE topology during normal operation is reported to the CE so that it can update its view and make any necessary changes to the forwarding tables to reroute packets. Phase one of the protocol, viz. the elements binding discovery occurs during the pre-association phase, while the topology discovery and maintenance phases occur during the post-association phase of the ForCES protocol.

The proposed discovery protocol is required to scale to a very large number of forwarding elements in the NE, with minimal impact on the resources. The following list provides some of the features and goals of the protocol.

. Determine appropriate CE-FE bindings
. Determine respective element capabilities
. Determine connectivity between elements
. React to changes in link connectivity
ForCES Discovery

2.1. Motivation

The ForCES architecture defines a network element (NE) as a single managed entity made up of a collection of FEs and CEs and is indistinguishable from other network elements in the network. This NE model definition leads to two types of links from the network’s perspective: internal (or intra-NE) links and external (or inter-NE) links. Intra-NE links are purely internal to the NE and are not exposed to the external world; whereas, inter-NE (or external) links are exposed to the external world and over which routing adjacencies (such as OSPF, IS-IS, BGP etc.) can be formed. An NE can contain FEs that have zero or more internal/external links; e.g. in Fig. 1, FE3 has two internal links and no external links while FE1 and FE2 have two internal links and one external link each. Note that this link definition does not include the control link connecting the FE to CE.

In a traditional IP network, packets are forwarded from one network element to the next using destination based forwarding, where the packet’s destination IP address is matched against the IP prefix forwarding table and the appropriate output port/link is chosen. However, a packet entering a ForCES NE may travel multiple FEs within the NE before it exits onto the output link. This requires that the packet be correctly routed from the ingress FE to the egress FE. This internal routing requires knowledge of the physical FE inter-connection topology so that the CE can appropriately setup internal forwarding tables at each FE to forward the packets. Note that traditional routing protocols such as OSPF, IS-IS, BGP etc. do not operate on internal links and hence cannot provide the required routing information to the forwarding tables. We, therefore, use a simple topology discovery protocol that only operates on internal links and provides the necessary routing information to route the packet from the ingress FE to the egress FE. Further, the protocol should be able to route around internal link failures, if a path exists. This makes the NE highly available and resilient.
3. Element Bindings and Topology Discovery Protocol

Since the ForCES working group is currently focusing on the case where the control and forwarding plane elements are single-hop away (close proximity), we restrict the description to this focus area. The protocol can be extended to multi-hop case as well, but is currently beyond the scope of this document. The protocol is expected to work on point-to-point as well as multi-access links. The discovery protocol should have the option of discovering either the IP layer topology or data-link layer topology. The need for the latter arises when an NE contains FEs with non-IP interfaces/links connecting each other (e.g. using Ethernet directly). Choosing which topology discovery mechanism to employ is done through either configuration or negotiation during protocol initialization. We shall focus on the IP topology discovery mechanism in this document and leave the data-link topology discovery issue as future study item.

3.1. Minimum requirements
In order for the protocol to work as described, the following assumptions are made.

- Each element has been configured with their respective IDs (CEID, FEID)
- Each element has been configured with IP addresses for topology discovery.
- The protocol is enabled on the required interfaces.

Note that these are configuration requirements and are satisfied by the respective managers (CEM/FEM).

3.2. Protocol Details

Each ForCES Protocol Element (PE can be either CE or FE) periodically advertises Hello messages to all its neighbors on configured interfaces. All protocol messages travel a single PE hop and are not routed to other elements. The messages are sent as IP datagrams (multicast/broadcast, where applicable, or unicast) to the neighboring elements over IP configured interfaces. For non-IP/data-layer discovery, Subnetwork Access Network [SNAP] frames may be used to ensure data-link layer independency and inter-operability. Each FE maintains the neighbor relationships as long as the Hello messages are received from the neighbor. If it does not receive a pre-determined number of back-to-back Hello messages from a given neighbor, it deletes the neighbor relationship from its database and reports this information to its CE. This ensures that the CE has the complete and up-to-date information of the underlying topology of the FE network.

The Hello message contains information useful for discovering and maintaining neighbor relationships. It contains the PE ID, type of protocol element (i.e. CE or FE), interval between any two messages, interval for deeming a neighbor inactive, capability information etc. This is, in some ways, similar to the capabilities of the OSPF Hello protocol, but with additional functionalities such as determining element bindings and helping in inter-FE discovery.

The state machine for the discovery protocol is shown in Figure 2.
3.2.1. Element Bindings Discovery

Dynamic element discovery occurs as a result of checking the contents of the received Hello messages from the neighboring elements. As part of the initial FE and CE configuration done by their respective managers (FEM and CEM), each element has the list of associations from which it is willing to communicate and establish connections. An FE is only required to accept requests from those CEs that are in the pre-configured list. Similarly, the CE is only allowed to control those FEs, and hence accept connections from those FEs, that are on its pre-configured list.

On receiving the Hello message, the PE determines whether the advertising neighbor is a CE or an FE by checking a flag within the packet or the split address space PEID (as explained in [ForCESP]). Once it determines the PE type, it matches the received ID with the set of allowed IDs from its pre-configured set. Upon a successful match, it notes the interface for communication with the remote PE. Note that an FE only matches with the list of CEs and, a CE only matches with a list of FEs. Hello messages received on a FE from a
neighboring FE will not match with the list of CEs on that FE, and hence no bindings occur. The result of the bindings is to discover the appropriate interface to use to communication with the CE (or vice-versa) and the capabilities offered by the PE. Note that the Hello messages continue to be exchanged on the CE-FE interface even during post-association due to the possibility that a new CE or a FE may come up over that interface and we, hence, need a mechanism to auto-discover this.

