SAVI
Internet Draft
Intended status: Standard Tracks
Expires: October 2010

J. Bi
CERNET
G. Yao
Tsinghua Univ.
J. Wu
CERNET
Fred Baker
CISCO

April 18, 2010

SAVI Solution for Stateless Address
draft-bi-savi-stateless-00.txt

Status of this Memo

Copyright Notice

Abstract

This document specifies the procedure for creating bindings between a stateless address (including Stateless Autoconfiguration Address, manually configured non-static address, etc) and an anchor (refer to [SAVI-framework]) on SAVI (Source Address Validation Improvements) device. The bindings can be used to filter packets generated on the local link with forged IP addresses. Different from other proposed solution for stateless address (e.g., SAVI-FCFS), this solution follows RFC4862 to arbitrate the ownership of address. Supplemental binding processes are also specified to cover conditions that cannot be handled by control packet snooping.

Table of Contents

1. Introduction

This document describes the procedure for creating bindings between stateless address and anchor (refer to [savi-framework]). Other related details about this procedure are also specified in this document.

These bindings can be used to filter packets with forged IP addresses. How to use these bindings is specified in [savi-framework], depending on the environment and configuration. The definition and examples of anchor is also specified in [savi-framework].

The binding process is partially inspired by the work of SAVI-FCFS. Different from a data trigger based procedure in SAVI-FCFS, this specification mainly focuses on control panel triggered process. Supplement binding processes are designed to cover deficiency of control packet snooping.

2. 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 [RFC2119].

3. Mechanism Overview

The mechanism specified in this document is designed to provide a host level source IP address validation granularity, as a supplement to BCP38 [BCP38]. This mechanism is deployed on the access device (including access switch, wireless access point/controller, etc), and performs mainly NDP/ARP snooping to set up bindings between stateless IP addresses and corresponding anchors. The bindings can be used to validate the source address in the packets.

4. Basic Principle of Address Ownership Arbitration

In stateless scenario, nodes can "assign" address to themselves, and perform unreliable duplicate detection to check whether the address is being used. For IPv6 address, collision happens with very little probability because of large address space, thus the unreliable nature of DAD is not serious in reality. However, the unreliability of DAD troubles source address validation. It is very hard, if not impossible, for a SAVI device to determine which node can use one address when conflict happens.

A wise principle, "First Come First Served" is currently used by SAVI-FCFS to determine ownership of address. This principle is

correct, however, the problem is, how to determine which node is the first to use an address. Because the unreliable nature of DAD, the first one assigned itself an address, may not be the one first using the address to send traffic sniffed by the SAVI device. SAVI-FCFS requires device to send detection probes to determine whether an address is being used by another node. However, this effort may be vain, because malicious node can reply any probe, and the probe is still unreliable to reach the possible target node.

After a long time attempt, we finally find that because of the unreliability of DAD and ND, there is no perfect arbitration policy. In another word, if a arbitration policy is perfect, it must rely on a reliable DAD. We can prove this through a simple deduction:

Perfect arbitrate=>

The one having assigned itself an address first gets the ownership of the address=>

The one having performed DAD first gets the address=>

The arbitrator must know who is the first to perform DAD=>

The DAD must be reliable to be sniffed by the SAVI device.

And the deduction can be reversed.

In the end, we found we were building tower on the quick sand. It concerns nothing about whether choosing data trigger or control packet trigger. Unless stateless assignment changes to be reliable, no solution can be secure.

In this document, we decide to follow RFC4862, which is the only stand track on stateless address assignment. This means, if a node finishes a successful DAD, including the DAD is performed by SAVI device in case of data trigger, the address MUST be bound with it. Then the ownership conflict is actually handled through allowing one address to be bound with multiple anchors. The bindings are only removed when the lifetime expires, which equals prefix life time learned from RA. Then we achieve a simple solution, whose security is based on the reliability of RFC4862.

5. Background and Related Protocols

This mechanism is an instance of a SAVI [savi-framework] solution, specialized for stateless addresses, including IPv6 Stateless

Autoconfiguration address, manually configured non-static IPv6 and IPv4 address.

