Deliverable FI3-D1.3.1
Analysis of dynamic path diversity tools

Hannu Flinck/NSN

Tivit Future Internet Program
(Tivit FI)

Period: 1.4.2011 – 30.9.2011

Tivit, Strategisen huippuosaamisen keskittymän tutkimusohjelma

Rahoituspäätös 1171/10, 30.12.2010, Dnro 2790/31/2010

www.futureinternet.fi

www.tivit.fi

This work was supported by TEKES as part of the Future Internet programme of TIVIT (Finnish Strategic Centre for Science, Technology and Innovation in the field of ICT).

Executive summary

Title: Analysis of dynamic path diversity tools

Content: This documents complements access selection studies of the multi-access use case surveying multi-path routing mechanism. Multi-path routing offers efficient use of available resources and it makes network overload situations more manageable. It can be considered as part of operator traffic engineering mechanisms. In this deliverable both intra- and inter-domain mechanisms are covered. Implications to WiFi access to 3GPP Enhanced Packet Core are also concluded.

Contact info: Hannu Flinck

1 Introduction

Multi-path routing refers to concepts where multiple parallel paths to same destinations are used at the same times. With it more efficient to use available network resources could be accomplished and network overload situations would come more manageable. Multi-path routing reduces the extent of the link capacity upgrade by almost a factor of two with respect to min-cost routing – which confirms it to be an appealing strategy for both current and future network architectures. In the long term multi-path routing capabilities should be integral part of the routing/control plane of the Internet.

In the packet-level load balancing, load balancing is performed for every single packet and the path may change from packet to packet. It is meant for fault-tolerance against link failures and to optimize transmission bandwidth. No state information from the routing protocols is required. Typical use case is the one where multiple parallel links are used between the same pair of routers. Only the current status of the next hop links suffices for packet scheduling.

Instead in multi-path routing the topology and link status information from the routing protocols are used for path selection. It can be applied at different levels of the protocol stack: session level, the IP level, at the MPLS or at the pseudowire levels. It can be deployed in intradomain or between domains (interdomain). Segments of a multiple path may overlap, may be edge disjoined or node disjointed depending how isolated the parallel paths are required to be. Multi-path routing is required to perceive flow level consistency, i.e. same flow should use same path to avoid any packet level reordering issues.

Currently multi-path routing is used for load balancing, improving resiliency, improving the convergence times of routing protocols as well as improving the resource usage within a network (e.g. to minimize that maximum utility of the network) mainly within intradomain settings. There are BGP extensions for multipath support but those are rarely used.

In this report we look at the key concepts of providing multi-path support mostly at the IP – layer. Multi-path routing extends the next hop selection mechanisms of the single shortest path routing schemes.

2 Benefits of Multipath routing

Site multi-homing is typically used for improving network resiliency and availability. While this remains primary objective for multi-homing, path diversity can be used for improving service experience and network utilization. In [1] the authors report that multi-homing with two upstream ISPs can improve performance at least 25% (measured in content download times) and the transfer speeds improve yet another 20% when site is multi-homing with 3 or more ISPs. They also find evidence of diminishing returns after more than four upstream providers are used for multi-homing. Thus, multi-homing can improve wide-area network performance and lower bandwidth costs. However, to collect these gains the upstream ISPs need to chosen properly. Failing to choosing the right set of providers cloud result into a performance penalty.

In multipath routing traffic is routed dynamically through multiple paths to a destination. This has the benefits of circumventing congested links that improves the end user perceived response time as well as providing resiliency for connectivity. Scheduling traffic between multiple paths through a network leads into more efficient usage of network resources. It has the potential for dynamic load balancing without a route change or need for a special load balancer devices. Instead, traffic is divided into different paths adaptively based on the network conditions. The concept is applicable both for intradomain and as well as interdomain routing.

