Analyzing Causes of Failures in the Global Research Network Using Active Measurements

Eugene S. Myakotnykh, Bjarne E. Helvik
Centre for Quantifiable Quality of Service in Communication Systems (Q2S) [1]
Norwegian University of Science and Technology, Trondheim, Norway

Jon Kåre Hellan, Olav Kvittem, Trond Skjesol, Otto J. Wittner, Arne Øslebø
UNINETT [2]

Abstract—**With the objective to better understand how the global Internet should achieve an availability in the order of five nines, i.e. be available 0.99999 of the time, active measurements were performed between Norway and China through the Global Research Network. End-to-end downtime statistics was continuously collected during a 3-month period up to mid February 2010. In addition to periodically sending probe packets between the two measurement systems, traceroute was used every two minutes to identify an exact IP-level path between the end-points. Also, TTL (time-to-live) counter in the IP-header, which is reduced by one on every hop, was analyzed for each packet. Causes of the observed network failures based on the collected data were identified and insight is gained into processes preceding and following communication downtimes. We distinguish inter- and intradomain failures and, when possible, identify an exact link or an Autonomous System where a certain event has happened. The study shows that the end-to-end path availability is mainly affected by interdomain failures and long BGP convergence time as well as series of events not straight forwardly explained by the anticipated (re)routing behavior.**

Keywords- dependability, failure analysis, failure detection, network measurements, Quality of Services, routing

I. INTRODUCTION

The Internet is a significant part of today's communication infrastructure. It is commonly recognized that such an infrastructure should provide transport services with at least an availability in the order of five nines, i.e., the service should be available at least 0.99999 of the time. We are not there yet, although measurements reported in this paper indicate that the global academic network is close to four nines. To improve further, it is necessary to identify the dependability bottlenecks and where unintended aspects of fault handling arise. The paper addresses these issues. Based on availability measurements across the global Internet, its objectives are to identify types of arising network failures, their outage (downtime) characteristics and to better understand processes preceding and following communication downtimes.

Active measurements are used for this purpose, based on probe packets spaced 10 ms apart and traceroutes every two minutes. The set-up of this network measurement project consists of two measurement nodes, one located at UNINETT facility in Trondheim, Norway, and one at CERNET (China Education and Research Network) in Beijing. UNINETT and CERNET are interconnected through the Global Research Network crossing several other European Research Networks like NORDUnet, Geant2 and TEIN3. Based on collected measurement data, we analyze a variety of routes and potential reasons of downtimes between the end-points. The analyzed period in this paper is from mid November 2009 till mid February 2010. Measurement data collected in the end-points is not sufficient to fully understand the cause of every single event. We seek to distinguish inter- and intradomain downtimes and, when possible, to identify a specific link or Autonomous System causing a certain packet loss period, as well as relate the observation to the anticipated fault handling.

The paper is organized as follows: the next section briefly revisits some papers related to this project's objectives and the measurement methodology. Section 3 presents the measurement data and explains how they were used to make a decision about the routing of the packets before and after a downtime. Section 4 describes the variety of paths between the end-points. Sections 5 focuses on deeper analysis of observed downtimes and investigate causes of the failures. Conclusions are drawn in Section 6.

II. BACKGROUND AND RELATED WORK

A. Measurement Methodology

Probe packets are sent between the measurement servers in Norway and China every 10 ms in both directions. Both servers have clocks synchronized using the NTP protocol. They are also located close to backbone networks with relative short distance to Tier 1 clock sources. Hence potential impairment introduced by access networks on packet loss and delay is minimized, and the impact on the end-to-end performance caused by routing events in the backbone can be investigated. The programs used to generate and receive packet streams are

[1] "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" appointed by The Research Council of Norway, funded by the Research Council, NTNU and UNINETT. http://www.q2s.ntnu.no

[2] UNINETT is a group of companies which supplies network services for universities, university colleges and research institutions in Norway and handles other national ICT tasks. The Group is owned by the Norwegian Ministry of Education and Research. http://www.uninett.no

978-1-4244-7286-4/10/$26.00 ©2010 IEEE

rude/crude [4]. Each packet is given a sequence number and timestamped at the sender as well as the receiver side. Hence packet delay and loss duration can be obtained. To help detect changes of routes, time-to-live (TTL) of the probe packets are also collected. As a reference, traceroute measurements are performed every two minutes. More details about the measurement setup and initial results of this study including downtime data collected in October-December 2009, observed delay-loss patterns analysis and network availability modeling, are presented in [1].

