Network Working Group
Internet-Draft
Expires: July 30, 2005

J. Arkko
Ericsson Research NomadicLab
C. Vogt
University of Karlsruhe
January 26, 2005

A Taxonomy and Analysis of Enhancements to Mobile IPv6 Route Optimization
draft-irtf-mobopts-ro-enhancements-00.txt

Status of this Memo

Copyright Notice

Abstract

The Mobile IPv6 protocol favors sending packets via the minimum routing path between a mobile node and its correspondent node over sending them through a home agent. This feature is called route optimization. Route optimization requires authentication and authorization of initially unacquainted and untrusted parties--a requirement that was rather new to the Internet community at the time Mobile IPv6 was designed. To solve the problem, the so-called return-routability procedure was built into Mobile IPv6. It lets the mobile node retrieve from the correspondent node two cryptographic tokens, which the mobile node can use to authenticate itself and prove its presence at a claimed new location after movement. Recently, a number of improvements or optional alternatives have been suggested to the standard procedure. Many of these improvements attempt further reduction of signaling messages and latency, but other improvements such as better security have also been suggested. This paper summarizes the goals for enhancements to route-optimization, discusses the security threats that such enhancements must consider, and categorizes the techniques that one can use for optimization. The paper highlights the key ideas of various recent proposals, and it evaluates the performance gain that such proposals can yield. It also discusses how significant enhancements to Mobile IPv6 are compared to ongoing optimization work in other parts of the network stack. Finally, the paper identifies needs for additional research.

Table of Contents

1. Introduction

An IP address traditionally combines identification and location semantics. The address prefix locates a node's point of network attachment. It is used by routers to forward IP packets towards the correct destination. At the same time, existing transport protocols and applications, commonly termed "upper layers", use the IP address as part of a session identification. This naturally rules out mobility: Whenever a mobile node moves from one IP-attachment point to another, its IP address must change to reflect the new location. The new "identity", however, causes sessions at upper layers to abort.

Protocol designers thus had to decide whether to change transport protocols and applications, or to come up with a new IP-layer protocol that could separate location from identification functionality in a way transparent to upper layers. Due to the prevalence of TCP and the significant base of existing applications, most people opted for the latter approach. Mobile IPv6 [29], its IPv4 counterpart [28], and the Host Identity Protocol [9] are three dominant mobility-management protocols that the IETF has developed to facilitate the continued use of existing transport protocols and applications in an Internet with mobility support. This document focuses on Mobile IPv6.

Mobile IPv6 uses two IP addresses per mobile node in an attempt to separate location semantics from identification semantics: a transient "care-of address" is used for the purpose of routing. It is re-configured whenever the mobile node moves to a new IP-attachment point. A static "home address" is configured with the network prefix from a non-mobile "home agent's" network. The home address doesn't change when the mobile node moves, and it can be used for session identification at upper layers.

The mobile node keeps the home agent up to date about its current care-of address. In the event that packets are sent to the mobile node's home address, the home agent captures them and tunnels them to the mobile node's care-of address. In the opposite direction, the mobile node may tunnel packets to the home agent who, in turn, decapsulates them and forwards them on to the correspondent node. This behavior was termed "bidirectional tunneling". It works fine even if the correspondent node is unaware of its peer being mobile. The correspondent node can just use the home address, and the home agent will take care that packets find their way to the mobile node's actual location. Obviously, this entails a lot of routing overhead in many common scenarios.

For better routing efficiency, Mobile IPv6 defines a second mode,

"route optimization", that allows two nodes to directly communicate. Route optimization requires the correspondent node to be aware of the mobile node's current care-of address. The mobile node informs the correspondent node whenever its care-of address changes.

All signaling between a mobile node and its home agent is authenticated, and optionally encrypted, through IPsec. The IPsec security associations can either be manually configured into the nodes, or they can be dynamically derived through IKE. The mobile node and home agent must also be configured with the material they need to identify themselves, and the home agent must be able to authorize a mobile node to use a particular home address.

Preconfiguration of home agents and mobile nodes requires administrative labor, but it is doable, because the association between a mobile node and its home agent, or set of potential home agents, is typically known in advance. On the other hand, when route optimization is used between an arbitrary pair of nodes, there is generally no relationship between the two nodes prior to communication. Empowering a node--not necessarily a mobile one--to redirect packets from one IP address to another hence poses two questions:

- When the correspondent node receives a command to redirect a mobile node's packets, how can the correspondent node be sure that it is the legitimate mobile node, rather than a malicious third node, which has send this command?
- How can the correspondent node rely on the mobile node actually being present at the IP address to which packets are to be redirected?

The first question identifies the need for a mobile node to authenticate itself during a correspondent registration. Without such authentication, a malicious node could interfere with a packet flow of another node, redirecting the flow to its own location for inspection purposes, or redirecting it to a random IP address for the purpose of denial of service against the legitimate recipient. The second question refers to spoofed care-of addresses: Probing a mobile node's presence at a care-of address is important to prevent malicious parties to redirect packets to other nodes that neither expect nor want those packets.

A variety of approaches have been proposed to solve the above-mentioned issues for the case of route optimization. People finally elected the "return-routability procedure" as a default mechanism for Mobile IPv6. The return-routability procedure delivers a pair of secret tokens to a mobile node's home and care-of addresses. The mobile node needs these tokens to prove that it is

the legitimate owner of the home address and that it is reachable at the care-of address. (Actually, the return-routability procedure is less strict: It only determines whether a node is on the path towards the two addresses, rather than that it actually holds the two addresses. This is a compromise that the procedure accepts.)

The return-routability procedure is run right before a mobile node registers a new care-of address with a correspondent node. It is also run periodically in case the mobile node does not move for a while. The advantage of the return-routability procedure is that it is lightway and does not require any sort of pre-shared authentication material. Moreover, it can be implemented in a stateless way at the correspondent node's side. On the other hand, the return-routability procedure usually consumes one round-trip time, which comes into addition to the rest of any pending registration. This can lead to a handover delay unacceptable for many real-time or interactive applications like Voice over IP (VoIP) and audio or video streaming. Also, the periodic repetitions imply a hidden signaling overhead that may interfere with mobile nodes who intend to sleep during times of inactivity. Finally, the security level of the return-routability procedure can be increased. It limits vulnerabilities to attackers that are on the path from the correspondent node to the mobile node or to the home agent. The residual vulnerabilities are similar to those that exist anyway in an Internet without mobility support. But still, mechanisms that use stronger, possibly cryptographic authentication can provide a higher level of security than the return-routability procedure does.

This paper describes and classifies strategies that can enhance or optimize Mobile IPv6 route optimization. Following this introduction, Section 2 discusses which new security threats mobility-management protocols need to take into account. Section 3 explains the current route-optimization protocol, identifies the goals and assumptions based on which it was developed, and briefly analyzes its security properties. A number of potential goals for enhancements (such as reducing latency) are discussed in Section 4. Section 5 reviews techniques that can be used to enhance or optimize Mobile IPv6 route optimization. Section 6 discusses how these techniques are applied in existing enhancement and optimization proposals, evaluates some of these proposals, and identifies opportunities for further research. The paper concludes in Section 8.

2. Mobility-Related Security Threats

Mobile IPv6 allows a node to redirect those packets, that a correspondent node would otherwise send to one IP address (the home address), to a second IP address (the care-of address).

Unfortunately, the ability for redirection can also be misused by a malicious node for an arbitrary pair of IP addresses unless appropriate precautions are taken.

Overall, there are three major families of mobility-related threats: impersonation attacks, resource-exhaustion attacks, and flooding attacks. The following subsections take a closer look at each of the categories. Threats are described in the light of Mobile IPv6, but some of them apply to other mobility-management protocols as well.

2.1 Impersonation Attacks

The probably most obvious issue with mobility is to ensure that only a mobile node itself has the ability to change its care-of address. If care-of-address registrations were unauthenticated, an attacker could easily impersonate an arbitrary victim. For instance, the attacker could contact the victim's correspondent node and register its own IP address on behalf of its victim. The correspondent node would assume that the victim's care-of address has changed, and it would redirect all packets intended for the victim to the attacker instead. The attacker could forward the packets to the victim after analyzing, or even tampering with, their payloads. In a related offense, the perpetrator could simply cause havoc at its victim by directing the victim's packets to a random or non-existent IP address. These attacks are jointly referred to as "impersonation attacks". Impersonation attacks can be prevented through proper authentication techniques that keep an outsider from assuming another node's identity.

It is important to recognize that impersonation attacks not only impact those nodes that have an interest in mobility. Although the attacker makes the correspondent node believe that the victim is mobile, neither the attacker nor the victim do have to be mobile. Indeed, mobile nodes, non-moving nodes with mobility support, as well as traditional stationary nodes are potentially endangered because they all share the same IPv6 identifier namespace. (Actually, even IPv4 nodes are jeopardized when IPv4-to-IPv6 translation occurs on the path between these nodes and their correspondent peers.) This

unfolds the need for mandatory protection of mobility-related signaling in order to safeguard the Internet community as a whole.

Beyond their large group of potential victims, mobility-related impersonation attacks allow an attacker to choose the location from where to wage its attack. For example, the impersonator could position itself at a place where it is easier to inject spoofed care-of-address registration packets into the network than anywhere on the direct path between the victims. The attacker may also move to a place where it can remain unrecognized. In contrast to this, in

the non-mobile Internet that we have today, an attacker can only listen to or tamper with packets while it is on the path between its victims. Similarly, a mobility-management protocol may give the attacker the possibility to shift the time for its attack. The attacker might be able to register false care-of addresses even before its victims' conversation begins, or attack a network long after it has visited it. In the non-mobile Internet, an attacker must strike at the same time as its victims communicate. The ability to choose the location and time for an attack constitutes a dangerous new degree of freedom for the attacker.

2.2 Resource-Exhaustion Attacks

Mobility support at correspondent nodes can become an issue if it takes a lot of processing capacity to handle an incoming care-of-address registration. During times of increased signaling load, a correspondent node may thus end up having to commit a significant fraction of its resources to mobility-related transactions. What is worse, an attacker may take advantage of this vulnerability. It could swamp the correspondent node with large quantities of bogus registrations messages, keeping it from doing useful work. Such denial-of-service attempts are called "resource-exhaustion attacks". Clearly, if mobility support is to be implemented on a large basis, handling care-of-address registrations must be lightweight in order to lessen the susceptibility to resource exhaustion. Another effective technique is to defer resource commitment until late in the registration process: Once the registrant has proven its identity or shown that it is willing to invest resources itself, it is less likely malicious. As a last resort, busy Internet servers should limit the resources they devote to registration processing, and they may give preference to those mobile nodes they know or have recently had meaningful communications with.

It is worthwhile to stress the trade-off between effectiveness of signaling authentication and resilience against increased signaling load. On one hand, a strong authentication mechanism can effectively prevent certain impersonation attacks. On the other hand, the resources a correspondent node must spend on the verification of a registering node's authenticity increases with the complexity of the authentication algorithm. The susceptibility to resource exhaustion thus grows with the level of protection against impersonation attacks.

2.3 Flooding Attacks

A third mobility-related security threat emanates from redirection-based flooding attacks. Redirection-based flooding

attacks are characterized by a victim being bombarded with unwanted packets at a rate that the victim, and possibly the victim's access network, cannot handle. As with impersonation attacks, it is important to compare existing flooding attacks in today's non-mobile Internet with redirection-based flooding attacks that could be made possible through an insecure mobility-management protocol.