In IPv6, IPv6 Stateless Autoconfiguration [RFC4862] is a widely deployed address assignment mechanism. A node can generate an address autonomously, and use Duplicate Address Detection described in [RFC4862] to auto-configure this address. [RFC4862] clearly requires that duplicated address detection must be performed on any IPv6 address, including DHCPv6 address. This is the basis of this control packet snooping based SAVI solution.

[RFC4861] specifies the Neighbor Discovery protocol, which is an essential part of IPv6 address assignment.

IPv4 doesn't have stateless auto-configuration mechanism, because of the high collision probability of auto-generated address. However, in some scenarios, interfaces are allowed to be configured addresses with a specified prefix, instead of assigning each interface a static address. In such scenarios, the address assignment method is regarded as stateless in this document.

[RFC5227] specifies the procedure to detect IPv4 address collision. It is not required currently. However, this feature is useful to determine the uniqueness of an IPv4 address on the link. Considering not all the operating systems support [RFC5227], this solution is designed to be compatible with operating systems not complying with [RFC5227].

6. Terminology

Main terms used in this document are described in [savi-framework], [RFC4862], [RFC826] and [RFC5227].

7. Conceptual Data Structures

This section describes the possible conceptual data structures used in this mechanism.

Two main data structures are used to record bindings and their states respectively. There is redundancy between the two structures, for the consideration of separation of data plane and control plane.

7.1. Binding State Table (BST)

This table contains the state of binding between source address and anchor. Entries are keyed on the anchor and source IP address. Each

entry has a lifetime field recording the remaining lifetime of the entry, and a state field recording the state of the binding.

Anchor	Address	State	Lifetime
A	IP_1	Bound	65535
A	IP_2	Bound	10000
B	IP_1	Bound	1

Figure 1 Instance of BST

7.2. Filtering Table (FT)

This table contains the bindings between anchor and address, keyed on anchor. This table doesn't contain any state of the binding. This table is only used to filter packets. An Access Control List can be regarded as a practical instance of this table.

Anchor	Address
A	IP_1
A	IP_2

Figure 2 Instance of FT

8. Binding States Description

This section describes the binding states of this mechanism.

DETECTION A gratuitous ARP or Duplicate Address Detection Neighbor Solicitation has been sent by the host (or the SAVI device).

BOUND The address has passed duplicate detection and it is bound with the anchor.

9. Stateless Scenario

Figure 3 shows the main elements in a stateless address allowed network. Nodes generate address themselves without the assistance of

any other server. Other address assignment mechanisms may be also used in such network.

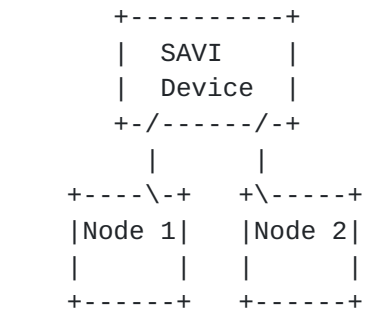

Figure 3 Stateless Scenario

10. Anchor Attributes

This section specifies the anchor attributes involved in this mechanism.

Anchor is defined in the [savi-framework]. Attribute of each anchor is configurable. In default, anchor has no attribute. An anchor MAY be configured to have one or more compatible attributes. However, an anchor MAY have no attribute.

If an anchor has no attribute, in this solution, Router Advertisement message from this anchor MUST be dropped. However, other packets SHOULD NOT be dropped.

10.1. SAVI-Validation Attribute

If and only if source address validation must be performed on the traffic from an anchor, this anchor MUST be set to have SAVI-Validation attribute. The filtering process on anchor with such attribute is described in section 12.

10.2. SAVI-RA-Trust Attribute

If and only if an anchor is associated with a trustable router, it SHOULD be set to have this attribute.

On a SAVI device enabled this solution, there may be no anchor with this attribute. This implies only link-local address is allowed, or prefix validation is not enabled, or only manually configured prefix is allowed. Router Advertisement message not coming from such anchors MUST be dropped.

10.3. SAVI-SAVI Attribute

If and only if an anchor is associated with another SAVI device, it SHOULD be set to have this attribute. All traffic from anchor with this attribute will be forwarded without check.

This attribute can also be set on other anchors if the administrator decides not to validate the traffic from the anchor. Note that DHCP server message and Router Advertisement will also be trusted.