Similar to [1], only downtime events exceeding 50 ms are analyzed in this paper. It is assumed that shorter loss periods can effectively be compensated on the receiver side. Also, 50 ms is regarded as the largest permissible time for interruptions from fault handling in SDH networks [2, 3].

B. Inter- and intradomain routing

A number of studies investigate relationships between network failures, routing behavior and end-to-end performance. Network downtimes may be caused by events inside an Autonomous System (AS), i.e. intradomain, or between ASes (including failure of an AS), i.e. interdomain. The length of a downtime will depend on the failure detection mechanism(s) available, the fault handling protocol (e.g. rerouting), and in some rare cases when the network gets disconnected, the fault rectification time(s). Interior Gateway Protocols (IGPs) such as OSPF [5] and IS-IS [6] are commonly used in the intradomain networks to reroute traffic after a failure. Studies show that the feasible convergence time for intradomain rerouting (time required by all routers in an AS to go back to steady state operation after a change in the network state) can potentially be less than one second [7, 8].

In addition to intradomain failures, routes between neighboring ASes may also experience disruptions. The Border Gateway Protocol (BGP) recovers communications in this case, and the process may take tens of seconds or even minutes [9, 10, 11]. Although a number of enhancements to speed up convergence have been proposed, for instance, [12, 13] and correlation between routing events and decreased end-to-end performance can be established [14, 15], causes and impacts of both inter- and intradomain routing events on end-to-end path performance still need to be better understood.

Although network measurements are used by researchers to investigate the Internet behavior, deep understanding of network failures is still partially limited by the lack of continuously collected measurement data. Intradomain, analysis of failures affecting connectivity is usually done by analyzing routing table updates for IS-IS [18, 19] and OSPF [20, 21, 22] within a certain AS. An important observation from [18] is that simultaneous failure events, e.g. multiple links experience downtimes at the same time, are common in intradomain networks. Mainly due to common causes, as much as 30% of the downtimes in the Sprint backbone network affect several links.

In 1997 Paxton [23] analyzed communications between multiple ASes. Repeated traceroutes between multiple Internet locations were applied to investigate routing behavior, symmetry and stability. Periods between traceroutes varied from a few hours to a few days. Measured network availability was around 99.5%, which is significantly lower compared to what we have observed for today's backbones. Other papers [14, 26] use active probe packets to study factors affecting end-to-end network performance. The approach adopted in this paper combines these two measurement methodologies and obtains finer granularity in the results by applying increased sampling rates. An alternative approach to gain insights into interdomain routing dynamics is adopted by the RouteViews [24] and the RIPE [25] projects, where passive measurements are used to collect BGP data from multiple routers. However, this approach yields less detailed insight into the impact on the end to end service.

Using active measurement this paper investigates communication failures and causes of the downtimes in the Global Research Network over a three month period. Continuous measurement experiments lasting for such a long period of time and having such a fine “granularity” (packets between the end-points are sent every 10 ms) are quite rare. This enables us to detect and measure not only outages between the two end-points in order of seconds, minutes or more, but also to detect impairments with sub-second durations. Thus we may simultaneously investigate causes and effects of intradomain impairments and also correlation and relationships between inter- and intradomain downtimes.

We monitor and investigate the real-time behavior of the network used for research purposes by universities and scientific organizations. This academic network is lightly loaded, so most route changes may be recognized by change in delay. Furthermore, the number of detected downtimes during the measurement period is moderate. Thus we can analyze every single loss event individually and get better understanding of routing processes taking place in the network. Substantial measurement data are collected in the study. It is shown below how the information can be used to understand and explain causes of communication failures between the end-points.

III. NETWORK DOWNTIME CHARACTERISTICS

Because routing policies may not be symmetric, the measurements are performed in both directions. Measurement results collected in the direction from Norway to China are used in this paper for further detailed analysis. Downtime periods exceeding 50 ms and detected from mid November 2009 till mid February 2010 are summarized in Table 1.

