Public Review for

Flow Labelled IP over ATM: Design and Rationale

Greg Minshall, Bob Hinden, Eric Hoffman, Fong Ching Liaw, Tom Lyon, and Peter Newman

Flow labelled IP over ATM was both an architecture and a technology to solve a number of problems that emerged from the train-crash that was the collision between Broadband-ISDN and the Internet.

It was not merely a hack to work around the mismatch in semantics between a virtual circuit switched underlay and a stateless datagram routed overlay. The intention was to take advantage of facilities offered by the underlay technology such as flow state maintenance and flowisolation and protection in switches, to offer QoS at various levels of granularity, whether for unicast or multicast. The paper was one of the seminal influences for what is now the burgeoning MPLS industry, although in the architecture described we see something simpler, more elegant, and, perhaps, more aligned to the traditional Internet design philosphy [Clark88], than to the telecommunications mindset in the hardcore Broadband Integrated Services Digital Networks (B-ISDN) Asynchronous Transfer Mode (ATM) proponents[1].

Of course there has been a long and valiant history of tweaking IP (and other datagram protocols) to run over virtual circuit (VC) networks, dating back to the need to run IP over X.25 in the 1970s-1980s. Typically, there was no advantage taken of the underlying circuit technology (indeed there was a disadvantage in having to use it), and so the main techniques developed involved heuristics for triggering circuit setup and tear-down from a shim module that sit between IP and the VC layer and monitored IP level activity. Alongside, if you were lucky, there might be a management knob to allow you to adjust the laxity or aggressiveness of the shim, depending on potential cost/performance tradeoff decisions: X.25, and later IP over dial-up typically incurred monetary costs by time and/or volume.

B-ISDN (and, guilty by association, ATM) was originally intended as an overarching system for networking, replacing IP. It was too complex and too late to survive in that role, although the underlying cell-switching technology is still a vital component in today's; network landscape (many xDSL lines use ATM as a multiplexing layer to allow low latency voice circuits to run without over-provisioning, alongside IP and video traffic - at lower speeds, such as the typical xDSL deployment of a handful of megabits a second, this is fairly crucial). However, in the panoply of B-ISDN there was a whole armoury of signalling protocols (UNI, NNI etc).

The authors have strong systems credentials as well as networking, and this shows in their interpretation of the problem, and the architecting of solutions including the details. They rightly ignored the complex smorgasbord of B-ISDN and seized on what ATM cell switching could offer[2]. Take-home messages from this paper are:

• You can separate routing from forwarding. If forwarding can all be done flexibly enough in hardware but cheaply and faster, then make sense of it. At the time, ATM switches did this nicely.

• The ideas here stem from understanding the nature of a flow in the IP world properly. Having noticed that one can infer the nature of a flow from its behaviour, one can build a system that *evolves* its cell-switching behaviour to match as the inference becomes stronger - this can apply at varying time scales. However, to bootstrap you have to get packets there - given IP is best effort, then a default VC is needed to get default packets to the default next hop with the default performance.

• Once you know the nature of a flow, it can be isolated and protected: VCs let you do this, but you need to set them up. A neat observation is that this can be done in the reverse direction to the user data plane. You need a protocol to do it, but you do not need PNNI or Q.2931 or SS7 (these protocols are what the telecom world had or was developing at the time but had a number of other purposes, akin to the Internet's RSVP). So the authors devised a simple enough protocol, but no simpler, called IFMP. It is analogous to today's LDP in the MPLS world. Actually,

[1] It is a matter of history now, but there is a clear distinction between the core concepts of cell switching that are embodied in ATM, and the more general network architecture of which ATM is only a part, which was known in the ITU as B-ISDN, and included a number of other control plane systems for signaling and routing and other functions.

[2] One of the authors worked in Cambridge previously at a time when Cambridge rings were being interconnected - this was very much a cell switched virtual circuit based network architecture, but the style of control plane was much more like what we see in this paper. If I was to hazard a guess, I would say that this is at least partly the root of some ideas here.

its simpler, and (as they say) more like an ICMP redirect:

• Receiver initiation lets you cope with multicast (c.f. RSVP).

• From time to time, you need to tell the switches more things to do, so you need a switch management protocol. This isn't just a MIB plus get/set, so maybe a bespoke general switch management protocol (GSMP) is needed, so they built one.

The fact that the authors really built this shows up in section 6 of the paper, where they list the grot (short for grotesque) that one has to cope with in reality. Problems such as what to do about TTL decrementing, IP header checksumming, fragmentation, and also cell interleaving in the ATM layer, are all things we are familiar with. Some of these "features" of IP are know frowned on, (indeed, left out in IPv6). The latter problem is due to the fixed small (surreally, 53 byte) cell size of ATM. This is something we don't have to suffer in modern switches since as you speed up (e.g. to Tbps), the header/payload cost/tradeoff starts to hurt both in capacity wasted and in VCI lookup rate.

So the router level gets an IP packet, finds the next hop is via a default VCI through a switch to another router. The packet gets to the far side after being shredded and un-shredded, and its noticed that this is part of a "longer lived flow" or indeed perhaps part of a long lived aggregate flow. A message is sent by the receiving router through the switches to request a new VC to be configured (or to add this flow to an existing one). Since the switch level probably has per-flow queues (in hardware) it can offer proper flow isolation and protection properties. Hence QoS. Most protocols at and above the IP level have some sort of setup and feedback packets (whether TCP or RTP based or DCCP, SCTP or other new spangled fine shiny transport protocols), and so one can induce various properties from packet analysis (or management) about whether to set things up non-default or not. Mice becoming elephants can receive this treatment naturally.

This paper has travelled forward in time by ten years as if by Tardis [TARDIS] to appear just when it is needed. The NSF FIND [FIND] program (and similar initiatives in other parts of the world) is finding many researchers revisiting network architectures, and while the intention is to have a "clean-slate" approach, looking at the tension between *two* architectures that is revealed here may be one way intellectually to create novel network system designs.

Of course, we know now that some of the things that the ATM switch layer was being used for here could indeed be done at line speed in a native IP forwarding device. Nevertheless, other ideas here remain highly valid. The systems ideas (defaults, late binding, simplicity of receiver based signalling), and the observations of things that cause problems to stateful networking and high speed hardware based forwarding (TTL, checksum, fragment/cell/framing) are all constant worries and concerns for the future network architect just as much as for the past.

I remember reading the Ipsilon white papers when they came out, and then watching the MPLS story unfold consequently with all its complexity. While that complexity is driven by many real-world (a.k.a customer) needs, the elegance and simplicity of this original work is beguiling and I encourage you to read it.

References

[CLARK] The Design Philosophy of the *DARPA* Internet Protocols, David D. Clark, Proceedings of ACM SIGCOMM, pp 106-114, August 1988, Stanford, CA, USA

[ATM] Asynchronous Transfer Mode: Solution for Broadband ISDN by Martin De Prycker, Ellis Horwood Ltd; 2nd ed.. edition (August 1993), ISBN: 0131785427

[FIND] http://find.isi.edu/

[TARDIS] http://en.wikipedia.org/wiki/TARDIS "The interior of the Tardis is very much bigger than the exterior..."

Public review written by
Jon Crowcroft
University of Cambridge, UK