This attribute is mutually exclusive with SAVI-Validation.

11. Prefix Configuration

Because Duplicate Address Detection doesn't have the function of checking the validity of the prefix of address, in this solution, it is SUGGESTED that correct address prefix SHOULD be configured. If the correct prefix is not configured, the SAVI device may bind anchors with addresses using bogus prefixes. Although a prefix level filtering mechanism, e.g., ingress filtering may be deployed on the layer 3 device of the network, it cannot prevent spoofing in the local link.

This document suggests set 3 prefix scopes:

IPv4 Prefix:

> The allowed scope of any kind of IPv4 addresses. It can be set manually.

IPv6 Prefixes:

> The allowed scope of SLAAC and manually configured IPv6 addresses. It can be set through snooping RA message from port with SAVI-RA-Trust attribute, DHCP-PD or manual configuration.

FE80::/64 MUST be set to a feasible prefix.

There is no need to explicitly present these prefix scopes. But these restrictions SHOULD be used as premier check in binding set up.

Refer to security consideration for other discussions.

12. Binding Set Up

This section specifies the procedure of setting up bindings based on control packet snooping. The binding procedure specified here is exclusively designed for anchor with SAVI-Validation attribute.

12.1. Process of Control Packet Snooping

12.1.1. Initialization

12.1.1.1. Trigger Event

A gratuitous ARP Request or Duplicate Address Detection Neighbor Solicitation is received from anchor.

12.1.1.2. Event Validation

The SAVI device checks the BST as follows:

1. Whether binding number limitation will be exceeded if a new binding entry is set up.

12.1.1.3. Following Actions

If the check is passed:

The packet MUST be forwarded.

An new entry MUST be generated, with state set to be DETECTION, and lifetime set to be MAX_ARP_DELAY or MAX_DAD_DELAY respectively.

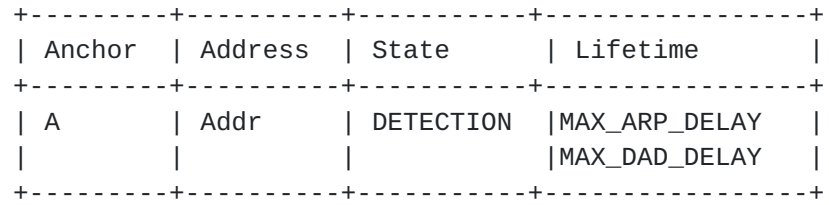

Figure 4 Binding entry in BST on detection

A new entry MUST be inserted into the FT.

12.1.2. State Transit from DETECTION

12.1.2.1. Trigger Event

A timeout event of an entry with state DETECTION occurs or an ARP Response or NA for an address in BST with state DETECTION is received.

12.1.2.2. Following Actions

If a timeout event of an entry with state DETECTION occurs, set the state of the entry to be BOUND. The lifetime of the entry is set to be the lifetime of corresponding prefix, which is learn from RA. If the address is link local address, the lifetime of the binding SHOULD be set to INFINITE.

Anchor	Address	State	Lifetime
A	Addr	BOUND	prefix lifetime

Figure 5 Binding entry in BST on finalization

If an ARP Response or NA for an address in BST with state DETECTION is received, remove the corresponding entry in BST and FT. The ARP Response or NA MUST be delivered to the client.

12.1.3. After BOUND

Once a binding entry is set up for an anchor, the binding will be used to filter packet with the anchor, as specified in section 13. The state of the binding entry will not be affected by any message.

If the lifetime of an entry with state BOUND expires, delete the entry in BST and Filter Table.

Switch port down event (or more general, anchor turns off-link) will change the corresponding entry, as described in section 15.

12.2. State Machine of DAD Snooping

The main state transits are listed as follows. Note that the anchor migration of binding entry is not included.

State	Message/Event	Action	Next State
-	Gra ARP/DAD NS	Generate entry	DETECTION
DETECTION	ARP RES/DAD NA	Remove entry	-
DETECTION	Timeout	Finish binding	BOUND
BOUND	Timeout	Remove entry	-

Gra ARP REQ: Gratuitous ARP REQUEST

ARP RES: ARP RESPONSE

DAD NS: Duplicate Address Detection Neighbor Solicitation

DAD NA: Neighbor Advertisement targeting at a tentative address

13. Supplemental Binding Processes

Supplemental binding processes are designed to cover situations that no DAD procedure can be sensed by SAVI device. The typical situations include: DAD NS loss, layer-2 path change and movement on local link without triggering DAD procedure.

