Internet Engineering Task Force
Internet Draft

IEPREP
H. Schulzrinne
Columbia U.

draft-ietf-ieprep-sip-reqs-02.txt
December 2, 2002
Expires: May 2003

Requirements for Resource Priority Mechanisms for the Session Initiation Protocol

STATUS OF THIS MEMO

Abstract

This document summarizes requirements for prioritizing access to circuit-switched network, end system and proxy resources for emergency preparedness communications using the Session Initiation Protocol (SIP).

1 Introduction

During emergencies, communications resources including telephone circuits, IP bandwidth and gateways between the circuit-switched and IP networks may become congested due to heavy usage, loss of resources caused by the disaster and attack during man-made emergencies, making it difficult for persons charged with emergency assistance, recovery or law enforcement to coordinate their efforts.

As IP networks become part of converged or hybrid networks along with public and private circuit-switched (telephone) networks, it becomes necessary to ensure that these networks can assist during such emergencies.

There are many IP-based services that can assist during emergencies. This memo only covers requirements for real-time communications applications involving SIP, including voice-over-IP, multimedia conferencing and instant messaging/presence.

This document takes no position as to which mode of communication is preferred during an emergency, as such discussion appears to be of little practical value. Based on past experience, real-time communications is likely to be an important component of any overall suite of applications, particularly for coordination of emergency-related efforts.

As we will describe in detail below, such SIP applications involve at least five different resources that may become scarce and congested during emergencies. In order to improve emergency response, it may become necessary to prioritize access to such resources during periods of emergency-induced resource scarcity. We call this "resource prioritization".

This document describes requirements rather than possible existing or new protocol features. Although it is scoped to deal with SIP-based applications, this should not be taken to imply that mechanisms have to be SIP protocol features such as header fields, methods or URI parameters.

The document is organized as follows. In Section 2, we explain core technical terms and acronyms that are used throughout the document. Section 3 describes the five types of resources that may be subject to resource prioritization. Section 4 enumerates four network hybrids that determine which of these resources are relevant. Since the design choices may be constrained by the assumptions placed on the IP network, Section 5 attempts to classify networks into categories according to the restrictions placed on modifications and traffic classes.

Since this is a major source of confusion due to similar names, Section 6 attempts to distinguish emergency call services placed by civilians from the topic of this document.

Request routing is a core component of SIP, covered in Section 7.

Providing resource priority entails complex implementation choices, so that a single priority scheme leads to a set of algorithms that

manage queues, resource consumption and resource usage of existing calls. Even within a single administrative domain, the combination of mechanisms is likely to vary. Since it will also depend on the interaction of different policies, it appears inappropriate to have SIP applications specify the precise mechanisms. Section 8 discusses the call-by-value (specification of mechanisms) and call-by-reference (invoke labeled policy) distinction.

Based on these discussions, Section 9 summarizes some general requirements that try to achieve generality and feature-transparency across hybrid networks.

The most challenging component of resource prioritization is likely to be security (Section 10). Without adequate security mechanisms, resource priority may cause more harm than good, so that the section attempts to enumerate some of the specific threats present when resource prioritization is being employed.

2 Terminology

CSN: Circuit-switched network, encompassing both private (closed) networks and the public switched telephone network (PSTN).

ETS: Emergency telecommunications service, identifying a communications service to be used during large-scale emergencies that allows authorized individuals to communicate. Such communication may reach end points either within a closed network or any endpoint on the CSN or the Internet. The communication service may use voice, video, text or other multimedia streams.

Request: In this document, we define "request" as any SIP request. This includes call setup requests, instant message requests and event notification requests.

3 Resources

Prioritized access to at least five resource types may be useful:

Gateway resources: The number of channels (trunks) on a CSN gateway is finite. Resource prioritization may prioritize access to these channels, by priority queuing or preemption.

CSN resources: Resources in the CSN itself, away from the access gateway, may be congested. This is the domain of traditional resource prioritization as MLPP and GETS, where

circuits are granted to ETS communications based on queueing priority or preemption (if allowed by local telecommunication regulatory policy). A gateway may also use alternate routing (Section 8) to increase the probability of call completion.