Three types of flooding attacks can be identified in today's Internet. The simplest one is a "direct flooding attack". Here, the attacker itself sends bogus packets to the victim. In an indirect "reflection attack", the attacker tricks a third node, the "reflection point", to send the packets. It typically uses a known protocol vulnerability to make the reflection point generate these packets [38]. For example, the attacker may send spoofed ICMP Echo Request packets to the reflection point, using its victim's IP address in the packets' IPv6 Source Address field. For each such request, the reflection point generates an ICMP Echo Reply message, which it sends "back" to the victim. The advantage of a reflection attack over a direct flooding attack is that the attacker is usually harder to track when flooding traffic comes from a third node. Another example for a reflection attack is TCP-SYN flooding. Here, the attacker sends TCP SYN packets, again with false source addresses, to the reflection point, which in turn sends TCP SYN-ACK packets to someone who does not expect these packets. Since most TCP servers are configured so that they re-send a TCP SYN packet multiple times when failing to receive an acknowledgement, this reflection attack can even produce a small amplification. Gaining higher amplification in today's Internet necessitates more complex

strategies like "distributed flooding attacks". In a distributed flooding attack, the attacker typically gains control over other nodes by spreading viral software. Then, at a certain point of time, infected nodes simultaneously commence a joint flooding attack against a common victim.

The introduction of mobility support renders amplified flooding attacks much less complex. Suppose a mobile node is allowed to change its care-of address without having to evidence that it is present at the new care-of address. Then, an attacker can subscribe, through its own IP address, to a large data flow (e.g., a video stream) offered by some server on the Internet. The attacker can easily accomplish the initial handshake procedure with the server while it uses its own IP address. Once data is flowing, the attacker can redirect the flow to the IP address of an arbitrary victim. The attacker can use the sequence numbers learned during the initial handshake procedure in order to spoof acknowledgements for packets that it assumes the server has sent to the victim. In this attack, not the attacker, but a faithful server on the Internet can be made generate packets used for an attack. The server does not have to be

infected with compromised code, and neither the victim nor the server has to be mobile. The attacker produces as little as spoofed feedback information to keep the data flow alive. To make matters worse, the attacker can redirect data flows from multiple servers to the victim.

Support for Mobile IPv6 route optimization is recommended to all IPv6 nodes [11]. The base of correspondent nodes that an attacker could exploit for a redirection-based flooding attack would therefore be immense. Also note that no distribution of viral software would be necessary. The severity of this new type of flooding is that it would provide potentially unbounded amplification at comparably low cost.

3. Mobile IPv6 Route Optimization

Route optimization requires the mobile node to register its current care-of address with both its home agent and correspondent node. The process of doing so is called a "home registration" and a "correspondent registration", respectively.

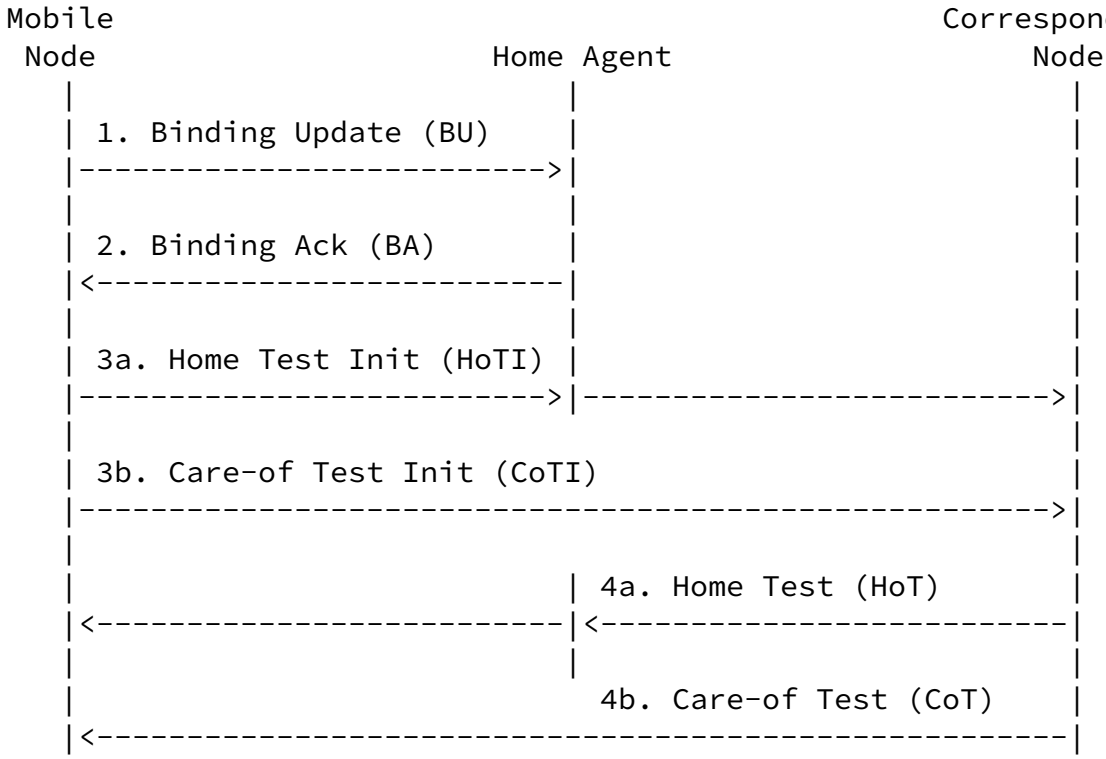
When a mobile node begins communicating with a particular correspondent node after a successful home registration, all packets

are initially routed through the mobile node's home agent, and bidirectionally tunneled between the home agent and the mobile node's current attachment point. For increased routing performance, the mobile node should do a correspondent registration as early as possible.

This section explains the standard Mobile IPv6 registration procedure in the case that route optimization is used. The goals and assumptions based on which this registration process was developed are presented thereafter. The section concludes with a security analysis of the Mobile IPv6 registration process.

3.1 Registration Procedure

A mobile node registers its current care-of address with its home agent and correspondent node. As a result, the home agent and correspondent node create "bindings" between the mobile node's home address and current care-of address. The following is a nutshell presentation of Mobile IPv6 home and correspondent registrations. Figure 1 illustrates this process. The interested reader is referred to RFC 3775 [29] for the complete specification.



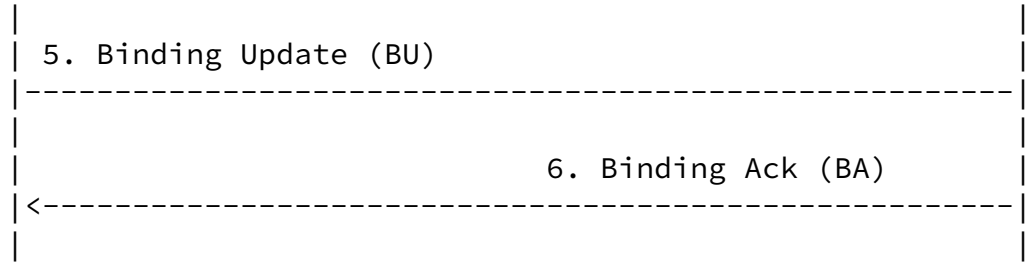


Figure 1: Mobile IPv6 Registration Procedure

When the mobile node detects that it has moved to a different access network, it configures a new care-of address. The mobile node then initiates a home registration by sending to the home agent a Binding Update (BU) message. The BU contains the mobile node's home address, current care-of address, and some supplementary information. If the home registration succeeds, the home agents returns a Binding Acknowledgement (BA) message, informing the mobile node that its home address is now bound to the new care-of address. The BA also specifies for how long the binding will stay in place.

RFC 3775 requires that the BU and BA be authenticated, and recommends that they also be encrypted, through IPsec. The mobile node and home agent may be pre-configured with the necessary security associations, or they may dynamically create them through IKE. In the latter case, the nodes have to be preconfigured with an identifier and the credentials necessary to prove their identity during the IKE authentication stage. This could be a preconfigured shared secret or

a public/private-key pair combined with a certificate that binds the public key to the identifier. Finally, the home agent needs sufficient information to authorize a mobile node to use a particular home address.

The correspondent registration consists of a return-routability procedure followed by the registration proper. The return-routability procedure, in turn, is a combination of two message exchanges, one exchange that goes through the home network and another direct exchange. The return-routability procedure aims to determine whether the mobile node is reachable at its home address and care-of address. The mobile node may initiate the return-routability procedure at any time after it has sent the BU to the home agent. It then sends a Home Test Init (HoTI) message and a Care-of Test Init (CoTI) message to the correspondent node. The HoTI is tunneled to the home agent, which forwards the message to the

correspondent node. The CoTI is sent to the correspondent node on the direct path.

When the correspondent node receives the HoTI, it generates a Home Keygen Token, which it returns to the mobile node's home address in a Home Test (HoT) message. The mobile node needs the Home Keygen Token to show that it is the legitimate owner of its home address. Similarly, when the correspondent node receives a CoTI, it sends a Care-of Test (CoT) message with a Care-of Keygen Token to the mobile node's care-of address. The mobile node needs the Care-of Keygen Token to prove its reachability at the new care-of address. The tokens are produced based on unpredictable nonces, the mobile node's home and care-of address, respectively, and some auxiliary data. Sufficient information is communicated as part of the registration protocol such that the correspondent node will eventually be able to recompute both tokens without having to explicitly store either of them. Note that, although some Mobile IPv6 implementations do the two message exchanges in sequence, a standard-conform and more efficient way is doing them in parallel. Home and Care-of Keygen Tokens are good for 3.5 minutes, so if the mobile node changes its care-of address again during this period, it may reuse its Home Keygen Token. The HoTI-HoT and CoTI-CoT exchanges are respectively called "home-address test" and "care-of-address test".

RFC 3775 recommends that the HoTI and HoT be authenticated, and optionally encrypted, through an IPsec tunnel between the mobile node and the home agent. It is the home agent's responsibility to update the corresponding security association to the new tunnel end point during the home registration.

Once the mobile node has received the HoT and CoT from the correspondent node as well as the BU from the home agent, it can send

a BU to the correspondent node, requesting the correspondent node to bind its home address to its current care-of address. The mobile node must compute a message-authentication code keyed with the Home and Care-of Keygen Tokens. When the correspondent node receives the BU, it can thus decide that the mobile node, first, owns the home address mentioned in the BU and, second, is reachable at the care-of address to which that home address is to be bound. The mobile node may optionally request the correspondent node to return a BA for confirmation by setting a flag in the BU. Note that the BA is mandatory for the home registration.

Bindings at correspondent nodes have a maximum lifetime of seven minutes. If a binding is not updated within this time, the mobile node must re-do the correspondent registration. This includes another run of the return-routability procedure.

3.2 Goals and Assumptions

An important objective for the development of Mobile IPv6 was to provide for a wide, preferably universal, support for route optimization. In fact, support for Mobile IPv6 and, thus, route optimization is recommended in the requirements suite for IPv6 nodes [11]. It was, and still is, believed that the additional routing overhead associated with bidirectional tunneling is too much of a burden on the core Internet given that the number of connected mobile nodes is expected to grow substantially within the next decades.

The aspiration for wide-scale deployment of route optimization has an impact on how a correspondent node can authorize a mobile node to use a particular home address. A mobile node must authenticate itself, preferably without any pre-configured keys, as the legitimate owner of the home address packets addressed to which it seeks to redirect to a certain care-of address. Without such authentication, any node--not necessarily a mobile one--could redirect any other node's packets. The challenge here is to bind to a given home address a property that only the mobile node owning that home address can have.