A central part of any multi-path routing scheme is the path selection mechanism among the multiple candidates. ECMP [2] uses equal cost paths and schedules flows statelessly between them. However, more sophisticated methods may use congestion feedback, probing of latencies and path tagging to select the paths. Path selection can also take into account relative loads on candidate paths. This easily leads to an optimization of some form of a utility function. The main challenge is the stability of the paths as in any load based routing solutions. Static multi-paths are used to remedy oscillations and to reduce reachability restoration times. Typical locations to deploy multi-path is at the customer-edge (intradomain case) and at AS level in peering points (interdomain case).

3 Mathematical formulation of the problem

Path selection problem is typically formulated as a multi-commodity flow optimization problem with convex objectives and linear constraints. The objective is to find the amount of traffic each for flow to minimize the maximum link utilization in the network [10]. Different schemes use different utility functions. Utility functions take into account whether the flow rates are determined independently over each other (uncoordinated case) or the rates are over the paths are functions of all paths (coordinated case).

4 Intradomain methods

OSPF that dominates the intradomain settings provides equal cost multi-path routing (ECMP). ECMP maintains flow coherency by routing packets belonging to a same flow through same next-hop. ECMP path selection is implemented by applying a hash over a 5 tuple consisting of the source and destination IP address, protocol type, TCP/UDP source port, and TCP/UDP destination port. There are several alternatives for the applied hash function (e.g. Modulo-N, Highest random weight) with different trade-offs [4]. Main design criteria involve selection of regions of the hash key results for the next-hops, obtaining the hash key and comparing the key to the regions to decide which next-hop to use.

The shortest path first (SPF) algorithm of OSPF computes all equal cost paths between the nodes of a network. ECMP algorithm needs only to know the number of equal cost paths that should be used for its scheduling. Typical and recommended link costs for SPF is to use inverse of the link bandwidth as the link weight. Clearly this approach is limited to static link weights and cannot take into account dynamically varying link loads. In [3] a method for optimizing OPSF link weights based on projected demands is developed. The authors find that in a practical use case their scheme is very close (with a few percent deviation) to the optimal case. The use of the scheme can provide even 50 – 110% increase of sustained traffic demand compared to static weight ECMP case.

Flow based path selection may not result into even traffic distribution as the traffic matrix is independent from paths and maybe dominated by certain flows. Dividing destination prefixes among available next hops provides a very course and unpredictable load split. To achieve even traffic load balancing with a hash based approach the number of different flows should be large. Especially very short prefixes are problematic. Quite often of traffic is destined to a single prefix leading to that some paths (next hops) will be favored and some would be underutilized.

Second issue with the hash based load balancing is a potential disruption (e.g. packet re-odering) of on-going sessions when new next hops are added to the FIB (for example a new interface is added to the router) as this changes the mapping of hash keys to the next hops [2]. The disruption caused by rearranging the hash key mappings is measured by the fraction of total flows whose path changes in response to some change in the router [4]. This can become problematic if one or more of the paths is flapping. Another concern with ECMP is that forwarding traffic to all possible paths is not optimal if the paths are not disjoint. If the paths overlap there is a likelihood that some links will be congested. To avoid this only a subset of the parallel paths should be used and the path selection should take into account congestion notifications.

There have been proposals for adjusting the ECMP hash boundaries based on the link loads to make path selection more responsive to traffic load. OSPF-OMP (as well as ISIS-OMP, MPLS-OMP) is one such scheme [5]. The load adjustment is achieved by changing how the hash keys map to the next hops by changing the hash key boundaries. The initial adjustment of the boundaries is small but increases exponential until the optimal traffic load is achieved. Once the traffic load decreases the increment is halved to reverse the adjustment to even distribution of the boundaries. Traffic load information is flooded within an OSPF area by using Opaque LSAs (e.g. LSA_OMP_LINK_LOAD). Forwarding is implemented as in ECMP except load is split unequally over the next hops. OSPF-OMP never reached RFC-level in the IETF. However, OSPF-TE [7] addresses some of same issues as OSPF-OMP.