Specifying CSN behavior is beyond the scope of this document, but as noted below, a central requirement is to be able to invoke all such behaviors from an IP endpoint.

IP network resources: SIP may initiate voice and multimedia sessions. In many cases, audio and video streams are inelastic and have tight delay and loss requirements. Under conditions of IP network overload, emergency services applications may not be able to obtain sufficient bandwidth in a best-effort network. While quality of service management is necessary to solve this problem, this is orthogonal to SIP, out of the scope for SIP, and as such these requirements will be discussed in another document.

Bandwidth used for SIP signaling itself may be subject to prioritization.

Receiving end system resources: End systems may include automatic call distribution systems (ACDs) or media servers as well as traditional telephone-like devices. Gateways are also end systems, but have been discussed earlier.

If the receiving end system can only manage a finite number of sessions, a prioritized call may need to preempt an existing call or indicate to the callee that a high-priority call is waiting. (The precise user agent behavior is beyond the scope of this document and considered a matter of policy and implementation.)

Such terminating services may be needed to avoid overloading, say, an emergency coordination center. However, other approaches beyond prioritization, e.g., random request dropping by geographic origin, need to be employed if the number of prioritized calls exceeds the terminating capacity. Such approaches are beyond the scope of this memo.

SIP proxy resources: While SIP proxies often have large request handling capacities, their capacity is likely to be smaller than their access network bandwidth. (This is true in particular since different SIP requests consume vastly different amounts of proxy computational resources,

depending on whether they invoke external services, sip-cgi [1] and CPL [2] scripts, etc. Thus, avoiding proxy overload by restricting access bandwidth is likely to lead to inefficient utilization of the proxy.) Therefore, some types of proxies may need to silently drop selected SIP requests under overload, reject requests, with overload indication or provide multiple queues with different drop and scheduling priorities for different types of SIP requests. However, this is strictly an implementation isssue and does not appear to influence the protocol requirements nor the on-the-wire protocol. Thus, it is out of scope for the protocol requirements discussion pursued here.

Responses should naturally receive the same treatment as the corresponding request. Responses already have to be securely mapped to requests, so this requirement does not pose a significant burden. Since proxies often do not maintain call state, it is not generally feasible to assign elevated priority to requests originating from a lower-privileged callee back to the higher-privileged caller.

There is no requirement that a single mechanism be used for all five resources.

4 Network Topologies

We consider four types of combinations of IP and circuit-switched networks.

IP end-to-end: Both request originator and destination are on an IP network, without intervening CSN-IP gateways. Here, any SIP request could be subject to prioritization.

IP-to-CSN (IP at the start): The request originator is in the IP network, while the callee is in the CSN. Clearly, this model only applies to SIP-originated phone calls, not generic SIP requests such as those supporting instant messaging services.

CSN-to-IP (IP at the end): A call originates in the CSN and terminates, via an Internet telephony gateway, in the IP network.

CSN-IP-CSN (IP bridging): This is a concatenation of the two

previous ones. It is worth calling out specifically to note that the two CSN sides may use different signaling protocols. Also, the originating CSN endpoint and the gateway to the IP network may not know the nature of the terminating CSN. Thus, encapsulation of originating CSN information is insufficient.

The bridging model (IP-CSN-IP) can be treated as the concatenation of the IP-to-CSN and CSN-to-IP cases.

It is worth emphasizing that CSN-to-IP gateways are unlikely to know whether the final destination is in the IP network, the CSN or, via SIP forking, in both.

These models differ in the type of controllable resources, identified as gateway, CSN, IP network resources, proxy and receiver. Items marked as (x) are beyond the scope of this document.

Topology	Gateway	CSN	IP	proxy	receiver
IP-end-to-end			(x)	(x)	x
IP-to-CSN	x	x	(x)	(x)	(x)
CSN-to-IP	x	x	(x)	(x)	x
CSN-IP-CSN	x	x	(x)	(x)	(x)

5 Network Models

There are at least four IP network models that influence the requirements for resource priority. Each model inherits the restrictions of the model above it.