TABLE I. OVERVIEW OF OBSERVATIONS

Downtime duration	Number of events
50 ms – 100 ms	23
100 ms – 1 s	18
1 s – 10 s	27
10 s – 100 s	11
More than 100 s	2
Total # of events	81
Total duration	1062 seconds
Path availability (%)	99.986%

The results are stable (approximately equal total downtime durations every month) and consistent with the earlier results described in [1]. Because the Global Research Network is lightly loaded, queuing effects are minimal most of the time and packet losses caused by congestion are quite rare. Overall end-to-end communications availability between the measurement systems is high and close to 99.99%. What can be done to increase the availability of the end-to-end communications even more to reach 5-nines? All the detected downtimes are investigated below in a more detail.

Fig.1 shows a downtime frequency curve [16] for the measurement data demonstrating end-to-end communications availability as a function of downtimes exceeding a certain threshold. T is downtime duration in seconds (log-scale); Availability is P(W≥T) (reversed log-scale), i.e. the network availability given downtime events exceeding a certain threshold T. The Figure confirms known conclusions that downtimes in the order of tens of seconds or minutes still have the most significant effect on the availability. But, the relatively large number of loss periods in the sub-second range may affect the quality offered by real-time services.

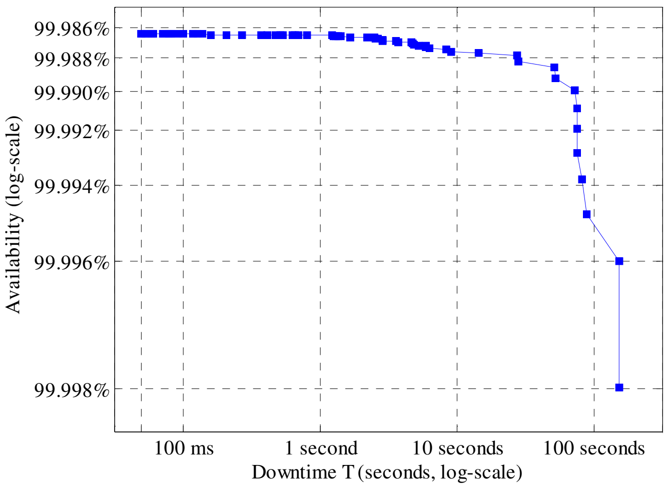

Figure 1. Downtime-frequency curves built based on the measurement results

Three sources of information are used in this paper to make a conclusion about a cause of a certain downtime:

(1) For each detected loss event, average packet delay before and after a downtime period is measured: 200 ms (20 packets) before and after each event were included in the analysis. Change in the network delay was computed as the difference between the two average delays. A delay change statistically different from zero indicates that a new route was chosen to deal with an interruption. Fig. 2a shows an example of downtime followed by a 100-millisecond change in the fixed part of the packet delay. Because the network is not congested, jitter (delay variation) in the network is very low enabling small changes in packet delay in the order of 1 ms to be detected (Fig. 2b). Hence, the measurement approach allow high-accuracy monitoring of changes in the end-to-end path.

Of course, not all loss events are followed by change in the packet delay. Many cases with no change in delay after a downtime were detected. Also, certain failures may be caused by temporary congestion where delay variation may be significant. In cases where congestion causes loss, continuous increase in packet delay is seen followed by a series of packet drops. Finally, a continuous decrease in delay is observed. Packet loss caused by congestion was observed when analyzing the measurement data. Only 4 observations during the whole 3-month measurement period were made. More details and deeper analysis of typical observed delay-loss patterns can also be found in [1].

