WHAT IS THE PLACE FOR USER-NETWORK SIGNALLING IN THE 21ST CENTURY?

Jon Crowcroft, Saleem N. Bhatti, Colin Perkins[♠]
{J.Crowcroft,S.Bhatti,C.Perkins}@cs.ucl.ac.uk

Abstract

In the provision of multiservice network services, much attention has been focused on the use of **user-network (UN) signalling**. UN signalling plays an important role in connection-oriented networks. In such networks, it can be used for admission control, providing resource allocation (resource reservation), enabling new services, accounting information (for generating bills) as well as allow collection of statistics that can aid in dimensioning and capacity planning. We argue that UN signalling is not required for IP-based networks and that; a) for an IP-based network other mechanisms may be more suitable for providing the features listed above; b) to insist that there is UN signalling in IP-based networks that reproduces the signalling mechanisms already used in the telco-network can hinder the deployment of new applications in IP-based networks.

Keywords: IP, signalling, multiservice networks

Introduction

As service providers move quickly to the provision of multiservice network offerings, there is concern that an IP-based network cannot support the rich signalling infrastructure that is currently deployed in telco networks[1]. There is a misconception that it is not possible to offer multiservice networks using an IP-based infrastructure without more comprehensive **user-network (UN) signalling**. Such signalling already exists in telco networks and it can be used directly for a wide range of purposes. In a traditional telco network, signalling protocols and information from signalling can be used for admission control, providing resource allocation (resource reservation), enabling new services, accounting information (for generating bills) as well as allow collection of statistics that can aid in dimensioning and capacity planning.

An IP-based network is connectionless and, at the IP-layer, there are really only two service primitives – one for sending data and one for receiving data. The introduction of UN signalling – signalling at the IP layer – would bring much additional complexity and packet processing requirements for the routers which are already heavily burdened with the current load of IP traffic.

We argue that there is little requirement for UN signalling in an IP-based network. While some form of signalling may be required for new applications, it is instructive to think carefully about the functions that signalling provides in a telco network and decide carefully where such functionality should be implemented (if at all) in an IP-based network. For an IP-based network:

- there may be other mechanisms or protocols that are more suitable for providing the features associated with traditional telco UN signalling;
- the introduction of UN signalling in IP-networks to reproduce the functions that exist in the telco-network can hinder the deployment of new applications in IP-based networks.

[♠] Computer Science Department, University College London, http://www.cs.ucl.ac.uk/, http://www-mice.cs.ucl.ac.uk/

[1] We use the term “telco network” to refer to the legacy PSTN/ISDN/GSM network infrastructure deployed in the public telephone network.

Admission control and resource allocation

In a telco network, admission control can be seen as a simple "YES/NO" decision made by the network. However, in a multiservice packet network, there are many possible outcomes to a user's attempt to use the network, e.g. choice of audio and video codecs, frame rates for video, etc. In a telco network a user typically has a fixed choice of quality of service (QoS), but in a multiservice packet network, the choice is a decision made according to a combination of the user's requirements, the capability of the application (and the end-system on which it runs) and the capability of the network at that time [BK98]. The way that admission control and resource allocation is performed is likely to be different for different instances of the same application, given different physical connectivity, different service providers, different users, etc. Such dynamism and heterogeneity is likely to be application-specific and hard to capture in a single UN signalling mechanism.

The Internet Engineering Task Force (IETF) has already tried to provide a network-level signalling protocol for per-flow resource reservation using the IETF INTSERV model[2] and RSVP [RFC2210]. While RSVP is arguably rich in functionality, it has poor scalability, and the Internet community acknowledges that it would be difficult to deploy INTSERV/RSVP on a large scale [RFC2208]. The main problems with the approach of INTSERV/RSVP with respect to signalling can be summarised in the bullet points below:

- the use of per-flow state would require backbone routers to hold state information for end-to-end flows across the Internet, which is clearly not scaleable without some sort of aggregation of state
- the soft-state approach, while allowing the network to be robust to end-system failures, means that there is the potential for much additional traffic due to signalling packets that are needed to refresh the soft-state
- the two-pass reservation set-up used in INTSERV/RSVP, while allowing the receiver to make the final decision on resources reserved, introduces additional packet exchanges for each flow without actually allowing end-systems to become aware of each others capabilities
- there is no integration with routing mechanisms so a change of network path (due to normal routing behaviour) requires additional signalling messages along the new path in order to reserve resources, with no guarantee of the same QoS on the new network path
- RSVP/INTSERV tries to combine user-network, network-network and user-user signalling, so there are practical problems in everyday use, e.g. how to set the duration of soft-state timers when user interaction is required before reservations are confirmed
- as charging is application-specific, RSVP routers must use the network-level information to provide application-specific accounting information – an unwelcome administrative task for busy routers