This process is designed to avoid permanent blocking. It is not supposed that binding can be trigger whenever a data packet with unbound address is received. Generally a number of packets and more time are needed to trigger a binding.

At least one of the following techniques MUST be implemented in SAVI device which deploys this solution:

1. Rate-limited Data Triggered Binding Process

2. Counter Triggered Process

3. Extended Control Packet Snooping Process

Other techniques may be prudently chosen as alternative if found to have equivalent or even better function to avoid permanently blocking after discussion, implementation and deployment.

13.1. Rate-limited Data Triggered Binding Process

13.1.1. SAVI-DataTrigger Attribute

If data trigger binding process is implemented as the supplemental binding process, an additional anchor attribute, named SAVI-DataTrigger, MUST be implemented.

This attribute is mutually exclusive with SAVI-SAVI.

Data triggered binding process will be performed on the anchor with such attribute.

13.1.2. Data Triggered Binding Process

If an anchor is set to have SAVI-DataTrigger attribute, data packet whose source address is not bound with the anchor, may not be filtered directly; instead, the SAVI device will check whether the address can be used by the client on the local link with limited rate:

1. If the address has a local conflict, meaning the DAD on the address fails, the packet MUST be discarded. The DAD procedure is performed by the SAVI device, through sending two or more DAD NS probes. The format and delivery of the DAD NS is specified in section 15. If the DAD is successful, the address MUST be bound with the anchor, with a lifetime equal to lifetime of corresponding prefix.

The data triggered process MUST be rate limited to avoid Denial of Services attack against the SAVI device itself. A constant DATA_TRIGGER_INTERVAL is used to control the frequency. Two data trigger processes on one anchor must have a minimum interval time DATA_TRIGGER_INTERVAL. This constant SHOULD be configured prudently to avoid Denial of Services.

Data triggered process is not strict secure. The node with data-trigger anchor has the ability to use the address of an inactive node, which doesn't reply to the DAD probe.

13.2. Counter Triggered Process

13.2.1. SAVI-CounterTrigger Attribute

If counter triggered binding process is implemented as the supplemental binding process, an additional anchor attribute, named SAVI-CounterTrigger, MUST be implemented.

This attribute is mutually exclusive with SAVI-SAVI.

Counter triggered binding process will be performed on the anchor with such attribute.

13.2.2. Counter Triggered Process

In this process, a counter is used to record the number of filtered packets by this solution or all the enabled SAVI solutions on anchor with SAVI-CounterTrigger attribute. A constant TRIGGER_COUNT is set with the counter.

Whenever the counter reaches TRIGGER_COUNT, this event MUST be handled by the SAVI device. The SAVI device performs following steps:

1. Set the counter to 0;

2. Perform DAD process on the source address of the packet triggering this event, through sending two or more DAD NS probe as specified in section 15. If the DAD fails, the packet MUST be discarded. If the DAD is successful, a binding MUST be set up on the anchor.

3. This event MUST be announced to network administrator. For example, a SNMP trap may be triggered; or an alert on console interface may be generated.

The constant TRIGGER_COUNT MUST be prudently configured to fit the specified deployment scenario. In extreme situation, it can be set to 1.

13.3. External Control Packet Snooping Process

13.3.1. SAVI-ExtSnooping Attribute

If extended control packet snooping is implemented as the supplemental binding process, an additional anchor attribute, named SAVI-ExSnooping, MUST be implemented.

This attribute is mutually exclusive with SAVI-SAVI.

Extended control packet snooping process will be performed on the anchor with such attribute.

13.3.2. Extended Control Packet Snooping

In this snooping process, other than DAD messages, other types of control packets processed by processor of SAVI device, if the source address is not bound, may trigger the device to perform binding process.

The control messages that MUST be processed include: (1) address resolution Neighbor Solicitation; (2) Neighbor Advertisement; (3) neighbor unreachability detection; (4) Multicast Listener Discovery; (5) Address Resolution Protocol. Other ICMP messages that may be processed by intermediate device may also trigger the binding process.