Pre-configured for ETS: In a pre-configured network, an ETS application can use any protocol carried in IP packets and modify the behavior of existing protocols. As an example, if an ETS agency owns the IP network, it can add traffic shaping, scheduling or support for a resource reservation protocol to routers.

Transparent: In a transparent network, an ETS application can rely on the network to forward all valid IP packets, however, the ETS application cannot modify network elements. Commercial ISP offer transparent networks as long as they do not filter certain types of packets. Networks employing firewalls, NATs and "transparent" proxies are not transparent. Sometimes, these types of networks are also called common-carrier networks since they carry IP packets without concern as to their content.

SIP/RTP transparent: Networks that are SIP/RTP transparent allow users to place and receive SIP calls. The network allows ingress and egress for all valid SIP messages, possibly subject to authentication. Similarly, it allows RTP media streams in both directions. However, it may block, in either inbound or outbound direction, other protocols such as RSVP or it may disallow non-zero DSCPs. There are many degrees of SIP/RTP transparency, e.g., depending on whether firewalls require inspection of SDP content, thus precluding end-to-end encryption of certain SIP message bodies, or whether only outbound calls are allowed. Many firewalled corporate networks and semi-public access networks such as in hotels are likely to fall into this category.

Restricted SIP networks: In restricted SIP networks, users may be restricted to particular SIP applications and cannot add SIP protocol elements such as header fields or use SIP methods beyond a prescribed set. It appears likely that 3GPP/3GPP2 networks will fall into this category, at least initially.

A separate and distinct problem are SIP networks that administratively prohibit or fail to configure access to special access numbers, e.g., the 710 area code used by GETS. Such operational failures are beyond the reach of a protocol specification.

It appears desirable that ETS users can employ the broadest possible set of networks during an emergency. Thus, it appears preferable that protocol enhancements work at least in SIP/RTP transparent networks and are added explicitly to restricted SIP networks.

The existing GETS system is an example of an "opportunistic" network, allowing use from most unmodified telephones, while MLPP systems are typically pre-configured.

6 Relationship to Emergency Call Services

The resource priority mechanisms are used to have selected individuals place calls with elevated priority during times when the network is suffering from a shortage of resources. Generally, calls for emergency help placed by non-officials (e.g., "911" and "112" calls) do not need resource priority under normal circumstances. If such emergency calls are placed during emergency-induced network resource shortages, the call identifier itself is sufficient to identify the emergency nature of the call. Adding an indication of

resource priority may be less appropriate, as this would require that all such calls carry this indicator. Also, it opens another attack mechanism, where non-emergency calls are marked as emergency calls. (If the entity can recognize the request URI as an emergency call, it would not need the resource priority mechanism.)

7 SIP Call Routing

The routing of a SIP request, i.e., the proxies it visits and the UAs it ends up at, may depend on the fact that the SIP request is an ETS request. The set of destinations may be larger or smaller, depending on the SIP request routing policies implemented by proxies. For example, certain gateways may be reserved for ETS use and thus only be reached by labeled SIP requests.

8 Policy and Mechanism

Most priority mechanisms can be roughly categorized by whether they:

- use a priority queue for resource attempts,
- make additional resources available (e.g., via alternate routing (ACR)), or
- preempt existing resource users (e.g., calls.)

For example, in GETS, alternate routing attempts to use alternate GETS-enabled interexchange carriers (IXC) if it cannot be completed through the first-choice carrier.

Priority mechanisms may also exempt certain calls from network management traffic controls.

The choice between these mechanisms depends on the operational needs and characteristics of the network, e.g., on the number of active requests in the system and the fraction of prioritized calls. Generally, if the number of prioritized calls is small compared to the system capacity and the system capacity is large, it is likely that another call will naturally terminate in short order when a higher-priority call arrives. Thus, it is conceivable that the priority indication can cause preemption in some network entities, while elsewhere it just influences whether requests are queued instead of discarded and what queueing policy is being applied.