The return-routability procedure was elected as the default authentication mechanism for Mobile IPv6 route optimization. It verifies home-address ownership through a routing property and so does without any pre-configured authentication material. When a mobile node shows that it can receive messages sent to its home address, this is understood as reasonable evidence that the mobile node is the legitimate home-address owner. Strictly speaking, the return-routability procedure checks only that the mobile node is somewhere on the path between the correspondent node and the home address. An on-path attacker could thus hijack a communications connection that is not protected otherwise. However, the problem

with on-path attackers is independent of mobility and already existed before the introduction of Mobile IPv6. Hence, the return-routability procedure does not create a new security threat.

Of course, there are alternatives to the return-routability procedure. Preconfiguring shared secrets into mobile nodes and

correspondent nodes is one, leveraging public-key cryptography is another. These approaches may in fact do a very good job in certain scenarios. However, both of them have deficiencies as far as general applicability goes. The following explains why neither approach was taken up as the default authentication mechanism in Mobile IPv6.

If a shared secret is pre-configured into a mobile node and its correspondent node, the correspondent node can authenticate the mobile node by having it encrypt a piece of random data and comparing the result with the expected ciphertext. This process is simple and appealing. The crux is that there is usually no existing relationship between an arbitrary pair of nodes before the nodes start communicating. Preconfiguration may hence not be feasible in many cases. And where preconfiguration does apply will it involve considerable administrative overhead, which makes the approach impractical except for some very limited scenarios. Also note that a security association alone does not show that a node owns a specific IP address. This property, however, is required for Mobile IPv6 route optimization, so an external mechanism is needed to authorize an authenticated mobile node to use a specific home address.

Public-key cryptography requires some external binding between a public key and the identity this public key is supposed to protect. Certificates issued by a trusted authority can usually do this job, although there is little experience with using home addresses as identifiers in the certificates. (E.g., the home address could be placed into a certificate's Subject AltName field.) Given a certificate that binds a public key to a home address, the owner of this home address can authenticate itself as such by signing some arbitrary piece of data with its private key. Since everybody can verify the signature with the mobile node's public key, this proves, in the end, that the mobile node actually knows the private key complementing the certified public key, and the certificate authority vouches that the public key, in turn, is associated with the home address. The issue with public-key cryptography in the case of Mobile IPv6 is the extraordinarily high number of potential home addresses. Many experts doubt that one could built a public-key infrastructure (PKI) of the size appropriate for Mobile IPv6. Furthermore, making certificates bind home addresses to public keys is per se an issue, because IP address assignment is typically handled by other network entities than the PKI nodes. A global PKI would also constitute an attractive target for attacks, endangering

the Internet as a whole.

It is important to recognize that some mobility-related attacks can be prevented through authentication and authorization to use a particular home address, while others cannot. A dominant threat uncovered is resource-exhaustion attacks. In fact, the stronger the authentication algorithms, the easier it is to exploit the resources of a node running these algorithms. In a resource-exhaustion attack, an attacker brings down its victim through massively sending it bogus requests. Protection against malicious resource exhaustion was another key driver in the Mobile IPv6 security-design process. The intent was primarily to safeguard those hosts which offer a popular or critical service without necessarily having to be mobile themselves. A Mobile IPv6 correspondent registration is robust against resource-exhaustion attacks in that it is of low complexity and delays state creation as well as computational tasks at the correspondent node until the mobile node has shown its credentials.

Redirection-based flooding attacks are another threat that cannot be encountered by authentication. Mandatory authentication may lessen the attractiveness of such flooding, but certainly cannot prevent it. There must hence be a different mechanism that prevents malicious use of care-of addresses. In Mobile IPv6, a correspondent node probes a mobile node's new care-of address before it sends packet there. This verification strategy operates end to end, and it is as such independent of any support from the network.

An alternative that does require network support is to enforce proper use of care-of addresses already at the mobile node's point of network attachment. The correspondent node may then simply believe in the validity of a care-of address without doing any verification itself. Many access networks today provide this service through ingress filtering [34]. However, the crux with verifying a care-of address at the fringe of the Internet is that an attacker can choose the location from where to wage a flooding attack. As long as there are access networks where ingress filtering, or an equivalent technique, is not deployed, an attacker can always avoid care-of-address verification. Designers therefore made Mobile IPv6 not to rely on ingress filtering. Certificate-based authorization of care-of addresses is also infeasible because care-of addresses change in a typically unpredictable way, whereas certificates are static.

It should be mentioned that care-of-address verification can be omitted in scenarios where the mobile node is considered trustworthy. For instance, RFC 3775 is based on the assumption that there is a trust relationship between mobile nodes and their respective home agents. The care-of-address test is hence omitted during home registrations. This is certainly a feasible hypothesis in many

cases, but one ought to bear in mind that, in some scenarios, it may be not. As an example, a mobile services provider may not be able to trust all individuals from its large customer basis, so it may probe a mobile node's care-of address even during a home registration irrespective of what RFC 3775 defines.

3.3 Security Analysis

To analyze the security of the return-routability procedure, one should evaluate its protection against the tree types of attacks described in section Section 2: impersonation attacks against third parties, resource-exhaustion attacks against mobile nodes or correspondent nodes, and flooding attacks against third parties.

In the context of Mobile IPv6, impersonation is an attack in which the perpetrator claims ownership over a victim's IP address, pretending that this IP address be its own Mobile IPv6 home address. The return-routability procedure can prevent such attacks unless the attacker is on the path from the correspondent node to the victim (in the case of a stationary victim) or from the correspondent node to the victim's home agent (if the victim is mobile). However, if an attacker happens to be on the critical path, it can spoof a HoTI on behalf of the victim, eavesdrop on the returning HoT, and thus illegitimately acquire a Home Keygen Token. The impersonator can produce its own Care-of Keygen Token by sending the victim's correspondent peer a tailored CoTI with a care-of address through which the impersonator is itself reachable. Having both tokens allows the attacker to send an authenticated BU on behalf of the victim.

The return-routability procedure's susceptibility to attacks from the routing path conforms with the objective to prevent new attack types that did not exist before the introduction of Mobile IPv6, but to disregard existing threats that are independent of whether mobility is supported or not. For instance, redirecting someone else's packets from outside those packets' routing path is generally impossible with plain IP, but a "man in the middle" may well launch a successful attack from a position on the routing path. This said, it stands to reason why the return-routability procedure prevents off-path attacks, but does little to stop on-path attacks.

Similarly, the return-routability procedure does not prevent an attacker from registering a care-of address which is located such that the attacker is on the path between the care-of address and the correspondent node. This attacker is in a position to launch a redirection-based flooding attack against the node using the target care-of address (or the entire network this care-of address belongs to). But here again could the attacker launch a comparable attack

already without the help of Mobile IPv6, simply by setting up an upper-layer connection with the victim's IP address. For instance, an on-path attacker could perform a TCP handshake on behalf of its victim, initiating, say, a large file download from an FTP server. With the TCP sequence numbers at hand, the attacker could also send acknowledgements on behalf of its victim to keep the data flow going or even accelerate it.

However, reducing the return-routability procedure's vulnerability to the routing path is insufficient to prevent a related style of attack that is called "space- and time-shift attacks". In these attacks, the perpetrator taps the critical wire in order to eavesdrop on or manipulate return-routability messages, and it then moves to a saver place and starts an impersonation attack from there. The attacker may also wait for a better point of time. It may even install a binding on behalf of a victim before the victim starts communicating. The mandatory, periodic registration refreshes defined by RFC 3775 mitigate the threat of space- and time-shift attacks.

The return-routability procedure is such that the correspondent node does not need to explicitly store the Home or Care-of Keygen Tokens sent to a mobile node. The information communicated in the protocol is sufficient for the correspondent node to re-calculate the token. This saves the correspondent node from attacks against its memory. On the other hand, it may open the door for attacks against the correspondent node's processing capacity. A token is a SHA-1 hash on the mobile node's and correspondent node's IP addresses, a random nonce, and a secret known only to the correspondent node. The computational overhead required to do the hash is rather moderate, although a correspondent node should implement its own policies to manage resources in a situation of increased processing workload.

4. Objectives for Enhancement

This section identifies areas in which route optimization, as specified in RFC 3775, may be incompatible with the requirements of certain applications. Objectives to enhance route optimization usually boil down to optimizations to the return-routability procedure, or alternative mechanisms that may replace the return-routability procedure. The enhancement objectives are herein discussed from a requirements perspective, such as the need for decreasing latency. The technical means to reach those objectives is not considered, nor is the feasibility of achieving them.

4.1 Latency Optimizations

A disadvantage of route optimization is that a mobile node must run a return-routability procedure before it can inform the correspondent

node about its new care-of address. Therefore, a correspondent registration consumes, at a minimum, one and a half round-trip times until the correspondent node receives the BU, assuming that the mobile node performs the home-address and care-of-address tests in parallel. An additional one-way time is needed until the first packet from the correspondent node, and possibly an optional BA, arrives at the new care-of address. Note that the CoTI, CoT, BU are transmitted on the direct path between the mobile node and the correspondent node, whereas the HoTI and HoT go through the home agent. The actual latency of the return-routability procedure is governed by the path with a longer round-trip time.

Direct communications to the correspondent node can optimistically start right after the Binding Update message has been sent (i.e., after one round-trip time), but more generally are delayed until the Binding Acknowledgement message is received (i.e., after two round-trip times).

Similarly, optimistic mobile nodes are allowed by RFC 3775 to start their return-routability procedure right after sending a Binding Update message to their home agent. They can so reduce the latency for the correspondent registration. But more generally, mobile nodes wait for the home registration to be completed and acknowledged before initiating the correspondent registration.

Depending on the type of application, the above delays can be an issue. Interactive, real-time applications, like voice over IP, are an example where the delays may cause perceptible quality degradations. Even fast bulk-data transfer can be affected if the lack of packets during the movement period is interpreted as congestion and leads to a new TCP slow start. There appears to be general consensus that faster mechanisms for route optimization are needed.

Note that the handover delay from an application's perspective is not just the latency of the IP mobility mechanism, but also includes delays at the IP layer and the link layer. In fact, the delays introduced by the rest of the stack can be significant as well

Section 6.3.1.

4.2 Signaling Optimizations

As mentioned in section Section 3, correspondent registrations have a maximum lifetime of seven minutes and must be refreshed in case they are not updated to a different care-of address in the meantime. The reason for this is to reasonably reduce the window of vulnerability to time- and space-shift attacks, where an attacker eavesdrops on unencrypted authentication material exchanged during the

return-routability procedure and launches an impersonation attack at a later time and from a different, probably more amenable location. Periodic re-registrations limit the likelihood and feasibility of such off-path attacks, since the attacker would have to get back on path whenever the authentication material is due to be refreshed.

A calculation in [2] shows that the seven-minute refreshment interval implies a signaling overhead of 7.16 bps when a mobile node communicates with a stationary node. The overhead doubles if both peers are mobile. On one hand, this signaling overhead is certainly negligible when the nodes actually communicate. On the other hand, it may cause problems for mobile nodes that are inactive and stay at the same location for a while, but still want to have route optimization ready with some correspondent node. These nodes typically prefer to go to standby mode to conserve battery power.

An example where the maintenance of route optimization in the absence of traffic may be useful is some sort of messenger service that mobile nodes can subscribe to. Here, having route optimization in place would allow the correspondent node in charge of sending the messages to instantaneously create an efficient connection. If the mobile node used bidirectional tunneling instead, the first packet that the correspondent node sent would be relayed through the home agent and trigger a correspondent registration upon arrival at the mobile node. Since a messenger service is likely to send a few packets per event only, the belatedly created new correspondent registration would probably remain completely unused.