OSPF-TE that is used for MPLS and GMPLS networks can convey more information about the topology and capacity of the network route and path selection purposes. It defines a set of Traffic Engineering LSAs to be used for traffic engineering purposes (e.g. for OSPF weight optimization):

- Maximum Bandwidth
- Maximum Reservable Bandwidth
- Unreserved Bandwidth
- Traffic Engineering Metric

However, OSPF-TE is suitable only for long term traffic loads because it requires recalculation of the shortest paths that ISPs tend to avoid. Rerunning of OSPF with new weights may cause transient loops and congestion. Therefore schemes that are more suitable to dynamic traffic changes have been developed.

Capability of hosts to influence hash based path selection is almost non-existent beyond selecting outgoing interface, the source address and ports for a session. This is because next hop selection is done by the routers inside the network by applying hash functions that are meant to randomize the next hop selection across all the flows. This randomization voids any attempt to select a particular path. Host's capacity to influence path selection of multiple routers in a chain is even more harder. Nonetheless, the authors of [6] propose a method for

path selection by hosts. The idea is based on an invertible hash function that is applied over the source and destination ports while the rest of the fields are used for an injective hash function. With this approach the hosts could select port values so that the combined hash (XOR of the two hashes) can be made specific of a path, resulting into a 32 bit path ID. The next hop selectors of routers are concatenated and keyed by the TTL value of the packet. Each router on the path uses different portion of the path ID based on the TTL- value. Naturally this implies changes in the host protocol stack as well as in the routers, making the proposal impractical.

Allowing (some part of) traffic to route over slightly longer paths can improve the overall through put of the network. Especially this is useful if there are congested links that should be bypassed. A set of tunneling based approaches have been developed to give better controllability of the traffic flows. Controllability of tunnels is attractive to network operators as evidenced by many TE-solutions that has adopted tunneling based approach. Where ECMP represents hop-by-hop based path selection, tunneling based schemes work at segments or at complete paths making traffic engineering easier. Depending on how the segments of tunnels are form together better performance can be achieved with the cost of tunnel creation logic. Experimentation shows that that even few multi-paths can achieve same or even better performance than ECMP [9]. The key elements of tunneling based schemes are an agent at the ingress and egress routers, load balancer function in the ingress that shifts traffic to less loaded links based on certain utility function and a closed-loop feedback that provides information for the dynamic path selection decision. Path stability is ensured by requiring that the feedback loop must work on shorter time scales than the load balancing function. The schemes differ on how the path selection and load balancing is done. Depending on the used algorithms different networking parameters are used for the decision making. Some schemes try to optimize a utility function that takes into account all paths and some make decisions based on traffic rates independently over each path. Also the way how path performance is monitored differ from passive monitoring to active probing of the available paths.

Tunneling based multi-path solutions may work within a domain or across domains. This depends on what kind of visibility path selection algorithm requires to the network topology and what kind of feedback is needed from the network elements to the agents. In the next we summarize two tunneling based approaches.

TeXCP [10] is a tunneling based solution for multi-path routing where the agents at the ingress routers react to congestion notifications from the core routers. Traffic is moved from over-utilized to under-utilized paths. The load balancing is devised so that even if the agents react on local information, the system balances the load across the whole network. It automatically prunes additional paths that do not reduce the maximum utilization and prefers shorter paths over longer ones. The load balancing period needs to be at least 5 times longer than the path utilization probing period. TeXCP algorithm has an issue with favoring shorter paths over longer one. It needs 6 parameters.

In another tunnel based approach, called MIRTO [11], traffic allocation to paths is inversely proportional to path cost using water filling procedure of max-min fairness. Best ranked paths are filled first. All ingress points are using a window flow control mechanism of additive increase multiplicative decrease similar to TCP. It requires per path congestion mechanism and fits well with ECN/MPLS congestion marking. Nothing more is required from the core nodes. However, this requirement makes the mechanism limited to intradomain only.

