# Incident Report: EQX-CHI-CR5 to CHIC-CR5 Instability

Last updated: 08/20/2019

## Metadata

**ESnet Ticket Number:** ESNET-20190816-001
**Incident Start Time:** 08/16/2019 06:27:51 PDT
**Incident Resolution Time:** 08/20/2019 15:31:00

**Outage Start Time:** 08/16/2019 06:36:20 PDT
**Outage Resolution Time:** 08/16/2019 08:48:05 PDT

## Incident description

- On the morning of August 16th, the EQX-CHI-CR5 router began to clock ingress FCS errors on port 1/1/1 (facing CHIC-CR5), on circuit ID ESNET-CHIC-EQCH-100GE-66221. The physical port, BFD, and IS-IS sessions remained up throughout the outage duration. iBGP sessions between ESnet routers which transited the link began flapping due to packet loss. This resulted in widespread routing instability.

## Impacted Sites

- The affected circuit connected ESnet backbone routers, so all East-West traffic through Chicago was potentially affected. Note that this circuit is part of an ECMP group, the loss of connectivity may have been partial instead of a total outage for some traffic patterns.
- Ames Lab shut down eBGP peering sessions w/ ESnet and relied on their backup service provider for connectivity during the outage.
- Mark Lukasczyk, BNL site coordinator called Mike O'C at 8:44AM Pacific. Just after we turned down the circuit. They saw the brief PMC notification, however they would have like to have had it earlier.

## Incident Resolution

- Temporary fix: After localizing the fault, the 100G link between CHIC-CR5 and EQX-CHI-CR5 (134.55.218.60/30) was taken out-of-service. All higher-layer routing protocols (iBGP) stabilized at that point.

- Permanent fix: Replacing the 100G IMM line card at CHIC-CR5 facing EQX-CHI-CR5 during emergency maintenance window on Tue Aug 20 2019 starting at 15:00:00 (US/Pacific) permanently resolved this issue.

## Incident Chronology*

* Note that this chronology is reconstructed based on electronic communication plus the best recollection of the incident participants.

| Date/Time (PDT) | Event/Action | Note |
|---|---|---|
| 08/16/2019 06:27:51 | First indication of incrementing errors on EQX-CHI-CR5 router | This was the first sampling interval in which our stats collection system measured errors. |
| 08/16/2019 06:36:20 | First log entry indicating a BGP failure on the FORR-RT2 router | |
| 08/16/2019 06:38:16 | First event logged by Spectrum that EQX-CHI-CR5 port 1/1/1 stopped responding to polls and kept flapping | This occurred over 60+ times but Spectrum alarms were not generated<br>Link didn't flap; This is one of many false positives in spectrum at this time due to packet loss caused by this incident. |
| 08/16/2019 06:50:00 | NOC contacted OCS (via Slack) | |
| 08/16/2019 06:59:00 | OCS indicates they are driving in to LBL; requests escalation to POW | |
| 08/16/2019 07:17:00 | NOC contacted POW (left message) | |
| 08/16/2019 07:20:00 | Alternate engineer online | |
| 08/16/2019 07:29:00 | NOC contacted POW (left message) | |
| 08/16/2019 07:36:36 | ServiceNow ticket created by NOC | ESNET-20190816-001 |
| 08/16/2019 07:48:00 | Major Incident declared<br>NOC manager engaged | |
| 08/16/2019 07:57:00 | POW on-line | |