Some namespaces may inherently imply a preemption policy, while others may be silent on whether preemption is to be used or not, leaving this to local entity policy.

Similarly, the precise relationships between labels, e.g., what fraction of capacity is set aside for each priority level, is also a matter of local policy. This is similar to how differentiated services labels are handled.

9 Requirements

In the PSTN and certain private circuit-switched networks, such as those run by military organizations, calls are marked in various ways to indicate priorities. We call this a "priority scheme".

Below are some requirements for providing a similar feature in a SIP environment; security requirements are discussed in Section 10. We will refer to the feature as a "SIP indication" and to requests carrying such an indication as "labelled requests".

REQ-1: Not specific to one scheme or country: The SIP indication should support existing and future priority schemes. For example, there are currently at least four priority schemes in widespread use: Q.735, also implemented by the U.S. defense network and NATO, has five levels, the United States GETS (Government Emergency Telecommunications Systems) scheme with implied higher priority and the British Government Telephone Preference Scheme (GTPS) system, which provides three priority levels for receipt of dial tone.

The SIP indication may support these existing CSN priority schemes through the use of different name spaces.

Private-use namespaces may also be useful for certain applications.

REQ-2: Independent of particular network architecture: The SIP indication should work in the widest variety of SIP-based systems. It should not be restricted to particular operators or types of networks, such as wireless networks or protocol profiles and dialects in certain types of networks. The originator of a SIP request cannot be expected to know what kind of CS technology is used by the destination gateway.

REQ-3: Invisible to network (IP) layer: The SIP indication must be usable in IP networks that are unaware of the enhancement and in SIP/RTP-transparent networks. Obviously, such networks will not be able to provide enhanced services.

This requirement can be translated to mean that the request has to be a valid SIP request and that out-of-band signaling is not acceptable.

REQ-4: Mapping of existing schemes: Existing CSN schemes must be translatable to SIP-based systems.

REQ-5: No loss of information: For the CSN-IP-CSN case, there should be no loss of signaling information caused by transiting the IP network if both circuit-switched networks use the same priority scheme. Loss of information may be unavoidable if the destination CSN uses a different priority scheme from the origin.

One cannot assume that both CSNs are using the same signaling protocol or protocol version, such as ISUP, so that transporting ISUP objects in MIME [3,4] is unlikely to be sufficient.

REQ-6: Extensibility: Any naming scheme specified as part of the SIP indication should allow for future expansion. Expanded naming schemes may be needed as resource priority is applied in additional private networks, or if VoIP-specific priority schemes are defined.

REQ-7: Separation of policy and mechanism: The SIP indication should not describe a particular detailed treatment, as it is likely that this depends on the nature of the resource and local policy. Instead, it should invoke a particular named policy. As an example, instead of specifying that a certain SIP request should be granted queueing priority, not cause preemption, but be restricted to three-minute sessions, the request invokes a certain named policy that may well have those properties in a particular implementation. An IP-to-CSN gateway may need to be aware of the specific actions required for the policy, but the protocol indication itself should not.

Even in the CSN, the same MLPP indication may result in different behavior for different networks.

REQ-8: Request-neutral: The SIP indication chosen should work for any SIP request, not just, say, INVITE.

REQ-9: Default behavior: Network terminals configured to use a priority scheme may occasionally end up making calls in a network that does not support such a scheme. In those

cases, the protocol must support a sensible default behavior that treats the call no worse than a call that did not invoke the priority scheme. Some networks may choose to disallow calls unless they have a suitable priority marking and appropriate authentication. This is a matter of local policy.

REQ-10: Address-neutral: Any address or URI scheme may be a valid destination and must be usable with the priority scheme. The SIP indication cannot rely on identifying a set of destination addresses or URI schemes for special treatment. This requirement is motivated by existing ETS systems. For example, in GETS and similar systems, the caller can reach any PSTN destination with increased probability of call completion, not just a limited set. (This does not preclude local policy that allows or disallows, say, calls to international numbers for certain users.)