Also, the accumulated signaling overhead for route-optimization maintenance generated by a large customer base may be an issue from a network provider's point of view. Not only do the signaling packets have to be routed through the network, part of them must also be processed by home agents. For example, of the 716 Mbps signaling

overhead generated by 100 million route-optimized mobile nodes, 220 Mbps goes through a home agent.

This discussion shows that there are scenarios in which an optimization for reduced signaling would be beneficial. These scenarios are important enough to have an impact on the deployment of Mobile IPv6.

4.3 Security Enhancements

The return-routability procedure is lightway and prevents mobility-related attacks reasonably well. In some cases, however, may better security be useful. One may in particular attempt to limit what on-path attackers can do. Attackers that operate in the same networks as one of the communication end points are also a

threat that one may want to avoid. There are existing proposals that offer higher security in Mobile IPv6 [24] and in other mobility-management protocols such as HIP [22].

However, even with better security for mobility management can the Internet as a whole not become any safer than the non-mobile Internet. For instance, on-path attackers can cause denial of service, or inspect and modify cleartext packets, already without misusing a mobility-management protocol. Applications that require strong security are therefore generally advised to end-to-end mechanisms such as IPsec or TLS. But even communications that are protected on an end-to-end basis are vulnerable to denial of service.

Better route-optimization security may become necessary in the future, if technological improvements remove some of the existing mobility-unrelated vulnerabilities of the Internet. For instance, the use of Secure Neighbor Discovery [20] in a network where one of the communication end points resides can remove some of the existing threats.

4.4 Applicability Enhancements

As per RFC 3775, a mobile node's home address and current care-of address are carried in all route-optimized packets. The course of the mobile node is therefore trackable, both by the correspondent node as well as by a third party. This can be an issue in situations where the mobile node prefers not to reveal its location. Location privacy, however, is inherently not supported by Mobile IPv6 route

optimization. A workaround is to fall back to bidirectional tunneling when location privacy becomes an issue. Packets that carry the mobile node's care-of address are then only transferred between the mobile node and the home agent, and they can be encrypted through IPsec ESP [27][10] on that path. However, the mobile node may have to periodically re-establish its IPsec security associations so that it cannot be tracked through its SPIs.

Scenarios where location privacy is desired are one example where Mobile IPv6 proves insufficient. Early improvement efforts have already started in this area [8][4]. There may also be other deployment scenarios where the applicability of Mobile IPv6 is limited and could be extended.

5. Enhancements Toolbox

This section introduces a number of techniques, a "toolbox", that can be used in the construction of an efficient and secure route-optimization protocol. The section starts with the standard mechanisms used in RFC 3775 and continues with additional techniques

that have been proposed as enhancements or alternatives.

It is important to mention that many enhancements techniques are insufficient or insecure when applied on their own, because the scope of each of them is usually limited to a certain sub-issue. It is the combination of a set of techniques that makes an efficient and secure route-optimization mechanism possible. Different techniques also have different trade-offs with respect to, say, universal applicability versus efficiency.

5.1 IP-Address Tests

An IP-address test can be employed to ensure that the peer is live and on the path to a specific destination address. RFC 3775 uses IP-address tests for two purposes: The home-address test provides evidence that a mobile node owns the home address it wants to use; the care-of-address test serves to preventing flooding attacks related to spoofed care-of addresses. As specified in RFC 3775, IP-address tests can be stateless for the correspondent node, making their use in denial-of-service attacks harder.

IP-ddress tests are a zero-configuration approach that is independent of ancillary infrastructure. The subsequent disadvantage is that

IP-address tests can only guarantee that a peer is on the path to the probed IP address, not that the peer truly owns this IP address. On the other hand, the types of attacks that an on-path attacker can do with route optimization are the same that an on-path attacker can anyway do without route optimization, so there is actually no significant new threat.

The use of two IP-address tests requires four messages. Both tests can be performed in parallel, so they can be completed in one round-trip time.

5.2 Protected Tunnels

An additional technique used in RFC 3775 is the protection of a part of the signaling communications through an authorized and, optionally, encrypted tunnel between a mobile node and its home agent. This prevents other nodes, close to the mobile node, from seeing a home-address test.

Given the starting point that we cannot assume a pre-existing end-to-end security relationship between the mobile node and the correspondent node, this protection exists only for the mobile node's side. In case the correspondent node is stationary, the path between the home agent and the correspondent node remains unprotected. An attacker on that path can still perform attacks, but these attacks

are similar to those that an on-path attacker can anyway do without route optimization. So, as with IP-address tests, the intent here is not to introduce any significant new threat to the Internet. The same is true in case the correspondent node is mobile. It then has its own home agent, and it is the path between the two home agents that stays unprotected.

5.3 Optimistic Behavior

RFC 3775 leaves quite a bit of freedom for a mobile node with respect to scheduling signaling and data packets. An optimistic mobile node can initiate the return-routability procedure right after sending the BU to its home agent, even before it has gotten a BA back.

The mobile node must wait for the home agent's BA before it can send a BU to the correspondent node. However, the mobile node may start sending data packets to the correspondent node right after it has sent this BU without having to wait for a BA.

The drawback of the described optimistic behavior is that a dropped, re-ordered, or rejected BU can cause data packets to be dropped. Such packet loss would also have an effect on pessimistic signaling, however. As a result, further experimentation and simulation may be needed to quantify to the effects of optimistic techniques under different conditions.

5.4 Proactive IP-Address Tests

The post-movement time period during which a mobile node and correspondent node cannot fully communicate is oftentimes called the "critical phase". Usually, the critical phase spans a home registration and a correspondent registration including a return-routability procedure. One technique to shorten the critical phase is to move some of these tasks to an earlier stage. In particular, the home-address test can be done proactively before a handover, instead of doing it afterwards, without violating the base specification. This is discussed in [25].

A Home Keygen Token is generally valid for 3.5 minutes. Consequently, the mobile node must initiate a proactive home-address test at least every 3.5 minutes if it seeks to have available a fresh Home Keygen Token at all times. This is especially helpful if the mobile node cannot foresee the next handover. Alternatively, the mobile node may be able to receive a trigger from its local link layer indicating that a handover is imminent. In this case, the mobile node may initiate the home-address test right in time instead of doing it periodically every 3.5 minutes. Note, however, that the mobile node must anyway re-initiate the correspondent

registration--and, thus, the home-address test--after the maximum binding lifetime of seven minutes. Link-layer triggers can therefore save the mobile node at most every second home-address test (unless they are combined with additional techniques such as [2]).

Another optimization possibility is performing a care-of address test before the movement. This is possible only if the mobile node is capable of attaching to two networks simultaneously.

5.5 Concurrent IP-Address Tests

If one assumes that a mobile node can attach to only a single network at a time, executing the care-of-address test proactively before a

handover is not an option. However, one may authorize a mobile node to start using a new care-of address right after the handover, without doing the care-of-address test first, with the restriction that a care-of-address test be initiated rightaway. The peers could then already exchange packets through the new care-of address while the test is being executed. In recent literature, one refers to the care-of address as "unconfirmed" when the correspondent node does not yet know the result of the concurrent care-of-address test, and one calls it "confirmed" thereafter. The lifetime of the associated binding can be limited to a few seconds as long as the care-of address is unconfirmed, and it can be extended once it becomes confirmed.

It is important to understand that concurrency is legitimate only for care-of-address tests. In contrast, home-address tests are done for mobile-node authentication, which must be done before any signaling messages are accepted. Authentication guarantees that only the legitimate mobile node can create or update a binding pertaining to its home address. However, both IP-address tests are in general simultaneously performed during the critical handover period, and one can expect the home-address test to have a longer latency than the care-of-address test. The full benefit of eliminating the care-of-address tests from the critical handover period by means of concurrency can therefore only unfold if some other mechanism is used to move the home-address tests out of the critical handover period as well. For instance, one can do the home-address tests proactively before a handover as suggested in Section 5.4, or one may use cryptographically generated home addresses as proposed further down in Section 5.9.

Concurrent care-of-address tests were first proposed in [25] where they were combined with proactive home-address tests. In [25], as soon as the mobile node has configured a new care-of address after a handover, it sends to the correspondent node an Early Binding Update (EBU) message. The mobile node signs the EBU with a

message-authentication code keyed only with the Home Keygen Token that the mobile node has previously retrieved through a proactive home-address test. Upon reception of the EBU, the correspondent node creates a tentative binding for the new care-of address, which can then be used while the care-of-address test is being executed. When the care-of-address is done, the mobile node sends a standard BU to the correspondent node, concluding the registration procedure.

From the reception of an EBU to the reception of the corresponding standard BU, the correspondent node cannot be sure whether the mobile node is actually present at the claimed new care-of address. A malicious node may misuse this property to redirect packets to a third party's IP address during this phase of uncertainty. Under many circumstances, this will not be acceptable even if the lifetime for an unconfirmed care-of address is tentative only, and there needs to be external protection. Techniques like those described in Section 5.7 or Section 5.8 can help.

5.6 Diverted Routing

Given that the per-movement signaling takes some time, a mobile node can optionally request its traffic to be routed through its home address while this signaling is being completed. The performance impact of this technique depends on the length of the critical phase as well as on the capacity and latency of the direct path and the path through the home agent. With respect to the packets that the correspondent node sends, the following analysis can be made.

The correspondent node does not know that the mobile node has moved until it has been told about this. It continues to send packets to the mobile node's old care-of address until that time. These packets are usually lost and must be transmitted by upper-layer mechanisms. In addition, even the request to delete or deactivate a binding requires some security-related signaling to prevent misuse by unauthorized nodes. Zero packet loss can generally only be achieved through local repair techniques in the mobile node's access network (cf Section 5.14), or if the mobile node can simultaneously attach to two IP networks.

Once the correspondent node knows that the old care-of address is stale, it can send further packets to the home address. If one assumes that the correspondent registration for the new care-of address involves messages through the home agent, it is obvious that at least some of these packets reach the mobile node before the new binding is set up. After all, signaling and data packets travel the same path.

It depends on the capacity and latency of the path through the home

agent relative to the latency of the direct path for how long the correspondent node should continue to send packets to the home address. If the former path has a high latency, it might be better

to queue some of the packets until the correspondent registration is complete and packets can be directly sent to the mobile node. One potential research direction is to look at whether the properties of the paths could be learned during the signaling and then used to decide the optimal time when the correspondent node should start queueing packets.

The situation for the packets that the mobile node sends is similar: Although the mobile node knows immediately that it has moved, RFC 3775 does not allow the mobile node to route-optimize packets from new care-of address until it has formally updated the correspondent node about the new care-of address. Of course, the mobile node may buffer packets until the correspondent registration is done so that no packets get lost.

Diverted routing appeared originally in [25] and has since been used also in [6].

5.7 Credit-Based Authorization

As described in Section 5.5, a new care-of address may already be used while the care-of-address tests is in progress. The prerequesite is that sufficient protection is provided against redirection-based third-party flooding. One way of doing this is authorizing a mobile node to receive packets at a new, unconfirmed care-of address based on credit that the mobile node has collected with a previous care-of address. (See Section 5.5 for a definition on when a care-of address is confirmed and when it is unconfirmed.) This mechanism has become known as Credit-Based Authorization (CBA) [26].