3.2.2. Topology Discovery and Maintenance

Since the CE needs to maintain consistent and up-to-date view of the inter-FE topology, it needs to obtain real-time information of the status of the internal links connecting the FEs. Since the topology discovery and maintenance occurs during the post-association phase, we make use of the event-notification and query/response messages [ForCESP] of the ForCES protocol to provide this information to the CE. It is important to note that each FE only maintains partial topology information obtained through neighbor relationship maintenance through Hello messages. The partial topology view seen by each FE is only the neighbor connectivity information. The CE has to derive the complete topology view of the interconnected FEs based on the partial topology information reported by each FE (or queried by the CE). This ensures that only the CE maintains all the intelligence and the protocol operation on the FEs is very simple and has minimal overhead.

The periodic Hello messages maintain PE neighbor relationships. Any change in the link or neighbor status causes the FE to generate an asynchronous/event-driven message to the CE indicating this change. The mechanism defined in [ForCESP] is used for delivering event-driven messages from the FE to the CE. This involves the CE subscribing to such event-driven messages from the FE.

The CE aggregates the partial topology information received from each FE and generates the inter-FE topology. With this complete knowledge of the inter-FE topology, it can now make appropriate updates to the forwarding tables on each FE to route IP packets inside the NE; from ingress FE to egress FE, assuming that the destination of the packet is not the current NE itself. Any changes in the internal link states (and hence the topology) requires that the CE reconfigure the forwarding tables on the FEs based on the most up-to-date information to ensure that the packets do not end up in a black hole.
The protocol message consists of a fixed length header (16 bytes) followed by one or more optional TLVs. The format of the message is as follows.

```
0                   1                   2                   3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|   Version   |    Flags      |          Packet Length        |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|          Checksum           |          Port ID              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                         PE ID                               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|        TLV-Type             |          TLV-Length           |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                    TLV-Value à                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|                            ...                              |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

Version: Version number of this protocol. Currently acceptable value is 0x01

Flags: These indicate whether the message is sent by a FE, (0x01) or CE (0x02). More options may be defined in the future.

Packet Length: The length of the protocol message in bytes, including the header and the following TLVs.

Checksum: 16-bit checksum for the protocol message. The checksum calculation does not include the IP header.

Port ID: This indicates the port on which this packet was sent out by the sender; useful for topology construction.

PE ID: This is the 32-bit identifier of the sender. It could either be CE ID or FE ID, depending on the sender.

The protocol header is followed by one or many TLVs. The following TLV types are defined:

Hello TLV: Indicates the Hello message as exchanged by the neighbors. The TLV defines the common hello parameters such as the
Capabilities TLV: Provides the capabilities information - TBD

Vendor specific TLV: TBD

Length: The length, in octets, of the value field of the TLV that follows.

Value: A variable length octet string containing instance specific information of the TLV.

3.3.1. TLV definitions

TBD

3.4. Inter-FE Topology Discovery Examples

The following examples illustrate the topology discovery mechanism. For sake of simplicity, we assume that there is only one CE per NE. The FEIDs of the FEs in the topologies below are FE1, FE2, FE3, and FE4. Each FE has port IDs labeled alphabetically. This is also the case with the CE.

```
-------------------
|     CE       |
-------------------
A    B    C    
/     |     \
/     A \v
/     ------- B \v
/     +--> FE3 <+-- \v
/     C     |     |
/     -------  |     \
A \v     |     v A
-------B     |     v E
|   FE1 <+--    +--> FE2 
| |     <----------------->
------- C   D     
E ^ D|     C ^ | B
| |     | |
| v     | v
```

FE3 Control Element reachability Table
During the element-binding phase, each FE sends out hello messages with its FEID and Port ID (as outlined earlier) to all of its neighbors. Since each neighboring FE also listens to such messages, it receives the hello message and adds it to the neighbor association table, which may look like that shown in Fig. (a). In the topology discovery phase, which is post ForCES association stage, the CE queries each FE about its neighbor table. The FE responds back with the partial topology information available through its neighbor relationships. Both the query and the response are carried by the ForCES protocol. The CE collects the partial topology information from all the FEs in the NE and aggregates this information to fully construct the inter-FE topology. Any changes to the FE neighbor table, e.g. when a link state changes, generates a trigger/update message to the CE. The new information is used to recalculate the new topology and subsequently the CE takes appropriate actions based on the new topology; such as updating the packet forwarding tables on the FEs.

The following examples show the neighbor association tables.

3.4.1. Forwarding Elements connected in a daisy chain

```
+-----------+       +-----------+       +-----------+
|           |       |           |       |           |
A ^ ^ B    ^    ^ D
/   /     /   /  
A  v       v A     v A
```

Figure (a) Full mesh among FE1, FE2, and FE3
(b) Multiple FEs in a daisy chain

3.4.2. Forwarding Elements connected in a ring
4. Security Considerations

Like all protocols, this protocol will have security issues as well. These issues will be researched in detail in future draft versions.

5. References

5.1. Normative


5.2. Informative

6. Authors’ Addresses

Furquan Ansari
101 Crawfords Corner Road
Holmdel, NJ 07733
USA
Phone: +1 732-949-5249
Email: furquan@lucent.com

Thyaga Nandagopal
101 Crawfords Corner Road
Holmdel, NJ 07733
USA
Phone: +1 732-949-3739
Email: thyaga@lucent.com

Hormuzd Khosravi
Intel
2111 N.E. 25th Avenue JF3-206
Hillsboro, OR 97124-5961
USA
Phone: +1 503 264 0334
Email: hormuzd.m.khosravi@intel.com

7. Full Copyright Notice

"Copyright (C) The Internet Society (2004). All Rights Reserved. This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of developing Internet standards in which case the procedures for copyrights defined in the Internet Standards process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the Internet Society or its successors or assigns."
8. Acknowledgement

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