Some schemes may have an open set of destinations, such as any valid E.164 number or any valid domestic telephone number, while others may only reach a limited set of destinations.

REQ-11: Identity-independent: The user identity, such as the From header field in SIP, is insufficient to identify the priority level of the request. The same identity can issue non-prioritized requests as well as prioritized ones, with the range of priorities determined by the job function of the caller. The choice of the priority is made based on human judgement, following a set of general rules that are likely to be applied by analogy rather than precise mapping of each condition. For example, a particular circumstance may be considered similarly grave compared to one which is listed explicitly.

REQ-12: Independent of network location: While some existing CSN schemes restrict the set of priorities based on the line identity, it is recognized that future IP-based schemes should be flexible enough to avoid such reliance. Instead, a combination of authenticated user identity, user choice and policy determines the request treatment.

REQ-13: Multiple simultaneous schemes: Some user agents will need to support multiple different priority schemes, as several will remain in use in networks run by different agencies and operators. (Not all user agents will have the

means of authorizing callers using different schemes, and thus may be configured at run-time to only recognize certain namespaces.)

REQ-14: Discovery: A terminal should be able to discover which, if any, priority name spaces are supported by a network element. Discovery may be explicit, where a user agent requests a list of the supported name spaces or it may be implicit, where it attempts to use a particular name space and is then told that this name space is not supported. This does not imply that every element has to support the priority scheme. However, entities should be able discover whether a network element supports it or not.

REQ-15: Testing: It must be possible to test the system outside of emergency conditions, to increase the chances that all elements work during an actual emergency. In particular, critical elements such as indication, authentication, authorization and call routing must be testable. Testing under load is desirable. Thus, it is desirable that the SIP indication is available continuously, not just during emergencies.

REQ-16: 3PCC: The system has to work with SIP third-party call control.

REQ-17: Proxy-visible: Proxies may want to use the indication to influence request routing (see Section 7) or impose additional authentication requirements.

10 Security Requirements

Any resource priority mechanism can be abused to obtain resources and thus deny service to other users. While the indication itself does not have to provide separate authentication, any SIP request carrying such information has more rigorous authentication requirements than regular requests. Below, we describe authentication and authorization aspects, confidentiality and privacy requirements, protection against denial of service attacks and anonymity requirements. Additional discussion can be found in [5].

10.1 Authentication and Authorization

SEC-1: More rigorous: Prioritized access to network and end system resources enumerated in Section 3 imposes particularly stringent requirements on authentication and authorization mechanisms since access to prioritized resources may impact overall system stability and

performance, not just result in theft of, say, a single phone call.

The authentication and authorization requirements for ETS calls are, in particular, much stronger than for emergency calls (112, 911), where wide access is the design objective, sacrificing caller identification if necessary.

SEC-2: Attack protection: Under certain emergency conditions, the network infrastructure, including its authentication and authorization mechanism, may be under attack. Thus, authentication and authorization must be able to survive such attacks and defend the resources against these attacks.

Mechanisms to delegate authentication and to authenticate as early as possible are required. In particular, the number of packets and the amount of other resources such as computation or storage that an unauthorized user can consume needs to be minimized.

Unauthorized users must not be able to block CSN resources, as they are likely to be more scarce than packet resources. This implies that authentication and authorization must take place on the IP network side rather than using, say, a CSN circuit to authenticate oneself via a DTMF sequence.

Given the urgency during emergency events, normal statistical fraud detection may be less effective, thus placing a premium on reliable authentication.

SIP nodes should be able to independently verify the authorization of requests to receive prioritized service and not rely on transitive trust within the network.

SEC-3: Independent of mechanism: Any indication of the resource priority must be independent of the authentication mechanism, since end systems will impose different constraints on the applicable authentication mechanisms. For example, some end systems may only allow user input via a 12-digit keypad, while others may have the ability to acquire biometrics or read smartcards.

SEC-4: Non-trusted end systems: Since ETS users may use devices that are not their own, systems should support authentication mechanisms that do not require the user to reveal her secret, such as a PIN or password, to the device.