CBA limits the data volume and rate that the correspondent node sends during a concurrent care-of-address test such that it does not exceed the data volume and rate that the mobile node has sent in the recent past. The intention here is not so much to prevent redirection-based flooding attacks altogether as to render impossible any kind of amplification that can be achieved through redirection. It is this inherent potential for amplification which constitutes the attraction to redirection-based flooding: While the attacker simply does an initial connection setup and a subsequent correspondent registration for packet redirection, it is the correspondent node which generates the packets (typically full-sized TCP segments) that the victim ends up having to receive. Transport-layer acknowledgements can generally be faked, so the attacker can easily keep the redirected data stream alive. In the end, the correspondent node spends, unknowingly, much

more resources on the flooding attack than the attacker itself. CBA renders such amplification impossible. This makes redirection-based flooding attacks very unattractive to the attacker because it would take the attacker less coordinative effort, and be at least equally effective, if it sent bogus packets to the victim directly.

Technically, CBA works as follows. A CBA-enabled correspondent node maintains a credit account for each mobile node it communicates with. The correspondent node increases the mobile node's credit by the size of each inbound packet received from the mobile node. When the correspondent node sends a packet to the mobile node, and a concurrent care-of-address test is in progress, the IP address to which the packet is sent depends on how much credit is left. If the credit is higher than the size of the outbound packet, that packet is directly sent to the mobile node's care-of address. However, in case the remaining credit is too small, the packet is sent to the mobile node's home address. Since the home agent has a trust relationship with the mobile node, it can forward these packets to the mobile node's care-of address without having to do a reachability check first. Exponential aging limits the lifetime of collected credit. This guarantees that the mobile node cannot collect credit over an extended time period at a very slow speed and use this credit, all at once, for a short but potent data burst towards a faked care-of address.

Allocating a mobile node's credit based on the packets that the mobile node sends and reducing the credit based on packets that the mobile node receives is defined as CBA mode 1. With applications that send comparable data volumes into both directions, CBA mode 1 works fine. On the other hand, CBA mode 1 may prevent the mobile node from collecting the amount of credit it needs for a handover when applications with asymmetric traffic patterns are in use. For instance, file transfers and media streaming are characterized by high throughput towards the client, typically the mobile node, and comparably little throughput towards the serving correspondent node. To better accommodate such applications, a second CBA mode was designed.

With CBA mode 2, credit allocation is based on packets that the mobile node receives from the correspondent node rather than on packets that the mobile node sends. New credit is allocated while the mobile node's current care-of address is confirmed; existing credit is used up while the care-of address is unconfirmed. Thus, its is the data flow from the correspondent node to the mobile node that is responsible for both credit allocation and reduction, resolving the issue with applications producing asymmetric traffic patterns.

It is less obvious why CBA mode 2 outrules flooding-attack amplification than it is for CBA mode 1. The key observation is that a mobile node invests comparable effort for packet reception as for packet transmission, in terms of bandwidth, memory, and processing capacity. It is therefore legitimate to give a mobile node credit for packets that it has received and processed. The question is, though, how the correspondent node can determine how many of the packets sent to a mobile node are actually received and processed by that mobile node. As mentioned above, relying on transport-layer acknowledgements is not an option as such messages can easily be faked. CBA mode 2 hence defines its own feedback mechanism, Care-of Address Spot Checks, which is much more robust to spoofing. With Care-of Address Spot Checks, the correspondent node periodically tags packets that it sends to the mobile node with a random, unguessable number, a so-called Spot Check Token. When the mobile node receives a packet with an attached Spot Check Token, it buffers the token until it sends the next packet to the correspondent node. The Spot Check Token is then included in this packet. Upon reception, the correspondent node verifies whether the returned Spot Check Token matches a token recently sent to the mobile node. New credit is allocated in proportion to the ratio between the number of successfully returned Spot Check Tokens and the total number of tokens sent. This implies that new credit is approximately proportional to the fraction of packets have made their way at least up to the mobile node's IPv6 stack. The preciseness of Care-of Address Spot Checks can be traded with overhead through the frequency with which packets are tagged with Spot Check Tokens.

5.8 Heuristic Monitoring

Heuristic approaches to protect concurrent care-of-address tests are conceivable as well. For instance, one may consider a lifetime limit for unconfirmed care-of addresses which, supplemented by a heuristic for misuse detection, can prevent, or at least effectually discourage, misuse of such addresses. The challenge here seems to be a feasible heuristic: On one hand, the heuristic must be sufficiently rigid to catch any malicious intents at the other side. On the other hand, it should not have a negative impact on a fair-minded mobile node's communications.

Another problem with heuristics is that they are usually reactive. The correspondent node can only respond to misbehavior after it appeared. If the response comes quickly, attacks may simply not be worthwhile. But premature actions should be avoided, of course. One

must also bear in mind that an attacker may be able to use different home addresses, and it in general hard for the correspondent node to see that the set of home addresses belongs to the same node. The attacker may also use multiple correspondent nodes for its attack in

an attempt to amplify the result.

5.9 Cryptographically Bound Identifiers

Cryptographically Generated Addresses (CGA) offer a method for binding a public key to an IP address. A CGA is an IPv6 address with a standard routing prefix and an interface identifier generated from a hash on the CGA owner's public key and some additional parameters. A CGA allows the owner to assert a claim on its address: It can sign a to-be-authenticated message with its private key and attach its public key along the parameters necessary to recompute the CGA. The recipient of this message can then verify whether the address is correct.

CGAs offer three main advantages: First, spoofing attacks against a CGA are much harder than attacks against a normal IP address. Though an attacker may always create its own CGA, it is unlikely to find a public/private key pair that produces someone else's CGA. Second, CGA-based authentication works entirely end-to-end; it does not depend on a certification infrastructure. Third, CGAs are syntactically just ordinary IPv6 addresses. Their additional semantics do not require any upgrade or modification to the Internet.

Many applications are conceivable where CGAs can be advantageous. In Mobile IPv6, CGAs can bind a mobile node's home or care-of address to its public key. CGAs were originally described in [36] and in [37], and they have later been used in [24], [19], [6], and others.

One limitation of CGAs is that the hash on the CGA owner's public key can only be 62 bits long. The rest of the address is occupied by a 64-bit routing prefix as well as the universal/local and individual/group bits. A brute-force attacker might thus be able to find a public/private key pair that produces a certain CGA. It could then claim ownership over this CGA. The threat can usually be contained by including the address prefix in the hash computation, so that an attacker, in case it did find the right public/private key pair, could not form CGAs for multiple networks from it.

Higher security can be achieved through longer cryptographically

bound identifiers. For instance, a node's primary identifier in the "Host Identity Protocol" (HIP) [22] is a 128-bit hash on the node's public key. It is used as an IP-address replacement at stack layers above IP. On the other hand, a 128-bit identifier is not routable, so there needs to be some external location mechanism if a node wants to contact a peer of which it only knows the identifier.

5.10 Pre-Configuration

The return-routability procedure was designed for communication peers that do not share an a-priori security association. In order to thwart off-path attacks nonetheless, the procedure can establish only very short-living security associations. However, in certain, restricted scenarios, it may be possible to pre-configure mobile and correspondent nodes with security associations. Such security associations can have much longer lifetimes because pre-configuration is inherently more secure than the plaintext token exchange from the return-routability procedure.

Pre-configuration has two major benefits. The first one is strong, cryptographic authentication and encryption, which can be applied to both signaling and data packets. The second advantage is lower signaling delay, because the additional round-trip time otherwise needed for the return-routability procedure can be spared. The obvious disadvantage of pre-configuration is its limited applicability.

It is important to recognize the necessity to unambiguously bind a security association to the home address that it is to protect. With regards to the return-routability procedure, this binding is realized by routing the HoTI and HoT through the home address. In the case of a pre-configured security association, the association must be related to the home address as part of the configuration. Note that this affects both secret-key and public-key cryptography.

Two proposals for pre-configuration are currently under discussion in the IETF as optional enhancements to RFC 3775. [15] re-uses most from the standard authentication and authorization procedure defined in RFC 3775. The only difference is that mobile nodes are endowed with the information they need to compute Home and Care-of Keygen Tokens themselves rather than having to obtain them through the return-routability procedure. [5] replaces the standard RFC-3775 concepts with IPsec and the Internet Key Exchange (IKE) protocol.

From a technical standpoint, a pre-configured security association can only replace a home-address test, not a care-of-address test. After all, a correspondent node cannot verify, based on the association alone, whether a mobile node is actually present at the announced care-of address. The problem can be circumvented by postulating that the correspondent node has sufficient trust in the mobile node to believe that the care-of address is correct. However, this assumption, which is made in [15], discourages the use of pre-configuration in scenarios where such trust is unavailable. For instance, a mobile-phone operator may be able to configure subscribers with secret keys for authorization to a particular

service, but it may not be able to vouch that all subscribers use this service in a trustworthy manner.

Another way to avoid the problem of care-of-address verification is to rely on access networks to filter out packets with incorrect IP source addresses (cf. Section 5.12). This approach is taken in [5]. However, the problem with local filtering is that it must be deployed everywhere an attacker may possibly access the Internet in order to be fully protective. Otherwise, an attacker can always find a place from where a spoofing attack is possible, endangering IP nodes anywhere. As things stand, the assumption that deployment of filtering techniques be universal is speculative.

Both of the above two assumptions can be eliminated through care-of-address tests, facilitating the use of pre-configuration in spite of lacking trust relationships or the existence of acess networks without local filtering techniques. Care-of-address tests can be made concurrent for higher efficiency.

5.11 Semi-Permanent Security Associations

A middle-way approach in between the return-routatibility procedure's short-term security associations and pre-configured permanent ones is to dynamically set up a semi-permanent security association upon first contact, and to use it to authenticate signaling over a longer period. Subsequent signaling can then be unambiguously bound to the initial contact.

On-demand security associations for IPsec are traditionally established by executing IKE between two peers. This works well when the negotiated keys are securely bound to the entity that they are to

protect. For instance, in HIP, the guarded entity is a 128-bit identifier which can be derived from the owner's public key through a hash function. In Mobile IPv6, however, the to-be-protected entity is a plain IPv6 home address, which is syntactically indistinguishable from other IPv6 addresses. Given that no direct authentication between the peers is generally feasible, there is no way for a mobile node to prove possession of its home address either. This would allow an attacker to do the IKE exchange with an arbitrary victim's IP address, and to discretionarily redirect the victim's traffic from that time on. Although the attacker would have to be on the path between the victim and the correspondent node while running IKE, it could move off the path once the keys have been exchanged. As the victim lacks the keys, it cannot even re-claim its IP address.

As a result, dynamic semi-permanent security associations must be bound to the right home addresses through some additional technique when used in the context of Mobile IPv6. For instance, they can be

combined with an initial, one-time home-address test, or IKE can be run through the home address. Another way is using CGAs as proposed in [6].

No matter how they are secured, dynamic semi-permanent security associations cannot be leveraged to prove the correctness of a care-of address. They hence fail to solve the problem with redirection-based flooding attacks, and should only be applied in conjunction with care-of-address tests. Semi-permanent security associations were first developed in [3] where they were called "Purpose Built Keys" (PBK).

5.12 Infrastructure

Infrastructure can provide assistance by vouching for the authenticity of a home address, the correctness of a care-of address, or the trustworthiness of a mobile node. Infrastructure can take many forms, such as a PKI tailored for route optimization, or an AAA infrastructure enhanced with the required features.

In its basic form, such an infrastructure hands out home addresses and associates a key with each home address. It may also produce certificates that can be used by others to verify the binding between a key and a home address, or it may provide a query interface where this verification is performed within the infrastructure.

