Incident Report

Major Network Incident 21st March 2013

22nd March 2013
Version 1.0

Introduction:

The report below is a breakdown of the incident that was experienced across our network from approximately 18.30hrs on Thursday 21st March 2013. We will look at the cause of the issue as well as the effects experienced by our customers. We will also give information about actions taken to prevent any re-occurrence.

Breakdown of the events:

At approximately 18.30hrs on 21st March 2013 we became aware of loss of service to a large number of customers, which appeared to be spread across all sites and services. Our systems and network team was immediately engaged and senior members of the team attended our Telford HQ to aid with the diagnostics and resolution. Initial investigations indicated that whilst we were able to see that the network was available internally and we could contact individual devices on the core, any layer 3 traffic was not flowing into and out of our network at any of our access points.

Our DSL based customers also experienced issues as they were also unable to pass traffic, or after rebooting their CPE, were unable to authenticate with our services.

The cause of the issue was traced at approximately 19.15hrs when our engineers identified a TCAM memory overrun in one of our core router logs. This indicated that the core routers were seeing excessive BGP routes, hence the overrun. This was then rapidly traced to our lab environment within our Telford HQ where a stress test of new hardware and equipment was taking place. The traffic and routes from the stress test that were causing the issue were being advertised on our core network, these would normally be filtered by our edge BGP filters. However a misconfigured filter was in place, which meant that the traffic was able to affect our core.

This stress test was immediately halted with the expectation that the network would return to normal. This did not happen. Further diagnostics showed that the core routers were no longer using hardware forwarding to exchange routes and had reverted to lower capacity software forwarding as a result of the overruns and problems seen. As a result services were not resumed upon the cessation of the stress test.

Our engineers unsuccessfully attempted to force the routers back into the normal hardware forwarding state. The decision therefore was taken to reboot our core routers individually as this would restore the service in a shortest possible timeframe. A series of reboots were therefore undertaken which saw the bulk of services restored by 20.45hrs, with two sites - Manchester TE and Colindale - remaining problematic.

Following the reboot of these routers, we were unable to contact them via the normal methods and hence attempted to via our out of bands (OOB) methods. This allowed us to contact the Manchester TE core router. A blade in this router had not returned to normal service and a further reboot corrected this. Manchester TE therefore became available at 21.29hrs. Attempts to access our Colindale equipment via OOB continued and, whilst we could connect to the equipment, we were unable to restore the service as a further reboot of the entire router showed a similar blade problem to that experienced in Manchester. Engineers were dispatched to site to ensure that all blades etc. were correctly seated and to work locally on the hardware.

Whilst the engineers were underway, further attempts to restore service continued and each individual blade was rebooted. This returned service to the Colindale equipment at 00.23hrs and the small number of remaining affected customers should have had normal service from this point. The dispatched engineers were instructed to return to base at this point.

Effects seen:

- IPv4 Layer 3 traffic into and out of our core network was unable to pass
- All services on core routers were unavailable during the reboot period
- Services on Manchester TE core and Colindale core were affected for extended time periods due to subsequent blade issues
- DSL services were unable to pass traffic and, given that customers authenticate via Layer 3 traffic, any attempts to connect / re-authenticate failed with a BT error page
- Our VoIP based phone system for technical support was unavailable to a proportion of our customer base
- Inability to access our NOC site during periods of the outage prevented us from using this communications channel to update customers

Actions taken:

The immediate action taken to rectify this problem was to correct the edge BGP filters between our lab equipment and our core. This is now in place and will remain indefinitely. We are currently reviewing whether a double level of filter can be introduced as an extra layer of protection and will implement this when possible.

We are committed to undertaking a review of our incident management which may involve the introduction of a logically separate NOC service and the transfer of our support desk to an alternative phone solution.

Summary:

We appreciate that this was a major incident for both us and our customers that rely on our network. We sincerely apologise for the incident and reassure you that it is being dealt with with the greatest priority.