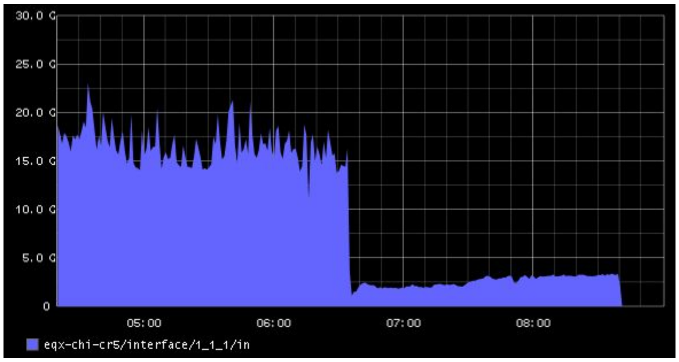
| | | |
|---|---|---|
| 08/16/2019 08:10:00 | NOC experiencing trouble reaching PMC (current instance is on east coast) | Current PMC instance is on east coast (pmc-east.es.net) |
| 08/16/2019 08:35:00 | PMC Outage Notification Sent | |
| 08/16/2019 08:40:40 | The port clocking ingress FCS errors (EQX-CHI-CR5 port 1/1/1) is manually shutdown by ESnet engineers, leading to recovery | |
| 08/16/2019 08:42:57 | netlog/syslog reports core routers are stable | Low-speed edge devices at PANTEX and ORAU excepted, which are slow to process the flood of BGP updates. |
| 08/16/2019 08:48:05 | PANTEX and ORAU routers stable after processing all pending BGP updates | Took ~8min for KRT queue to clear (‘show krt queue’) |
| 08/16/2019 09:12:14 | Initial notification to ASCR | |
| 08/16/2019 09:15:00 | Nokia support case open | |
| 08/16/2019 09:17:39 | Initial notification sent to ESCC | Wrong end timestamp mentioned |
| 08/16/2019 09:48:55 | Followup notification to ESCC with additional details and corrected timestamp | |
| 08/16/2019 09:49:48 | Followup notification to ASCR | |
| 08/16/2019 10:27:00 | ESnet engineers collected Ciena PM data. Determined source of FCS errors to be coming from the CHIC-CR5 router.<br>*[for 15-min granularity, must be captured within 8-hours of failure]* | Optical system counted and transparently passed the errored frames from CHIC-CR5 to EQX-CHI-CR5 |
| 08/16/2019 12:??:?? | ESnet engineers reset CFP on CHIC-CR5 (‘clear mda’) | Determined this did not fix the issue.<br>Line card has *not* yet been reset (‘clear card’). |
| 08/17/2019 09:45:01 | Planned maintenance announced for permanent fix, to occur on Tue Aug 20 | |

| | | |
|---|---|---|
| | 2019 at 15:00:00 (US/Pacific). | |
| 08/20/2019 15:31:00 | 100G CFP replaced, no change. Line card replaced, problem was resolved. Test stream shows no FCS errors. | FCS errors still accumulated rapidly after replacing just the pluggable optic. |
| 08/20/2019 15:57:00 | Link metrics updated (not ECMP) | |

**Response and Observations**

ESnet monitoring system (Spectrum) did not alarm specifically on the incrementing FCS errors. There were many false positive alarms in Spectrum that were side-effects of the actual root cause.

Fault localization was exceedingly difficult given the circumstances. Just before 08:40:40, ESnet engineers observed that the traffic on the CHIC to EQX-CHI link had experienced an abnormal dropoff, we recognized immediately that this is a sign of soft failure and noted that it went almost flat precisely at when the iBGP failures occurred. We confirmed a high rate of errors via the CLI interface and proceeded to shutdown the router interface with a high degree of confidence that core routing would stabilize afterwards.



[dropoff in traffic on the affected link due to the soft failure]

As a result of a previous outage (ESNET-20160424-002), a feature on our Nokia routers (“CRC Monitor”) was considered. “CRC Monitor” will automatically transition a physical port taking these errors to self-admin down. This would have minimized the the outage impact of this incident as well as sent up a beacon to identify where the issue was occurring. We elected not to implement “CRC Monitor” at the time due to concerns of unanticipated consequences from this automated response (e.g. network partitioning in the worst case).

Our execution of our major incident communication plan was slower than it should have been. The first notification to ESCC and ASCR stakeholders was **after** the outage was resolved.

## Action Plan

- Action: investigate why spectrum did not alarm on root cause FCS errors and take corrective action if defect found
- Action: investigate possible integration of other data sources (snmp, netbeam) which did observe the incrementing errors
- Action: review major incident communication plan and its implementation during this incident with a goal of speeding up initial notification to ASCR, ESCC stakeholders.
- Action: investigate creation of an ESnet "Status" page.
- Action: reevaluate the decision re: “CRC Monitor” for backbone Nokia interfaces.