Setting up this type of infrastructure has generally been considered impossible for general Internet use. One of the problems is the current separation of IP-address assignment and security infrastructures. However, Certificate-based Binding Updates (CBU) [23] are a useful simplification: A home agent is assigned a certificate that binds the home-network prefix to the home agent's public key. Correspondent nodes can trust the home agent based on the certificate, and the home agent vouches for the trustworthiness of the mobile nodes it serves. The advantage is that, rather than having to issue a certificate per mobile node, only a certificate per home-network prefix is required. This makes the infrastructure problem more tractable. Furthermore, the home agent may help to establish an end-to-end security association between the mobile node and the correspondent node so that subsequent messages can be securely exchanged on the direct path between the communicating peers. This allows for improved signaling delay during later handovers.

The infrastructure can also help to separate the trustworthy mobile nodes from the non-trustworthy ones. Together with an identifier for each mobile node, this could be used to retroactively track down misbehaving nodes.

AAA architectures are another approach to mobile-node authentication. A home AAA server, which may or may not be co-located with the home agent, can then contact a remote AAA server in the correspondent node's network. Note that this moves some of the authentication overhead from the correspondent node to the remote AAA server. An AAA-based approach can also dynamically assign mobile-node requests to different correspondent nodes while keeping secret authenticating material local at a single AAA server.

Verification of care-of addresses may be based on infrastructure in the mobile node's local access network. For instance, as a care-of address normally appears as a BU's IP source address, the infrastructure can verify that the IP source addresses of all packets leaving the network are correct. Ingress filtering [34] provides this feature to the extent that only the prefix of an outbound packet's IP source address is inspected. The pertinence of this check strongly depends on whether it is done directly in the access router or further up the packet's path.

The problem with verifying a care-of address at the fringe of the Internet is that there is no way for a remote correspondent node to

decide whether a packet has really undergone a check or not. And although many access networks today already do ingress filtering, an attacker can always find a network where such a technique is not deployed.

A more secure approach to care-of-address verification is hence to let the correspondent node itself verify a mobile node's care-of address by querying a piece of infrastructure, located in the mobile node's access network, about the mobile node's presence at a particular care-of address. For this, the mobile node would have to be identifiable by means known to both the correspondent node and the infrastructure. If the care-of address is cryptographically generated (cf. Section 5.9) and configured through Secure Neighbor Discovery [20], the mobile node can be securely identified by the care-of address alone. Otherwise, if CGAs are unavailable, an additional PKI or AAA architecture is needed to distribute the required credentials.

5.13 Local Mobility

Mobile IPv6 handles all mobility on an end-to-end basis. However, it may sometimes make sense to handle part of a mobile node's movements entirely within the local access network. This can yield performance improvements in terms of signaling overhead and handover latency.

Hierarchical Mobile IPv6 [18] is an optimization of RFC 3775 for local mobility support. It introduces the concept of a regional

Mobile Anchor Point (MAP) that acts as a local home agent towards visiting mobile nodes and proxies them towards their home agents and correspondent nodes. When a mobile node enters a visited network, it configures an "on-link care-of address", like in RFC 3775, and a wider-scope "regional care-of address" from a MAP's network. The mobile node registers a binding between the two care-of addresses with the MAP, and it uses the regional care-of address for communicating with nodes outside the local network. When the mobile node moves to a different network within the same MAP's realm, it configures a new on-link care-of address, but keeps its regional care-of address. For the outside world it thus seems as if the mobile node was stationary within the MAP's network.

The mobile node may in addition register the regional care-of address with its own home agent and its correspondent nodes. This allows the mobile node to roam between the domains of different MAPs without

breaking ongoing communication connections.

5.14 Local Repair

When a mobile node moves from one IP-attachment point to another, some packets are likely to be still in flight towards the old location. Local repair techniques can be used to forward these packets to the new IP-attachment point. This can be done through a tunnel or a host route between the old and new access router.

Local repair usually implies that packets are buffered at the old or new access network. If the mobile node leaves the old access network without telling its new care-of address, it must signal this information back subsequently. In-flight packets arriving at the old network should in this case be buffered until the mobile node's new location is known. Alternatively, the mobile node may be able to proactively determine its new care-of address before it moves. In-flight packets can then immediately be forwarded to the new location, where they probably have to be buffered for a little while until the mobile node arrives. A protocol that supports both modes is defined in [17].

5.15 Assisted Auto-Configuration

The local network may assist a mobile node in finding a new access router to which it can handover. For instance, in [17], the mobile node can search for a suitable access point and ask its current access router to proxy-advertise the router to which this access point is attached.

Additionally, the local network may support the mobile node in configuring a new care-of address from its old link. In [17], after

the mobile node has determined the access router that it wants to handover to, it may suggest a new care-of address. The candidate care-of address is signaled to the prospective access router which performs DAD on the care-of address and signals the result back to the mobile node. The new access router also performs Proxy Neighbor Discovery in case the new care-of address is available.

Assisted auto-configuration can have enormous performance benefits, especially when combined with local repair techniques. A disadvantage is the investments for setting up the required infrastructure. Also, local support must be provided in both the old

and the new access network, so handovers between different administrative domains may be problematic.

5.16 Processing Improvements

One goal for designing the return-routability procedure was low computational complexity. The processing overhead for route optimization should thus be acceptable in general. However, mechanisms alternative to the return-routability procedure, such as public-key cryptography, may be more taxing on processing resources. Here, it may help to replace RSA algorithms with ECC techniques. Moreover, even the low-complexity cryptographic algorithms used in the return-routability procedure may be too expensive for very busy servers.

5.17 Delegation

Given that the home agent does not need to move or conserve battery energy, it can be used for performing computationally expensive tasks. It can also be used for parts of the signaling in order to reduce communications over slow wireless links. Some work relating to delegation has been done in [24], [23], and [32].

6. Analysis

This section analyzes the techniques presented in Section 5 in relation to each other and in the context of the enhancement objectives described in Section 4. The techniques are categorized first. Some recent proposals for route-optimization enhancement, which rely on one or combine multiple of these techniques, are subsequently evaluated. The section concludes with a number of opportunities for further research in the area of route optimization.

6.1 Categorization of Techniques

Local techniques include support for micro-mobility, route repair, and assisted auto-configuration. They seek to reduce or eliminate

long end-to-end round-trip times or delays caused by standard auto-configuration techniques. Local techniques have traditionally been implemented in a manner that requires configuration, but there appears to be no fundamental reason why this would have to be so.

IP-address tests, which may be proactive or concurrent, credit-based

authorization, heuristic monitoring, CGAs, semi-permanent security associations, and cryptographically bound identifiers are end-to-end techniques. They are independent of local network support and are thus very flexible. They may also be inexpensive to deploy. The cost for these advantages is typically a smaller performance gain compared to local mechanisms due to global, thus potentially long, round-trip times.

Zero-configuration techniques require no prior configuration or assistance from a managed infrastructure. IP-address tests, diverted routing, credit-based authorization, heuristic monitoring, CGAs, semi-permanent security associations, cryptographically bound identifiers, processing improvements, and delegation are zero-configuration techniques. Mechanisms that require pre-configuration or some kind of infrastructure are not among zero-configuration techniques.

6.2 Evaluation of Recent Proposals

6.2.1 Local Assistance

There are currently two proposals in the IETF for local mobility assistance, Hierarchical Mobile IPv6 and Fast Handovers for Mobile IPv6. Hierarchical Mobile IPv6 (HMIPv6) [18] screens a mobile node's movements within a MAP domain from nodes not inside that domain. In case standard Mobile IPv6 is used end-to-end, this eliminates most of the global signaling: While its regional care-of address does not change, a mobile node does not need to update its home agent or correspondent nodes beyond the mandatory periodic refreshments. Not having to signal on a global basis also reduces handover latency.

Updates to the home agent and to correspondent nodes are necessary only when the mobile node leaves the current MAP domain. The mobile node may then register a new regional care-of address with a different MAP if one is available. Note that switching MAPs usually requires the mobile node to signal more than if standard Mobile IPv6 was used alone. Local and end-to-end signaling then comes together because, as it stands, a mobile node must contact the new MAP separately from its home agent and correspondent nodes. Due to dependencies between a MAP registration and a contemporary home or correspondent registration, a mobile node may want to wait for the MAP registration to complete before it initiates the standard Mobile

IPv6 registration procedure. Handover latency is then increased in

addition to the signaling overhead. Future work could thus go into the integration of MAP registrations with standard Mobile IPv6 signaling.

Another interesting research opportunity seems to be a mechanism that tells neighboring MAPs from different administrative sites that their domains overlap. The MAPs could then mutually advertise each other throughout certain parts of their domains.

The cost for an inter-MAP handover in terms of signaling load and delay strongly depends on the network topology. In an optimal layout, a MAP is located somewhere on the path from the mobile node to its home agent and correspondent nodes. This may be the case if the MAP is co-located with an Internet gateway. Then, switching MAPs is cheap. On the other hand, if the MAP is way off the direct path between a mobile node and its peers, the additional overhead might become noticeable. Indeed, a good topological layout is crucial for the performance of HMIPv6.

The second local optimization, Fast Handovers for Mobile IPv6 (FMIPv6) [17], streamlines the router-discovery and IPv6-address-configuration processes, and it facilitates lossless handovers. FMIPv6 assumes that access routers are capable of matching a neighboring access point to the access router to which it attaches. The capability is a prerequisite for proxy router discovery. Yet, it is still an open issue how access routers learn about this information. Manual configuration is one option, though it can be extremely expensive. More attractive would be an automated mechanism that allows access routers to dynamically recognize access points to which mobile nodes may want to switch. A related issue is how such knowledge can be securely obtained across the borders of administrative sites. These are opportunities for future research.

Note that Hierarchical Mobile IPv6 is applicable even when Mobile IPv6 is not used beyond the local domain. I.e., a mobile node may use its regional care-of address as a temporary home address. The mobile node would thus appear to a correspondent node as a stationary node in the MAP's network. This allows the mobile node to keep its movements private as long as it stays within the same MAP domain. FMIPv6 can be used in combination with standard Mobile IPv6, HMIPv6, or both, but it cannot be used without either.

Of course, there are disadvantages with HMIPv6 and FMIPv6. Local optimizations in general require investments in the access network and thus imply additional costs for the network operator. Also, as mentioned, localized mobility support may even cause increased overhead in certain situations. And local mechanisms are to date

ineffective for handovers across administrative domains unless providers have mutual agreements to interconnect their networks.

6.2.2 Pre-Configuration

The Home Keygen Token exchange from the return-routability procedure is the default authentication technique used in Mobile IPv6. It facilitates reasonable security even between nodes that have no pre-existing relationship. On the other hand, nodes that do share a common secret should be allowed to omit the home-address test. This can be beneficial in certain scenarios where the home-address test causes a long handover latency due to packet redirection through the home agent. Note that a shared secret cannot replace the care-of-address test. Eliminating the care-of-address test requires some sort of trust into mobile nodes not to spoof a care-of address. The pre-configuration approach standardized in the IETF [15] uses a shared secret between the mobile node and the correspondent node, and it assumes that mobile nodes are trustworthy.

The assumption that mobile nodes be fair-minded turns out to be quite far stretching. On on side, it affords the elimination of the entire return-routability procedure, not just the Home Keygen Token exchange. As explained in [15], and as it can also be inferred from figure Figure 1, this cuts the average handover latency in half. On the other hand, the assumption significantly limits the applicability of the optimization. There are certainly scenarios where pre-configuration per se would be possible, but no trust model can be assumed. For instance, an ISP may configure its media servers with the keys of its customers. The customers could then use pre-configured Mobile IPv6 for communications with the media servers, but some customers might misuse the lack of a care-of-address test to wage a redirection-based flooding attack against a third party. This example reveals the difference between a security association for authentication and a trust relationship for misuse prevention.