The SAVI device MUST perform DAD to check if the address has a local conflict. The format and delivery of DAD probe is specified in section 15.

A minimum time interval EXT_SNOOPING_INTERVAL MUST be set to limit the rate of such triggering process.

Note that this process may not be able to avoid permanent block, in case that only data packets are sent by node. Generally, this mechanism is still practical, because data packet sending without control plane communication is rare and suspicious in reality. Normal traffic will contain control plane communication packets to help traffic setup and fault diagnosis.

14. Filtering Specification

This section specifies how to use bindings to filter packets. Because the Filtering Table is an allow-table, packet with source address not in the table will be filtered.

Filtering policies are different for data packet and control packet. ND messages that may cause state transit are classified into control packet. Neighbor Advertisement and ARP Response are also included in control packet, because the Target Address of NA and ARP Response should be checked to prevent spoofing. All other packets are considered to be data packets.

14.1. Data Packet Filtering

Data packets with an anchor which has attribute SAVI-Validation MUST be checked.

If the source of a packet associated with its anchor is in the FT, this packet SHOULD be forwarded; or else the packet MUST be discarded.

14.2. Control Packet Filtering

For anchors with SAVI-Validation attribute:

The source address of IPv6 NS and IPv4 gratuitous ARP MUST pass the check on FT.

The target address and source address in all the Neighbor Advertisement packets and ARP replies MUST also pass the checks on FT.

For other anchors:

All RA packets MUST be from anchor with the SAVI-RA-Trust attribute.

15. Format and Delivery of Probe Messages

The SAVI device MAY send detection probes on behavior of node to determine whether the assigned address is duplicated in case of data

trigger is enabled. Currently no other probes are designed in this solution.

15.1. Duplicate detection

Message Type: DAD NS, Gratuitous ARP Request

Format:

Link layer source - link layer address of host;

Link layer destination - For IPv6, use multicast address specified in [RFC3307]; For IPv4, use broadcast address;

IP source - Unspecified address for IPv6; The tentative address for IPv4;

IP destination - For IPv6, multicast address specified in section 5.4.2 of [RFC4861]; For IPv4, the tentative address;

Delivery:

MUST not be delivered to the host which the SAVI device is performing DAD on behavior of.

16. Binding Remove

If the lifetime of an entry with state BOUND expires, the entry MUST be removed.

17. Handle Anchor Off-link event

Port DOWN event MUST be handled if switch port is used as anchor. In more general case, if an anchor turns off-link, this event MUST be handled.

Whenever an anchor with attribute SAVI-Validation turns down, the bindings with the anchor MUST be kept for a short time.

To handle movement, if receiving DAD NS/Gra ARP request targeting at the address during the period, remove the entry.

If the anchor turns on-link during the period, recover bindings. It may result in some security problem, e.g., a malicious node immediately associates with the anchor got off by a previous node, then it can use the address assigned to the previous node. However,

this situation is very rare in reality. Authors decide not to handle this situation.

18. About Collision in Detection

The SAVI device may receive a response in detection. Some related details are specified here.

18.1. The Result of Detection without Host Aware

In case the SAVI device send detection packet instead of the host, the host will not be aware of the detection result. If the detection succeeds, there is no problem. However, if the detection fails, the packets from the host with the assigned address will be filtered out. This result can be regarded as a reasonable punishment for not performing duplicate detection and using a collision address. The SAVI device MAY choose to notice the client that the assigned address has been used, through a NA message. This mechanism is not required in this solution.

19. Filtering during Detection

In this mechanism, whenever a DAD NSOL is received, this address will be allowed immediately even before duplicate detection is completed. This design is in consideration of a host may start to send packets straightway without detection. Also this design is to be compatible with optimistic DAD [RFC4429].

However, this feature may allow an attacker to send quantities of packets with source addresses already assigned to other nodes.

20. Binding Number Limitation

It is suggested to configure some mechanism in order to prevent a single node from exhausting the binding table entries on the SAVI device. Either of the following mechanism is sufficient to prevent such attack.

1. Set the upper bound of binding number for each anchor with SAVI-Validation.

2. Reserve a number of binding entries for each anchor with SAVI-Validation attribute and all anchors share a pool of the other binding entries.