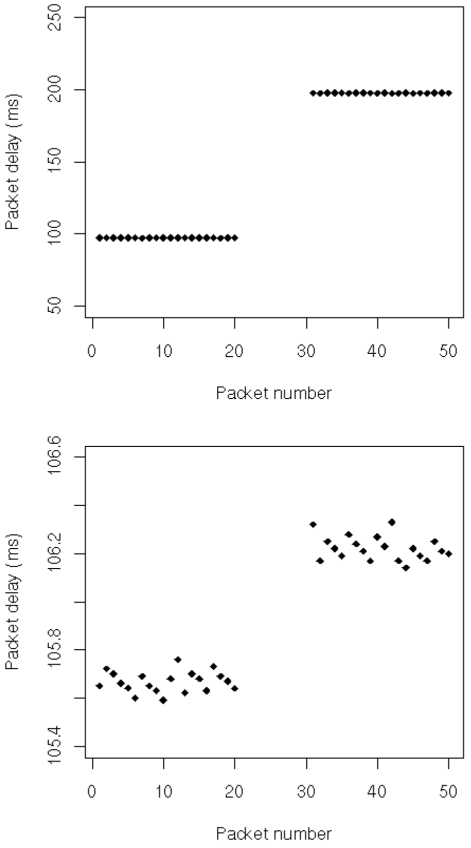

Figure 2 (a,b). Network downtime followed by the change in the fixed part of packet delays

(2) Traceroutes are collected every two minutes between the end-points. Based on this information, we can analyze the variety of end-to-end paths and potential reasons of failures between the end-points. The two-minute interval is too long to get information about routes just before and immediately after each downtime, but the collected statistics is still helpful and used in the analysis. Because congestion in the network is usually not significant, traceroute data collected during the long period of time enables us to associate a certain end-to-end path between Norway and China with a certain delay. Even if a route change after a certain loss period was not detected by the traceroute (because the route has changed again to the initial state a few seconds after the downtime), measured packet delays after the failure may be matched with delay of paths found by traceroute in general and hence indicate which path was used for traffic rerouting.

(3) Each IP-packet header contains a field called time-to-live (TTL). The TTL field is reduced by one by every host routing a packet towards its destination. If this field reaches

zero before the packet arrives at its destination, the packet is discarded. The purpose of the field is to avoid a situation in which undeliverable packets keep circulating in the network. To capture all traffic received by certain ports of the measurement systems *tcpdump* is run on both servers. The TTL counter from each packet is extracted. Changes in the TTL field after a certain loss period proves that a new route is used to deal with the failure. The opposite is not true. If the end-to-end path for a packet has changed, it does not mean that TTL should also change. Lengths of the failed and operative paths can be equal. However, in many cases, by combining together all the collected statistics, we may explain the processes that follow communication downtimes.

IV. End-to-end paths Analysis

Traceroute files collected in the direction from Norway to China were processed and visualized. Based on more than 60 thousand observations, Fig. 3 shows a graph with all network elements observed between the measurement points. The bold lines in the figure show the two paths used most frequently - 99.9% of time. The rest of the links were used very rarely. For instance, the second route between the GEANT and the TEIN3 networks was observed in about 500 traces, i.e., in less than 0.01% of the total experiment duration. Other links in the NORDUnet network were seen in 20-25 traces. Links in the UNINETT network were seen in just 5 traces while links in the Internet2 network (through the US) only two times.

The end-to-end delay on the most frequently used paths was around 100 ms. Increases in end-to-end delay up to 400 ms were measured when different paths were used. And due to the light loading of the network and the dominance of the propagation delay, by examining end-to-end packet delay it was often possible to interpret which paths were in used even without traceroute taken at certain moments of time.

V. Investigating Causes of Downtimes in the Global Research Network

A. "Long" downtimes

Fig. 4 shows all detected downtimes exceeding 10 seconds (13 events) and two events between 8 and 10 seconds ranked from long to short. The reason for these selections as "long downtimes" will be explained below. The longest two measured events had durations close to 160 seconds. Six downtimes were around 80 seconds. Using the packet delay statistics, traceroutes and TTL data, an attempt is made to better understand causes of the failures.

Based on data investigations, the downtimes in Fig. 4 can be divided into two groups. All the longest downtimes in the approximate range from 70 to 160 seconds were followed by change in the fixed part of packet delay. The change was in the range from 4 ms up to approximately 200 ms. This indicates that a new route was chosen to deal with the failure. Closer investigation of traceroute files confirms that the events are caused by interdomain rerouting. Fig. 5 demonstrates an example. A 76 second downtime was caused by BGP tables update because the most frequently used link between the UNINETT and NORDUnet networks became unavailable due to maintenance activity in NORDUnet. TTL counter did not change in this case because the end-to-end path length remained the same.