For IP-based networks, the favoured model for large-scale deployment at the time of writing is that of Differentiated Services [RFC2475] as is currently being defined within the IETF DIFFSERV WG[3]. This uses a fixed number of per-hop behaviour (PHB) definitions that indicate how packets are handled by the network on a hop-by-hop basis. In the DIFFSERV scheme, the semantics of the PHBs are implemented in routers and other network elements. An 8-bit field in the header of each IP packet identifies the PHB used for each packet. This is the only "signalling" involved. Admission control is performed using a service level agreement (SLA) at provider/user interface to police traffic. PHBs identify traffic with the same handling requirements so the admission control and resource control is on aggregated traffic flows and not per application instance as in INTSERV. This offers better scaling properties. DIFFSERV relies on correct policing of traffic within the network in order in order to avoid congestion due to malicious traffic sources or misbehaving traffic sources. Traffic conditioners in routers can be programmed to delay or drop packets from misbehaving aggregate flows but there is no explicit congestion signal back to the source. So, how does the network signal the presence of congestion to the end users?

[2] http://www.ietf.org/html.charters/intserv-charter.html

[3] http://www.ietf.org/html.charters/diffserv-charter.html

Congestion control

A very important function in all packet switched networks (be they VC-based networks – such as ATM and frame relay, or connectionless – such as IP-based networks) is the provision of congestion control mechanisms. In IP-based networks, currently there is no *explicit* congestion signalling[4]. When an IP-based network gets congested, packets are dropped. TCP uses these dropped packets as *implicit* congestion control signals coupled with a packet acknowledgment (ACK) scheme. TCP has its own sophisticated congestion control mechanism based on the detection of missing/delayed/lost ACKs. There are two main problems with relying solely on packet drops. Firstly, on some links (e.g. wireless links), it is not possible to distinguish between transient packet corruption (such as hand-off between cells in a wireless network) and real congestion. In TCP, mechanisms such as SACK [RFC2018] can help here. Secondly, for real-time applications that use UDP and no ACK packets, there is no general way of "detecting" congestion in the network. So, the Internet community is considering the use of Explicit Congestion Notification (ECN) for IP [RFC2481] using simple bit flags in the IP packet header. However, this is not really user-network signalling – rather it is network-user signalling! There is a large body of work to show that such simple bit-flags in packet headers can be used as effective congestion control mechanisms (e.g. Jain and Ramakrishnan's original work in [JR88]), and indeed simple bit-flags are already used in frame relay – the forward explicit congestion notification (FECN) and backward explicit congestion notification (BECN).

Also, there are activities within the Internet community to make real-time traffic adaptive in a TCP-like way to provide network-friendly back-off behaviour of real-time flow rates during congestion [TCPF].

Charging

It is argued that another mechanism enabled by signalling is charging. However, it is possible to charge for network usage in an IP-based network by looking at the source address of the IP packet. There will be a requirement to authenticate the IP packet but that is increasingly becoming the requirement of the communication itself. Indeed, in recognition of this, the IETF IPSEC WG[5] has produced a standard for authentication of IP packets [RFC2402], which can be applied to IPv4 and is mandatory for IPv6. (Authentication of packets would also be useful in reducing the kind of denial of service attacks that have been reported recently.) Since each data packet may have to be authenticated anyway, any additional signalling just consumes network capacity unnecessarily, and the signalling messages would also need to be authenticated, adding to the processing required for each flow. Separating signalling from data means that you need to certify that both are related! This is easy in a monopoly-like access system such as the telco networks, but highly impractical with end-to-end signalling messages across the Internet.

At the edge of the network, the network service provider could (authenticate and) police packets from their users. In the core of the network, providers deal on large aggregates traffic between their networks using SLAs for policing. This is the only way that charging can scale to the kind of volume expected across the Internet. It is likely that even if RSVP-like signalling does make a comeback on an edge-to-edge basis or user/provider interface, the core network will not see user-network signalling messages.

Additionally, many of the functions and services that are traditionally charged for in telco networks (e.g. call forwarding, voice-mail, etc.) are application specific and so map to application-level functions in IP-based applications. So, application-level signalling seems the most appropriate place for such functions and for the charging of such functions.

Dimensioning and capacity planning

For a telco network offering a single service, it is possible to use logs and statistics of UN signalling messages to assess future dimensioning and perform capacity planning. This is because the signalling messages each refer to a single flow and all flows are can be thought of as being constructed of (multiples of) identical lower-level multiplexes, e.g. phone calls, ISDN calls. However, we have already stated that in a multiservice

[4] The ICMP Source Quench has practical problems when used and so is not widely deployed.