In an effort to extend pre-configuration to scenarios where no trust relationship can be presupposed, one may combine it with care-of-address tests. Of course, using a care-of-address test partly vitiates the handover-latency improvement that can be reached otherwise. But there may still be a positive impact on handover latency, because pre-configuration eliminates the triangular home-address test through the home agent, whereas the care-of-address test uses the direct, and typically faster, path between the communicating nodes. For increased performance, a concurrent care-of-address test can be used in combination with credit-based authorization or heuristic monitoring. It should also be noted that pre-configuration facilitates stronger authentication mechanisms than the return-routability procedure, and thus the use of route

optimization may become more suitable for applications with high security requirements.

These things said, it seems like a good idea to make the pre-configuration protocol bendable to different environments. Is there a small group of people who trust each other? Then, of course, group members are unlikely to spoof care-of addresses, and the care-of-address test may be omitted. Or is the group of users large and users are primarily unknown to each other like the customer base of a big ISP? Then, proper authentication does usually not imply trust, and it may not be feasible to use pre-configured keys without checking a mobile node's reachability. Traceability techniques are not necessarily a compensation for the missing care-of-address test, because they are a reactive measure, whereas a care-of-address test is a proactive one.

6.2.3 CGA-Based Optimizations

CGAs can guarantee that a mobile node is the legitimate owner of its home address. They provide higher assurance than the pure use of routing paths. This facilitates a significant reduction in the number of signaling messages per correspondent registration as well as the periodicity of these registrations. In addition, the public keys used in the CGA technique allow packets to be transferred privately, a feature that can be used for both data encryption and for other route-optimization enhancements.

CGA may also be used to reduce the risk of flooding attacks via care-of addresses, as attackers should be unable to generate a private/public-key pair of which the public key hashes to a particular victim's IP address. However, only the interface identifier of a CGA is cryptographically generated, so flooding a network or a link is still an issue. As a result, CGAs should be employed together with a care-of-address test in scenarios where redirection-based flooding attacks are a concern. Similarly, an initial home-address test is typically required, too, in order to avoid that the deletion of a binding causes a flood upon a faked home address.

To resolve collisions in case CGAs are used as care-of addresses, a collision count is part of the input to the CGA hash function. Increasing the collision count by one changes the result of the hash function, so new CGAs can be successively tried until an unused one

is found. On the other hand, the collision count also helps an attacker in faking a CGA: It may use the same private/public-key pair to efficiently generate multiple CGAs. For this reason, the collision count should usually be limited to a few bytes only.

6.2.4 Credit-Based Authorization

With CBA, a new care-of address can be probed in parallel with already using it. This eliminates the handover latency that the reachability check entails when performed during the critical handover period. Depending on how much credit a mobile node has at the time of handover, packets are either routed via its home network or directly to its care-of address. The performance benefits depend on the relative delay characteristics of the direct path between the communicating peers and the path through the home agent. They also depend on how fast the transport-layer protocol in use detects and retransmits lost packets.

CBA depends on the mobile node being able to collect an amount of credit high enough to bring it through the next handover. How easy it is for a mobile node to acquire new credit depends on the CBA mode, on the credit-aging factor, and on the application's traffic patterns. (See Section 5.7 for a description of the two CBA modes.) In general, applications with symmetric traffic patterns make it easy for the mobile node to get sufficient credit. CBA mode 1, however, can be problematic when most data is sent from the correspondent node to the mobile node unless the credit-aging factor is very small. The reason is that, in this mode, earned credit is proportional to packets sent to the correspondent node, whereas spent credit is proportional to packets going into the opposite direction. Also, CBA does not work well in either mode if the mobile node moves so rapidly that it does not manage to refill its credit account between two successive handovers.

Note that CBA is an optimization technique which can be integrated into any mobility-management protocol that verifies IP addresses through probing. Protocols that may benefit are, for instance, other Mobile IPv6 optimization techniques described in this paper such as CGA-based ones (cf. Section 5.9). Also, when Mobile IPv6 is used with pre-configured shared keys between mobile nodes and correspondent nodes (cf. Section 6.2.2), but reachability checks are still prescribed to ensure that properly authenticated mobile nodes do not lie about their care-of address, CBA may be applied to eliminate the additional cost that the reachability checks would

otherwise entail. Moreover, in scenarios where a home agent cannot trust in the correctness of the registered mobile nodes' care-of addresses, CBA-based concurrent care-of-address tests could be proscribed even for home registrations. The same is true for Hierarchical Mobile IPv6, which, as it stands, assumes that a MAP can be confident that mobile nodes use correct on-link care-of addresses, and so gets around the care-of-address test.

Furthermore, CBA may also be used to parallelize reachability checks

in the Host Identity Protocol (HIP) [22] and current Mobike proposals. Note, however, that mobility-management protocols in general do not have an equivalent to Mobile IPv6's static home addresses through which mobile nodes stay reachable at all times. Hence, temporarily routing a mobile node's packets through a static IP address in case this mobile node runs out of credit during a handover may not always be an option.

A nice property of CBA mode 1 is that it does not require support from the mobile node. The mobile node neither needs to understand that CBA is effective at the correspondent node, nor does it have to have an idea of how much credit it currently has. A legitimate question is whether this is true even if the correspondent node may temporarily send packets to the home address of a mobile node that has run out of credit. After all, RFC 3775 suggests that mobile nodes interpret the reception of a tunneled packet as a hint to initiate a new correspondent registration, which would obviously be inappropriate in the described situation. The answer is that all correspondent registration are subject to rate limiting, so mobile nodes can be expected not to misunderstand tunneled packets during the handover procedure.

Care-of Address Spot Checks must be responded to by a mobile node. Consequently, CBA mode 2 can only be implemented transparently to the mobile node in scenarios where packet loss is not an issue and Care-of Address Spot Checks can be omitted.

Last, but not least, an interesting question is whether CBA mode 2 could be misused by an attacker that has an asymmetric connection to the Internet. Wide-spread digital subscriber lines (DSL), for instance, typically have a much higher download rate than upload rate. The limited upload rate would render most denial-of-service attempts through direct flooding meaningless. But the strong download rate could be misused to illegitimately build up credit at

one or many correspondent nodes. The credit could then be used for a more potent, redirection-based flooding attack. The reason why this has so far not been considered an issue is that, in order to build up enough credit at the remote end, the attacker would first have to expose itself to the same packet flood that it could then redirect towards the victim. This is obviously true with respect to data volume. Since credit is aged over time, it also applies to the data rate.

6.2.5 Prefix-Based Certificates

The Mobile IPv6 base specification avoids strong authentication cryptography for signaling between mobile nodes and correspondent nodes. One reason for this is the impracticality of a global trusted

PKI that could approve bindings between the mobile nodes' identities and public keys. Another reason is that limited power resources and processing capacity at the mobile nodes generally rule out any complex cryptographic operations. Robustness to resource-exhaustion attacks requires a similar restrictiveness on the correspondent-node side.

CBU circumvents the lack of a global PKI. The reduction in the number of potentially required certificates makes certificate-based approaches to mobile-node authentication more feasible than it is today. The approach also avoids heavy computations at mobile nodes since public-key cryptography is handled by the home agent. On the other hand, the processing overhead at correspondent nodes actually increases compared to standard correspondent registrations. This may not be an issue since resources at stationary correspondent nodes are usually higher than those of many mobile devices. But it may be an issue if the correspondent node is a popular web server or other central resource that cannot afford doing complex cryptographic operations. One should, however, bear in mind that the increased overhand implies a higher risk to resource-exhaustion attacks.

CBU does not solve the issue with care-of-address spoofing: A vouching home agent does not prevent a malicious mobile node from faking its care-of address. The culprit could cheat its home agent, or it could cooperate with it. This said, CBU should be combined with a care-of-address test that rules out redirection-based flooding attacks. A combination of concurrent care-of-address tests and CBA (cf. Section 5.7) can be used to keep the signaling delay during handover as low as it currently is in [23].

There is ongoing work on the integration of AAA with Mobile IPv6 [13]. The current focus is on authentication between mobile nodes and home agents with the intention to replace the IPsec-based authentication protocol for home registrations. But the concept of security proxies proposed in [23] may as well be re-used for enhancements to the AAA infrastructure.

6.3 Future Research

Mobility-related optimizations are currently actively studied by many researchers at different protocol layers. The preceding sections identify ideas and existing proposals for enhancing route optimization. While some of the basic methods are fairly well understood and are being deployed, there are a number of interesting, newer approaches that deserve to be studied in more detail. This section discusses research directions that appear fruitful, or necessary in the future, and that go beyond the existing proposals described so far.

6.3.1 Research at Other Protocol Layers

The efficiency and security related to movements does not depend on Mobile IPv6 route optimization alone, even if researchers often pose their analysis in that light. A movement that is visible at the IP layer involves all lower layers as well. This includes layer 2 attachment procedures; layer 2 security mechanisms such as negotiation, authentication, and key agreement; IPv6 Router and Neighbor Discovery; as well as IPv6 Address Autoconfiguration and Duplicate Address Detection. A complete network attachment typically requires over twenty link- and IP-layer messages, assuming that features necessary for a commercial deployment (such as security) are turned on.

A significant research question is the performance of the network-access stack as a whole. Current protocol stacks have a number of limitations in addition to the long attachment delays [31], such as denial-of-service vulnerabilities, difficulties in trusting a set of access nodes distributed to physically insecure locations, or the inability to retrieve sufficient information for making a handoff decision.

A number attempts are ongoing to improve various parts of the stack, mostly focusing on handover performance. These include link-layer

enhancements, parameter tuning [39], network-access authentication mechanisms [1], fast-handover mechanisms [35], AAA architectures [21], and IP-layer attachment improvements [12]. It is uncertain how far this optimization can be taken by only looking at the different parts individually. An integrated approach may be necessary to gain more significant improvements [32].

It is also unclear at this time which components are the most critical ones. [31] suggests that mobility-related signaling contributes only under 10% of the overall delay in an IEEE 802.11 environment. The most significant delays are caused at the link layer and for IPv6 attachment. However, the results are not conclusive due to the high deviation between the measurements. The results can also be affected by a number of conditions, such as the availability of specific link-layer optimizations, or the type of security mechanism used for Mobile IPv6 home registrations.

6.3.2 Further Route Optimization Research

The primary driver to improve route optimization appears to be better efficiency for a few usage scenarios, such as fast movements or the ability to reduce signaling frequencies for hosts in standby mode. Ongoing work addresses these aspects already quite well, and many of the suggested methods are reasonably stable in this regard. It is

expected that further, perhaps smaller improvements will continue to be achieved through research and parameter tuning far into the future. This is particularly true because the optimizations are often targeted to a specific usage scenario, and may not give the same improvement in other situations. As a result, the publication of a few enhanced methods for different scenarios seems more reasonable than expecting to define a final, all-encompassing method.

The development of an infrastructure-based route-optimization method is clearly a longer-term project. At this stage, it is not even clear that such a mechanism is needed. The Certificate-Based Binding Update Protocol shows some promise in this area, particularly if it can be combined with other mechanisms that use certificates, such as Secure Neighbor Discovery [20]. Pre-configuring keys into end hosts [15] is simple and efficient, but the number of scenarios where it applies is likely to be very limited only.

The following is a list of interesting ideas for new route-optimization research.

- Local mobility or local repair optimizations that require no configuration.

- Care-of-address verification mechanisms that employ lower-layer assistance or Secure Neighbor Discovery.