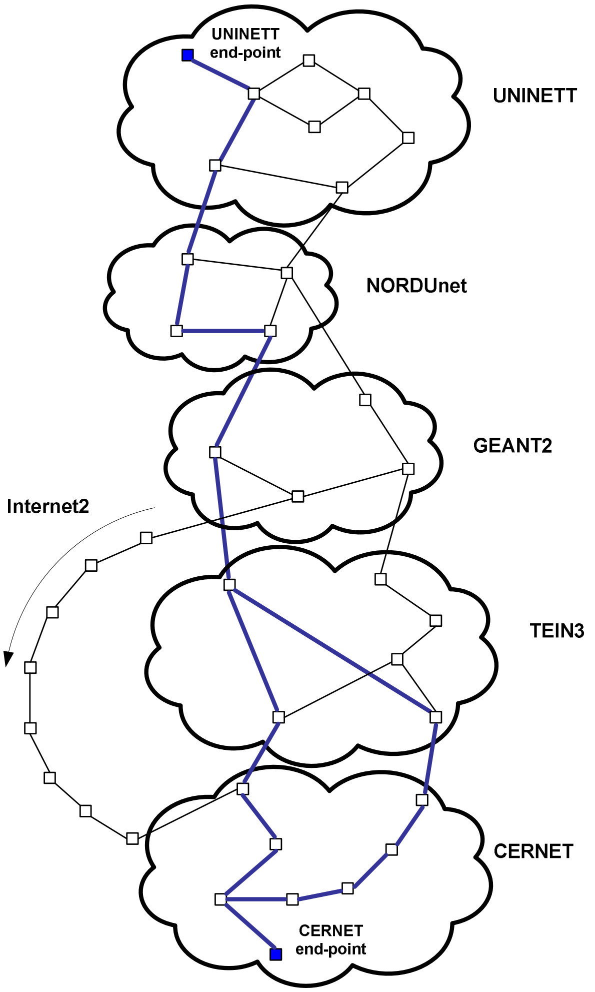

Figure 3. Routing topology

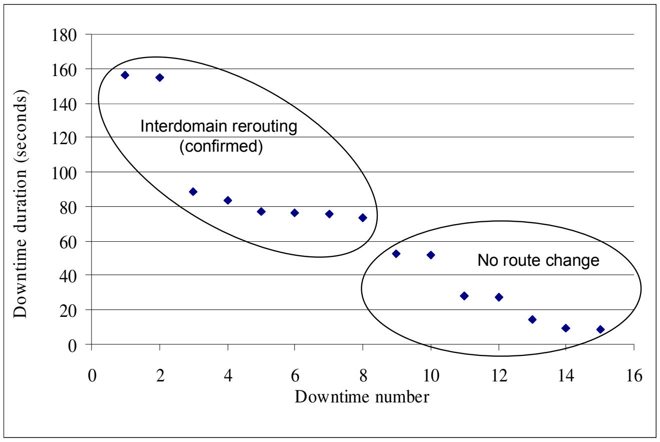

Figure 4. Downtimes exceding 8 seconds (ranked from high to low)

The shorter "no route change" downtimes shown in Fig. 4 were not followed by change in the fixed part of network delay and the end-to-end path of packets before and after the loss periods remained the same. These observations deviate from the anticipated routing behavior since the down times are quite long, up to about 50 seconds, and it is not obvious why an alternative path was not chosen to reroute packets around the