[5] http://www.ietf.org/html.charters/ipsec-charter.html

network, each flow can be quite different, depending on user requirements, as well as application capability and network capability. Therefore, collection of statistical information and data for network usage is by using network management tools that can measure flows and monitor network activity.

So, in large IP-based networks, there is still a need to monitor and analyse network traffic. As well as presenting some interesting and challenging engineering problems, such measurement and analysis of traffic in a scaleable way is still a research issue. There is much collaborative work in this area, e.g. Cooperative Association for Internet Data Analysis (CAIDA)[6]. To introduce signalling systems used in telco networks for this purpose before it is understood if such mechanisms would actually be useful may not be beneficial to the scaling of IP-based networks.

Note that it is the simplicity of IP, the relative homogeneity of the IP service and the *lack* of complex signalling at the IP layer that gives better scope for growth and dimensioning of the network. It also makes it possible to have easy introduction of heterogeneity and richness of services and protocols at the application-layer, as we will discuss in the next section.

Application-level signalling

The Internet community sees most signalling needs as being application-specific rather than network specific. So, the signalling mechanisms are not part of the network layer but are application-layer protocols. This allows IP-applications to evolve at their own pace, without relying on any specific network-layer mechanisms and allows many different types of applications (all with very different signalling requirements) to co-exist. IP-based applications are being developed rapidly and often go through several evolutions in quick succession, with updates to their signalling requirements being part of that evolutionary process. For example, for voice-over-IP (VoIP), the ITU Recommendation H.323 [H.323] has gone from version 1 in 1996 to version 3 in 1999, with changes (improvements) to the signalling in each version. As the signalling is at the application layer, it is possible to introduce different versions of H.323 into networks, as required, with many versions of H.323 in operation concurrently, without any changes to the network elements. This kind of service deployment would be much harder if each change to the signalling protocol required upgrades to all network elements. Also, other VoIP or session control signalling protocols, such as the Session Initiation Protocol (SIP) [SIP], can be operated in parallel over the same network.

The application-layer is also the place to deal with application-specific heterogeneity, such as different QoS for end users in a large multicast conference. Such mechanisms as conference control protocols (e.g. [KC98]), user clustering and transcoding gateways (e.g. [KHC98]) and layered codecs (e.g. [MJV96]) can support such mechanisms with application-specific signalling.

The application layer is also the appropriate place to account for the invocation and usage of application-specific services and functions. So using signalling information at the application layer rather than network-level signalling information for accounting purposes seems more appropriate for charging and billing in IP-based networks.

Enabling new services and applications

As increasing numbers of users start to use IP-based networks and the Internet, there is a need to have scalability and stability at the network layer. Any changes to the IP layer affects a huge number of users. However, there is also a need to allow new applications and services to be introduced dynamically and quickly. If changes to the network layer are required for new services and applications, then these two requirements are in contradiction – it is not possible to have stability and dynamism in the network layer at the same time! Therefore, the focus within the Internet community has been to aim for stability and performance at the network layer, and move “intelligence” related to applications and services into the application layer. This allows new services to be introduced without changes to the network, by making available the application-level entities – servers and clients – to provide those new services.

[6] http://www.caida.org/

However, this is an oversimplification. There are some changes being proposed to the network layer (e.g. IPv6[7], DIFFSERV, IPSEC), but these are not in support on any *one* particular application, but rather to provide general mechanisms to support *many* different applications. This approach produces a scaleable and extensible infrastructure for the introduction of new services and applications. As discussed in the previous section, new protocols or extensions and enhancements to existing protocols can be introduced easily. However, this means that the network is not really optimised for the operation of any *one* application, so there is often a requirement for much new functionality to be introduced at the application layer. The provision of VoIP is a good example of this.

Also note that there are some application-level features available in IP-based applications (for example VoIP) which could be harder to support in a traditional telco network but are more easily supported in an IP-based network, e.g. multicast communication. There may not be a direct mapping of signalling messages between the telco networks and an IP-based network in such cases, as we explore in the next section.

IP-network/telco-network integration

Interworking and integration between public switched networks (such as PSTN and ISDN) and the IP-based networks will require mapping of signalling messages. However, it only makes sense to map those elements of signalling protocols that can be supported by IP-based applications. It may not make sense to map all signalling messages from say, SS7 or Q.931, directly across to IP. In fact this is visible in the evolution of H.323 from version 1 to version 3, which shows how the use of Q.931 and H.245 has been streamlined and adapted in order to improve performance.