SEC-5: Replay: The authentication mechanisms must be resistant to replay attacks.

SEC-6: Cut-and-paste: The authentication mechanisms must be resistant to cut-and-paste attacks.

SEC-7: Bid-down: The authentication mechanisms must be resistant to bid down attacks.

10.2 Confidentiality and Integrity

SEC-8: Confidentiality: All aspects of ETS are likely to be sensitive and should be protected from unlawful intercept and alteration. In particular, requirements for protecting the confidentiality of communications relationships may be higher than for normal commercial service. For SIP, the To, From, Organization, Subject, Priority and Via header fields are examples of particularly sensitive information. Callers may be willing to sacrifice confidentiality if the only alternative is abandoning the call attempt.

Unauthorized users must not be able to discern that a particular request is using a resource priority mechanism, as that may reveal sensitive information about the nature of the request to the attacker. Information not required for request routing should be protected end-to-end from intermediate SIP nodes.

The act of authentication, e.g., by contacting a particular server, itself may reveal that a user is requesting prioritized service.

SIP allows protection of header fields not used for request routing via S/MIME, while hop-by-hop channel confidentiality can be provided by TLS or IPsec.

10.3 Anonymity

SEC-9: Anonymity: Some users may wish to remain anonymous to the request destination. For the reasons noted earlier, users have to authenticate themselves towards the network carrying the request. The authentication may be based on capabilities and noms, not necessarily their civil name. Clearly, they may remain anonymous towards the request destination, using the network-asserted identity and general privacy mechanisms [6,7].

10.4 Denial-of-Service Attacks

SEC-10: Denial-of-service: ETS systems are likely to be subject to deliberate denial-of-service attacks during certain types of emergencies. DOS attacks may be launched on the network itself as well as its authentication and authorization mechanism.

SEC-11: Minimize resource use by unauthorized users: Systems should minimize the amount of state, computation and network resources that an unauthorized user can command.

SEC-12: Avoid amplification: The system must not amplify attacks by causing the transmission of more than one packet or SIP request to a network address whose reachability has not been verified.

11 Security Considerations

Section 10 discusses the security issues related to priority indication for SIP in detail and derives requirements for the SIP indicator. As discussed in Section 6, identification of priority service should avoid multiple concurrent mechanisms, to avoid allowing attackers to exploit inconsistent labeling.

12 Acknowledgements

Fred Baker, Scott Bradner, Ian Brown, Ken Carlberg, Janet Gunn, Kimberly King, Rohan Mahy and James Polk provided helpful comments.

13 Bibliography

[1] J. Lennox, H. Schulzrinne, and J. Rosenberg, "Common gateway interface for SIP," RFC 3050, Internet Engineering Task Force, Jan. 2001.

[2] J. Lennox and H. Schulzrinne, "CPL: A language for user control of internet telephony services," Internet Draft, Internet Engineering Task Force, Nov. 2001. Work in progress.

[3] E. Zimmerer, J. Peterson, A. Vemuri, L. Ong, F. Audet, M. Watson, and M. Zonoun, "MIME media types for ISUP and QSIG objects," RFC 3204, Internet Engineering Task Force, Dec. 2001.

[4] A. Vemuri and J. Peterson, "Session initiation protocol for telephones (SIP-T): (SIP-T): context and architectures," RFC 3372, Internet Engineering Task Force, Sept. 2002.

[5] I. Brown, "A security framework for emergency communications," Internet Draft, Internet Engineering Task Force, June 2002. Work in progress.

[6] J. Peterson, "A privacy mechanism for the session initiation protocol (SIP)," Internet Draft, Internet Engineering Task Force, June 2002. Work in progress.

[7] M. Watson, "Short term requirements for network asserted identity," Internet Draft, Internet Engineering Task Force, June 2002. Work in progress.

14 Authors' Address

Henning Schulzrinne
Dept. of Computer Science
Columbia University
1214 Amsterdam Avenue
New York, NY 10027
USA
electronic mail: schulzrinne@cs.columbia.edu