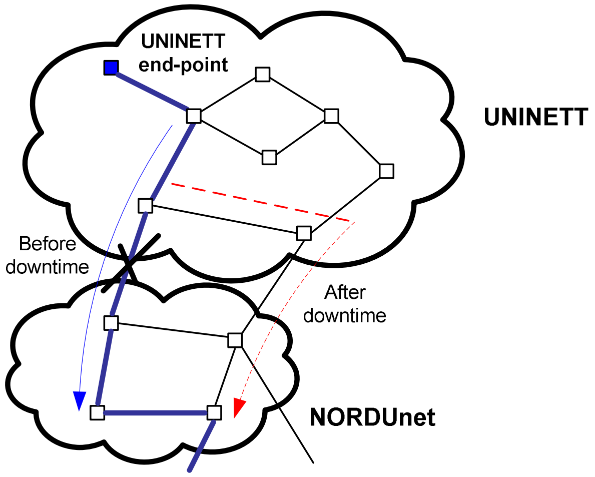

Figure 5. Example of interdomain rerouting

failed element. Tentative explanations are: a) Investigations of BGP convergence by Chang [17] show that a significant number of BGP routing updates indicate temporary path changes, but they ultimately converge on a path that is identical from the previously used path. b) Simultaneous failures were observed. The latter alternative is regarded as most likely since in 6 cases out of the 7 cases, the downtimes were preceded by shorter loss periods, as shown in Fig. 6. Also, 3 such events happened the same day and were only a few minutes apart; two of the other down times also occurred during one day. Indirectly this indicates a possibility of simultaneous link failures in a certain network segment.

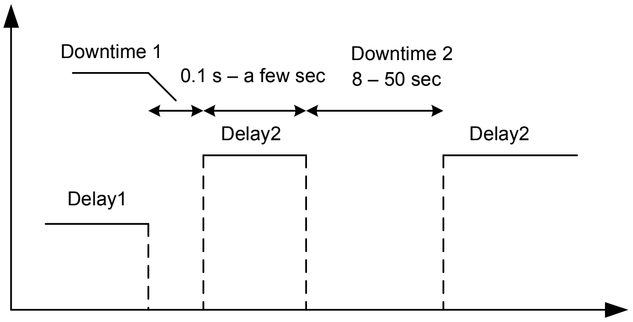

Figure 6. Concecutive downtimes

Fig. 7 demonstrates an example of such a scenario. Before the first failure, packets are sent along Route 1. Then, a failure occurs causing an observed service downtime slightly higher then one second while the new path, Route 2, is selected. Next, after a few hundreds milliseconds, a new failure with long down time occurs taking down route 2 for 52 seconds causing a similar service downtime. After the correction of the latter failure, packets were again sent along the same route, i.e., Route 2. The network topology shows that one more path, denoted Internet2, was available through the United Stated and Japan to China. The Internet2 route was not used in the given case, maybe because it had a convergence time longer than 52 seconds. As already mentioned, the Internet2 route was observed used only a few times during the 3-month period.

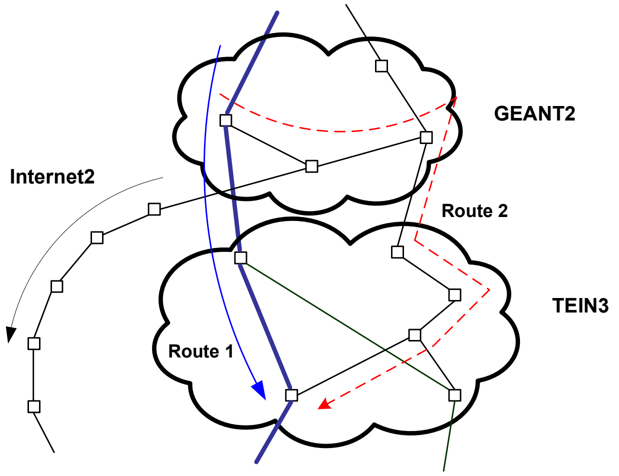

Figure 7. No path change after a downtime

B. *"Medium" downtimes*

All downtimes in this group are characterized by change in the fixed part of network delay indicating that a new path was chosen to reroute traffic around the failed network element. Durations of all these events exceeded 1 second, except for two caused by temporary congestions. The changes in the end-to-end delay were in the range of 4 to 300 ms. Traceroute statistics is available in more than 50% of the cases. The rest of the observed reroutes could not be confirmed by traces because a second reroute event switched back traffic to the initial path after a few seconds. However, the changes in the TTL counter and in the packet delay indicate that a new path was in use.

The collected traceroute data show two typical kinds of packet rerouting. The first type is shown on Fig. 8. It clearly demonstrates that the downtime happened because of intradomain rerouting in the NORDUnet network.