3. Limit DAD NSOL rate per anchor, using the bound entry number of each anchor as reverse indicator.

21. MLD Consideration

The SAVI device MUST join the tentative address multicast group whenever perform duplicate detection on behavior of host.

22. Link Layer Address Binding Toleration

As packet is possible to get lost on the link, and the first packet, which is generally DAD for link layer address, has higher lost probability because of the link initialization or authentication mechanism, a more tolerable mechanism for link local address MUST be used to avoid false positive.

Whenever a control message with link local source address is processed by this solution (ND, NA), if the address is not bound, the SAVI device MUST perform DAD to check the uniqueness of the address. If no collision is found, a binding entry for the link local address MUST be inserted into the binding table.

Other layer 3 control messages, including MLD, MAY also be used to trigger this process.

23. Handle Layer 2 Path Change

Layer 2 path change is an important challenge on this control plane based solution. The SAVI device MUST be sensitive to any layer 2 path change. Whenever a layer 2 control protocol frame, including STP, RSTP, TRILL, is received from some anchor, which announces a layer 2 incoming path is changed to the anchor, data packet trigger process MUST be enabled on the anchor for a period. Although generally such events can be handled through pre-configuration of data-trigger attribute, the future layer 2 protocol may be flexible and hard to handle through manual configuration.

24. State Restoration

If a SAVI device reboots accidentally or designedly, the states kept in volatile memory will get lost. This may cause hosts indirectly attached to the SAVI device to be broken away from the network, because they can't recover bindings on the SAVI device of themselves. Thus, binding entries SHOULD be saved into non-volatile storage whenever a new binding entry changes to BOUND state or a binding with state BOUND is removed, unless other alternatives specified here is implemented.

If binding is saved into non-volatile memory, immediately after reboot, the SAVI device MUST restore binding states from the non-

volatile storage. The lifetime and the system time of save process MUST be stored. Then the device MUST check whether the saved entries are obsolete when rebooting.

The possible alternative is:

If the network enables 802.1ag, the bindings can be recovered with the help of the first hop routers through snooping unicast Neighbor Solicitations sent by routers based on the Neighbor Table.

25. Constants

MAX_ARP_DELAY	Default 1s but configurable
MAX_DAD_DELAY	Default 1s but configurable
DATA_TRIGGER_INTERVAL	Device capacity depended and configurable
TRIGGER_COUNT	Device capacity depended and configurable

26. Security Considerations

For prefix level granularity filtering is the basis of host level granularity filtering, to learn and configure correct prefix is of great importance to this mechanism. Thus, it's important to keep RA and DHCP-PD secure. [draft-ietf-v6ops-ra-guard-03] describes a mechanism to improve the security of RA message.

27. IANA Considerations

There is no IANA consideration currently.

28. References

28.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

28.2. Informative References

[RFC3307] B. Haberman, "Allocation Guidelines for IPv6 Multicast Addresses", RFC3307, August 2002.

[RFC4861] T. Narten, E. Nordmark, W. Simpson, and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC4861, September 2007.

[RFC4862] Thomson, S., Narten, T. and Jinmei, T., "IPv6 Stateless Autoconfiguration", RFC4862, September, 2007.

[RFC5227] S. Cheshire, "IPv4 Address Conflict Detection", RFC5227, July 2008.

29. Acknowledgments

Thanks to Christian Vogt, Eric Levy-Abegnoli, Mark Williams, Erik Nordmark, Marcelo Bagnulo Braun, Alberto Garcia, Jari Arkko, David Harrington, Pekka Savola, Xing Li, Lixia Zhang, Robert Raszuk, Greg Daley, Joel M. Halpern, Mikael Abrahamsson, John Kaippallimalil and Tao Lin for their valuable contributions.

Authors' Addresses

Jun Bi
CERNET
Network Research Center, Tsinghua University
Beijing 100084
China
Email: junbi@cernet.edu.cn

Guang Yao
CERNET
Network Research Center, Tsinghua University
Beijing 100084
China
Email: yaog@netarchlab.tsinghua.edu.cn

Jianping Wu
CERNET
Computer Science, Tsinghua University
Beijing 100084
China
Email: jianping@cernet.edu.cn

Fred Baker
Cisco Systems
Email: fred@cisco.com