There are various IETF working groups looking at the integration of IP-based networks and telco based signalling:

- SIP: it is possible to tunnel specific signalling messages from telco networks in SIP INFO messages
- MEGACO[8]: support for “dumb” terminal devices – H.323 and SIP assume intelligent user terminals such as workstations
- IPTEL[9]: provision of IP ⇔ PSTN gateways and protocols for gateway location
- PINT[10]: integration of PSTN and Internet applications at the service level where IP-based applications can request PSTN-based services
- SPIRITS[11]: in compliment to PINT, how service requests from a telco Intelligent Network (IN) system might be mapped to IP-based applications and services

These groups are at pretty early stages in their work. In the Internet domain, these are all application-level signalling protocols.

Conclusion

The traditional telco network is based on user-network (UN) signalling for a specific service resource that is described simply at lower layers, i.e. a physical-level multiplex giving access to network transmission capability. Even on a virtual circuit (VC) based network, while we are not always assured fixed resources, we have some confidence of the kind of service we will get from the network. UN signalling in a telco network makes requests for these relatively fixed resources. In an IP-based network, users may need different resources at different times for the same application. Different parts of the network may or may not be able to support the needs of the application at that time. Also many applications may be in operation at one time. UN

[7] http://www.ietf.org/html.charters/ipngwg-charter.html

[8] http://www.ietf.org/html.charters/megaco-charter.html

[9] http://www.ietf.org/html.charters/iptel-charter.html

[10] http://www.ietf.org/html.charters/pint-charter.html

[11] http://www.ietf.org/html.charters/spirits-charter.html

signalling in IP-based networks will not scale across the backbone networks without the use of aggregated state information. The introduction of signalling protocols at the network layer may hinder evolution of IP-networks and introduction of new applications and services.

We have argued that the traditional telco uses for signalling – admission control, providing resource allocation (resource reservation), enabling new services, accounting information (for generating bills), dimensioning and capacity planning – can be fulfilled by other mechanisms in IP-based networks. There is a need, however for some network-user signalling in the form of congestion control flags, as well as network-level policing of traffic. Neither of these two functions require UN signalling. Other telco UN signalling functions can be handled in application-specific signalling protocols in IP-based networks.

References

[BK98] S. N. Bhatti, G. Knight, “QoS Assurance vs. Dynamic Adaptability for Applications”, Proc. NOSSDAV98 - 8th International Workshop on Network and Operating System Support for Digital Audio and Video, New Hall, Cambridge, UK, 8-10 July 1998

[H.323] ITU Rec. H.323, “Packet-based Multimedia Communication Systems”

[JR88] R. Jain and K. Ramakrishnan, “Congestion Avoidance in Computer Networks with A Connectionless Network Layer: Concepts, Goals, and Methodology,” Proc. Computer Networking Symposium, Washington, D.C., April 11-13, 1988, pp. 134-143

[KC98] N. Kausar & J. Crowcroft, “General Conference Control Protocol”, IEE Telecommunication 1998, Edinburgh, 1998.

[KHC98] I. Kouvelas, V. Hardman & J. Crowcroft, “Network Adaptive Continuous-Media Applications Through Self Organised Transcoding”, Proc. NOSSDAV98 - 8th International Workshop on Network and Operating System Support for Digital Audio and Video, New Hall, Cambridge, UK, 8-10 July 1998.

[MJV96] Steven McCanne, Van Jacobson and Martin Vetterli, “Receiver-Driven Layered Multicast”, Proc. SIGCOMM96, Palo Alto, California, pp. 117-130, August 1996.

[RFC2018] M. Mathis, J. Mahdavi, S. Floyd, A. Romanow, “TCP Selective Acknowledgement Options”, RFC2018, October 1996.

[RFC2208] A. Mankin, Ed., F. Baker, B. Braden, S. Bradner, M. O`Dell, A. Romanow, A. Weinrib, L. Zhang, “Resource ReSerVation Protocol (RSVP) -- Version 1 Applicability Statement Some Guidelines on Deployment”, RFC2208, September 1997.

[RFC2210] J. Wroclawski, “The Use of RSVP with IETF Integrated Services”, RFC2210, September 1997.

[RFC2402] S. Kent, R. Atkinson, “IP Authentication Header”, RFC2402, November 1998.

[RFC2475] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, “An Architecture for Differentiated Service”, RFC2475, December 1998.

[RFC2481] K. Ramakrishnan, S. Floyd, “A Proposal to add Explicit Congestion Notification (ECN) to IP”, RFC2481, January 1999.

[SIP] http://www.cs.columbia.edu/sip/

[TCPF] http://www.psc.edu/networking/tcp_friendly.html