- The introduction of optimizations developed in the context of Mobile IPv6 to HIP or other mobility protocols, or to link-layer mobility solutions.

- The extension of the developed techniques to full multi-addressing, including also multi-homing.

- Further development of techniques that are based on "asymmetric cost wars" [33], such as CBA.

- Integrated techniques taking into account both link and IP layer mobility tasks.

6.3.3 Experimentation and Simulation

As discussed earlier, the contribution of different stack parts to the overall movement latency is still unclear. The following is a list of areas where measurements and experimentation can yield further, valuable insight.

- Measurements of a realistic network scenario, enabling all features that would likely be needed in a commercial deployment. These features include link-layer access control, for instance. Similarly, it is necessary to consider support for existing enhancement proposals.

- Measurements and simulations of the performance impacts that existing enhancement proposals have on the different parts of the stack.

- Measurements and comparisons of different implementations that are based on the same specification. For instance, it would be valuable to know how much implementations differ with regards to the use of parallelism that RFC 3775 allows in home and correspondent registrations, or with respect to early packet transmission before reception of a BA.

- Measurements of the impact that network conditions such as packet loss can have on existing and new route-optimization mechanisms.

- Statistical data collection on the behavior of mobile nodes in different networks. Route-optimization techniques behave differently depending on what the frequency of movements is, or what traffic streams appear during a mobile node's lifetime.

- Measurements or simulations of the performance that existing route-optimization schemes show under different application scenarios, such as the use of applications with symmetric vs. asymmetric traffic patterns.

7. Security Considerations

Security issues related to route optimization are an integral part of this paper and are as such discussed throughout the paper.

8. Conclusions

Mobility-related optimizations are currently actively studied by many researchers. Some of the basic techniques--such as the return-routability procedure, pre-configured keys, or CGAs--are either already being deployed or can expected to be in the near future. A growing number of new proposals are being studied that attempt to optimize these basic techniques further, or to make them better applicable to a particular scenario.

Many of the current proposals are mature enough to withstand close scrutiny. Their relative advantages are rather subjective, however. For instance, some proposals are very efficient, but have a high cost

in terms of configuration, whereas others do not require configuration, but are slower. It hence appears likely that more than one new method will have to be standardized. Deployment experience is also important, so publication of a few alternative methods as RFCs would be desirable.

It is interesting to see that most if not all current proposals had predecessors that were shown to be insecure. For instance, the initial return-routability procedure as well as the first versions of CGAs were published before the threat of flooding attacks was fully understood. Concurrent care-of-address tests were also first

suggested with insufficient protection against flooding attacks. And several proposals employing semi-permanent security associations have initially suffered from impersonation attacks. This shows the need to reserve a sufficient amount of time for community analysis and review of new methods.

Another interesting observation is that most mature proposals combine a number of techniques and do not rely on any single approach. This is due to the intricate nature of the problem: to build a mechanism that is efficient and at the same time avoids a quite significant number of potential security vulnerabilities.

On the other hand, it is also necessary to avoid overly expensive or complex solutions. For instance, in evaluating the security needs for route optimization, it is important to compare these needs to other vulnerabilities, e.g., denial-of-service attacks, that already exist for on-path attackers in an Internet without mobility support. Of course, a mobility-management protocol should not make these vulnerabilities worse. But since the issues already exist, it is not necessarily a requirement that they be solved by a mobility-management protocol.

A significant research question is the overall performance of the network stack in a mobile setting. This includes mobility management at the IP layer, but is not limited to it. Current network-access protocol stacks have a number of limitations, such as long attachment and movement latencies or significant denial-of-service vulnerabilities. It is uncertain whether further, significant benefits can be achieved if one continues to look at the different parts of the network stack individually. Most likely, a more comprehensive approach is needed. It is also unclear at this time which components of the network stack are the most critical ones to optimize.

Other significant research questions include what effect network conditions such as packet loss can have on current proposals, and to what degree proposals depend on specific application patterns. Our

current understanding about the different traffic patterns and their effects on mobility is limited, and experiments, modelling, and simulations will be needed.

9. References

[1] Institute of Electrical and Electronics Engineers, "Local and Metropolitan Area Networks: Port-Based Network Access Control", IEEE Standard 802.1X, September 2001.

[2] Arkko, J. and C. Vogt, "Credit-Based Authorization for Binding Lifetime Extension", Internet-Draft draft-arkko-mipv6-binding-lifetime-extension-00, May 2004.

[3] Bradner, S., Mankin, A. and J. Schiller, "A Framework for Purpose-Built Keys (PBK)", Internet-Draft draft-bradner-pbk-frame-06, June 2003.

[4] Daley, G., "Location Privacy and Mobile IPv6", Internet-Draft draft-daley-mip6-locpriv-00, January 2004.

[5] Dupont, F. and J. Combes, "Using IPsec between Mobile and Correspondent IPv6 Nodes", Internet-Draft draft-dupont-mipv6-cn-ipsec-01, June 2004.

[6] Haddad, W., Madour, L., Arkko, J. and F. Dupont, "Applying Cryptographically Generated Addresses to Optimize MIPv6

(CGA-OMIPv6)", Internet-Draft draft-haddad-mip6-cga-omipv6-02, June 2004.

[7] Haddad, W. and S. Krishnan, "Optimizing Mobile IPv6 (OMIPv6)", Internet-Draft draft-haddad-mipv6-omipv6-01, February 2004.

[8] Haddad, W., "Privacy for Mobile and Multi-homed Nodes: MoMiPriv Problem Statement", Internet-Draft draft-haddad-momipriv-problem-statement-00, October 2004.

[9] Moskowitz, R., "Host Identity Protocol", Internet-Draft draft-ietf-hip-base-00, June 2004.

[10] Kent, S., "IP Encapsulating Security Payload (ESP)", Internet-Draft draft-ietf-ipsec-esp-v3-08, March 2004.

[11] Loughney, J., "IPv6 Node Requirements", Internet-Draft draft-ietf-ipv6-node-requirements-11, August 2004.

[12] Moore, N., "Optimistic Duplicate Address Detection for IPv6", Internet-Draft draft-ietf-ipv6-optimistic-dad-01, June 2004.

[13] Patel, A., Leung, K., Khalil, M., Akhtar, H. and K. Chowdhury, "Authentication Protocol for Mobile IPv6", Internet-Draft draft-ietf-mip6-auth-protocol-00, July 2004.

[14] Patel, A., "Problem Statement for bootstrapping Mobile IPv6", Internet-Draft draft-ietf-mip6-bootstrap-ps-00, July 2004.

[15] Perkins, C., "Preconfigured Binding Management Keys for Mobile IPv6", Internet-Draft draft-ietf-mip6-precfgKbm-00, April 2004.

[16] Nikander, P., Arkko, J., Aura, T., Montenegro, G. and E. Nordmark, "Mobile IP version 6 Route Optimization Security Design Background", Internet-Draft draft-ietf-mip6-ro-sec-01, July 2004.

[17] Koodli, R., "Fast Handovers for Mobile IPv6", Internet-Draft draft-ietf-mipshop-fast-mipv6-02, July 2004.

[18] Soliman, H., Castelluccia, C., Malki, K. and L. Bellier, "Hierarchical Mobile IPv6 mobility management (HMIPv6)",

Internet-Draft draft-ietf-mipshop-hmipv6-02, June 2004.

[19] Aura, T., "Cryptographically Generated Addresses (CGA)", Internet-Draft draft-ietf-send-cga-06, April 2004.

[20] Arkko, J., Kempf, J., Sommerfeld, B., Zill, B. and P. Nikander, "SEcure Neighbor Discovery (SEND)", Internet-Draft draft-ietf-send-ndopt-06, July 2004.

[21] Arbaugh, W. and B. Aboba, "Experimental Handoff Extension to RADIUS", Internet-Draft draft-irtf-aaaarch-handoff-04, November 2003.

[22] Moskowitz, R., Nikander, P. and P. Jokela, "Host Identity Protocol", Internet-Draft draft-moskowitz-hip-09, February 2004.

[23] Bao, F., "Certificate-based Binding Update Protocol (CBU)", Internet-Draft draft-qiu-mip6-certificated-binding-update-02, August 2004.

[24] Roe, M., Aura, T., O'Shea, G. and J. Arkko, "Authentication of Mobile IPv6 Binding Updates and Acknowledgments", Internet-Draft draft-roe-mobileip-updateauth-02, March 2002.

[25] Vogt, C., Bless, R., Doll, M. and T. Kuefner, "Early Binding Updates for Mobile IPv6", Internet-Draft draft-vogt-mip6-early-binding-updates-00,

February 2004.

[26] Vogt, C., Arkko, J., Bless, R., Doll, M. and T. Kuefner, "Credit-Based Authorization for Mobile IPv6 Early Binding Updates", Internet-Draft draft-vogt-mipv6-credit-based-authorization-00, May 2004.

[27] Kent, S. and R. Atkinson, "IP Encapsulating Security Payload (ESP)", RFC 2406, November 1998.

[28] Perkins, C., "IP Mobility Support for IPv4", RFC 3344, August 2002.

[29] Johnson, D., Perkins, C. and J. Arkko, "Mobility Support in

IPv6", RFC 3775, June 2004.

[30] Arkko, J., Devarapalli, V. and F. Dupont, "Using IPsec to Protect Mobile IPv6 Signaling Between Mobile Nodes and Home Agents", RFC 3776, June 2004.

[31] Alimian, A. and B. Aboba, "Analysis of Roaming Techniques", IEEE Contribution 11-04-0377r1 2004.

[32] Arkko, J., Eronen, P., Nikander, P. and V. Torvinen, "Secure and Efficient Network Access", Extended abstract to be presented in the DIMACS workshop, November 2004.

[33] Arkko, J. and P. Nikander, "Weak Authentication: How to Authenticate Unknown Principals without Trusted Parties", Proceedings of Security Protocols Workshop 2002, Cambridge, UK, April 16-19, 2002.

[34] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.

[35] Mishra, A., Shin, M., Arbaugh, W., Lee, I. and K. Jang, "Proactive Key Distribution to Support Fast and Secure Roaming", IEEE Contribution 11-03-084r1-I, January 2003.

[36] Nikander, P., "Denial-of-Service, Address Ownership, and Early Authentication in the IPv6 World", Proceedings of the Cambridge Security Protocols Workshop, April 2001.

[37] O'Shea, G. and M. Roe, "Child-proof Authentication for MIPv6", Computer Communications Review, April 2001.

[38] Paxson, V., "An Analysis of Using Reflectors for Distributed Denial-of-Service Attacks", Computer Communication Review 31(3)., July 2001.

[39] Velayos, H. and G. Karlsson, "Techniques to Reduce IEEE 802.11b MAC Layer Handover Time", Laboratory for Communication Networks, KTH, Royal Institute of Technology, Stockholm, Sweden, TRITA-IMIT-LCN R 03:02, April 2003.

Authors' Addresses

Jari Arkko
Ericsson Research NomadicLab
FI-02420 Jorvas
Finland

Email: jari.arkko@ericsson.com

Christian Vogt
Institute of Telematics
University of Karlsruhe
P.O. Box 6980
76128 Karlsruhe
Germany

Email: chvogt@tm.uka.de
URI: http://www.tm.uka.de/~chvogt/

Appendix A. Acknowledgements

The authors wish to thank Gabriel Montenegro and Rajeev Koodli for their support, review, and suggestions related to this paper.

Intellectual Property Statement

specification can be obtained from the IETF on-line IPR repository at http://www.ietf.org/ipr.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

Disclaimer of Validity

Copyright Statement

Acknowledgment

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