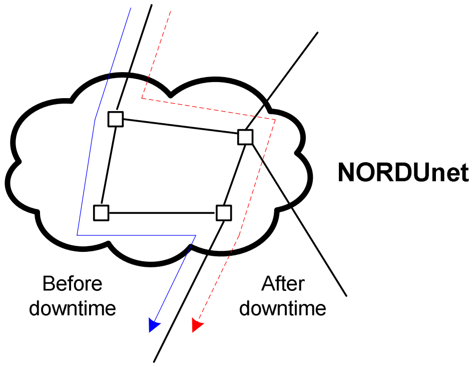

Figure 8. Intradomain rerouting

The second type has already been briefly discussed and shown of Fig. 7. After an approximately one second downtime period, the route between the Geant2 and TEIN3 networks has changed. The new route went through the same networks, but different edge routers. Routing behavior on Fig. 7 is quite similar to that on Fig. 5 and it is not evident from the collected measurement data why in the first case failure durations were around a few seconds, and in the second case it was 76 seconds. However, in the second case, information about BGP updates was available. Evidently the downtimes were handled by different routing protocols.

C. *"Short" downtimes*

About 40% of downtimes in the sub-second range were followed by change in the fixed part of packet delay and are similar to those discussed in the previous section. The same end-to-end path was used before and after failure events for the remaining 60% of the loss events. These events may have been temporary link-level failures, where the transmission over the link is restored before rerouting is initiated, e.g. due to the Dead timer for OSPF [5] and Hold time for IS-IS [6]. Hence the outcome depends on the router configuration. For instance the default value for IS-IS Hold timer is 1 second [6], whether the link failure is explicitly notified from layer 2 network management, e.g. from SDH transport, or not. The longest service downtime without a subsequent new path observed in the short downtimes category was 270 ms.

VI. Conclusion

This paper presented results of active network measurements collected during a three-month period between Norway and China. Communications between the end-points went through the Global Research Network consisting of several Research Networks like GEANT2, TEIN3 and others. The data included packet delay, loss, periodic traceroutes and number of hops in a used path (TTL). Because the network was lightly loaded, service downtimes were consistently due to network element failures and network miss-operation of various kinds rather than congestion. Having end-to-end delays dominated by propagation times made it feasible to investigate in detail causes of every downtime event.

End-to-end network availability is mainly affected by long service downtimes (exceeding 10 seconds). The longest of them is close to 160 seconds. It was observed that these events were followed by change in a routing path, and investigation of traceroute files confirmed that the events are caused by interdomain rerouting. Some of the long service downtimes did not result in change of the end-to-end path. This may be due to simultaneous link failures, operational activities in the network, (poor) configuration of the routing systems or miss-operation, or most likely a combination of these. In order to increase the end to end service availability further toward five nines, it is commonly recognized that the interdomain rerouting times should be shortened. However our investigation shows that we also have a number of "irregular events" which cause considerable downtimes due to lack of action, i.e., no rerouting. Removal of such cases presents a considerable potential for improvement in availability.

The rest of the observations are inline with what is expected. A substantial fraction of downtimes up to 270 ms was not followed by change in traffic route most likely due to a temporary link failure. In the remaining cases, new paths were chosen typically by intradomain reroutes.

VII. References

[1] E. Myakotnykh, B. Helvik, O. J. Wittner , O. Kvittem, J. K. Hellan, T. Skjesol, A. Øslebø, "An empirical analysis of dependability characteristics on a global route", International Workshop on Quality of Services (IEEE IWQoS 2010), Beijing, China, June 16-18, 2010.

[2] ITV-T Rec. G.841, "Types and Characteristics of SDH Network Protection Architectures," 1996.

[3] P. Francois, C. Filsfils, J. Evans, and O. Bonaventure, "Achieving sub-second IGP convergence in large IP networks," ACM SIGCOMM Comput. Commun. Rev., vol. 35, no. 3, pp. 35–44, 2005.

[4] "Rude/crude real-time udp data emitter/collector," http://rude.sourceforge.net/

[5] J. Moy, "OSPF: Anatomy of an Internet Routing Protocol", Addison-Wesley, 1998.

[6] H. Gredler , W. Goralski, "The Complete IS-IS Routing Protocol", SpringerVerlag, 2004.

[7] P. Francois, C. Filsfils, J. Evans, and O. Bonaventure, "Achieving sub-second IGP convergence in large IP networks," ACM SIGCOMM Comput. Commun. Rev., vol. 35, no. 3, pp. 35–44, 2005.

[8] A. Shaikh and A. Greenberg, "OSPF monitoring: Architecture, design and deployment experience," in Proc. USENIX 1st Symp. Networked Systems Design and Implementation (NSDI '04), San Francisco, CA, Mar. 2004, pp. 57–70.

[9] C. Labovitz and A. Ahuja, "The impact of Internet policy and topology on delayed routing convergence," in Proc. IEEE INFOCOM, Anchorage, AK, Apr. 2001, vol. 1, pp. 537–546.

[10] F. Wang, J. Qiu, L. Gao, J. Wang, "On Understanding Transient Interdomain Routing Failures", IEEE/ACM Trans. on Networking, Vol. 17, No. 3, June 2009.

[11] A. Sahoo, K. Kant, and P. Mohapatra, "Characterization of BGP recovery under Large-scale Failures," in Proc. ICC 2006, Istanbul, Turkey, Jun. 11–15, 2006.

[12] D. Pei, B. Zhang, D. Massey, L. Zhang, "An analysis of convergence delay in path vector routing protocols," Computer Networks, vol. 30, no. 3, Feb. 2006, pp. 398–421.

[13] D. Pei, M. Azuma, N. Nguyen, J. Chen, D. Massey, and L. Zhang, "BGP-RCN: Improving BGP convergence through root cause notification," Comput. Networks, vol. 48, no. 2, pp. 175–194, Jun. 2005.

[14] F. Wang, Z. M. Mao, J. Wang, L. Gao, and R. Bush, "A measurement study on the impact of routing events on end-to-end Internet path performance," in Proc. ACM SIGCOMM, Pisa, Italy, 2006, pp. 375–386.

[15] S. Agarwal, C.-N. Chuan, S. Bhattacharyya, C. Diot, "The Impact of BGP Dynamics on Intra-Domain Traffic", In Proceedings of ACM SIGMETRICS, New York, USA, June 2004.

[16] J. Kilpi, I. Norros, and U. Pulkkinen, "Downtime-frequency curves for availability characterization", in The 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN2007), Edinburgh, U.K., Jun. 2007.

[17] Chang, D. F., Govindan, R., Heidemann, J., "The temporal and topological characteristics of BGP path changes", In Proceedings of IEEE ICNP, November 2003.

[18] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C. Chuah, and C. Diot, "Characterization of failures in an operational IP backbone network", IEEE/ACM Transactions on Networking, Vol. 16, No. 4, pp. 749-762, Aug. 2008.

[19] A. Alaettinoglou and S. Casner, "Detailed analysis of ISIS Routing Protocol on the Qwest backbone," NANOG [Online]. Available: http://www.nanog.org/mtg-0202/ppt/cengiz.pdf

[20] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental study of Internet stability and wide-area network failures," in Proc. FTCS, Jun. 1999.

[21] D.Watson, F. Jahanian, and C. Labovitz, "Experiences with monitoring OSPF on a regional service provider network," IEEE ICDCS, May 2003.

[22] A. Shaikh, C. Isett, A. Greenberg, M. Roughan, J.Gottlieb, "A case study in OSPF behavior in a large enterprise network", in ACM SIGCOMM Internet Measurement Workshop, Marseille, France, November 2002.

[23] V. Paxson, "End-to-end routing behaviour in the Internet", IEEE/ACM Transactions on Networking, 1997, 5, (5), 601–615.

[24] U. of Oregon. RouteViews Routing Table Archive. Available from: http://www.routeviews.org/.

[25] RIPE. Routing Information Service Project. Available from: http://www.ripe.net/.

[26] F. Wang, N. Feamster, L. Gao, "Measuring the contributions of routing dynamics to prolonged end-to-end Internet path failures", in GLOBECOM, Los Alamitos, November 2007.