5 Interdomain methods

The precondition of multi-path routing in interdomain setting is that the upstream provider exports multiple routes to the site that wants to use multiple paths, or alternatively the site in question is multi-homed, i.e. has multiple upstream providers that can be used to carry traffic. The first step is to select upstream providers. The needed number of upstream providers depends on cost and reliability of the providers among other things. Once the upstream providers have been selected the configuration tends to be static and is in time scales of months or longer. Load balancing the traffic between the selected upstream providers is a dynamic allocation problem similar to the one in intradomain methods with reaction times of order of minutes. Many studies [e.g. 13, 14, 15] show that 3 to 4 upstream providers are enough to gain full benefit from multi-homing and multi-path routing.

By default BGP distributes only one path that has been selected to be the best according to its route selection process. The default BGP behavior doesn't allow the advertisement of multiple paths for the same address prefix, or Network Layer Reachability Information (NLRI). In fact, a route with the same prefix/NLRI as in a previously advertisement replaces the previously advertized route. An extension to BGP to support for multiple paths for the same address prefix without the new paths implicitly replacing any previous ones is proposed in "add-path" ietf draft [16]. Advertising multiple paths in BGP requires modifications to the protocol information elements and its route selection. The add-path proposal defines a path identifier that is unique and local between the BGP speakers. Each readvertising BGP speaker must generate its own path identifier for a route in addition to the announced prefix. Each prefix with appended path identifier is considered as a unique route independent from other instances of the same prefix also from the route selection point of view. Clearly, BGP needs to maintain per path advertisement state which adds memory and CPU cost with this approach. Add-path proposal has not been implemented in commercial routers due to need to upgrade all BGP speakers in a deploying domain.

Another proposal to introduce multi-path support into BGP is "Distribution of diverse BGP paths" proposal [17]. Here only route reflectors need to be updated. The idea is to create planes of route reflectors so that each route reflector selects the best path according to the plane where it is configured. The best path reflector announces the best path, the second plane announces the second best path and so forth. The applications for using multi-path capable BGP are not limited to multi-path load balancing that is the focus of this deliverable but also includes fast connection restoration with back up paths, BGP control plane churn reduction and local recovery during network failures.

MIRO [18] is a tunneling based proposal to add multi-path support into BGP system of interconnected autonomic systems (AS). Default routes are announced through BGP but those ASs that need alternative paths use bilateral negotiation to asks another to advertise alternate routes. In responding to a request for alternative paths an AS may provide additional paths obtained from another negotiation as new candidates. An AS trying to achieve high performance might query all immediate neighbors and 2-step away neighbors. Another AS trying to avoid an insecure AS might consult a public Internet topology graph and exclude some ASes that do not comply its policies. Tunnels are created tbetween ASs to alternative egress links based on result of the negotiation. The downstream AS provides a unique tunnel identifier (based on link ID or router ID) to the upstream AS, independent of which AS initiated

the negotiation. End user packets are tunneled using IP-in-IP encapsulation. The tunnel remains active until one of AS tears it down or it expires due to inactivity.

CDN networks typically (e.g. Akamai's Sure Route) use multi-path content aware routing that is an overlay on top of the global Internet. An overlay network is similar to tunneled network, but also includes intermediate nodes that make application level routing decision between the alternative tunnels. Naturally there is an issue of relay node placement in the network topology. The advantage of overlay routing is that it can provide a greater number of diverse paths than what BGP policy complaint routing would provide. For many BGP policy complaint default routes there exists alternative indirect paths that can offer better performance. Overlays can improve significantly round trip time (on average 33% shorter) and provide better through put (15% on average) [12]. However, when a site is multi-homed with 3 upstream sites overlay routing offers marginal benefits compared to the case where there it is not applied but instead intelligent route selection is applied. The benefits of overlay routing are based on its ability to find better performing routes (e.g. "shorter") outside of BGP policies and to react and avoid congestion on the selected paths.

Locator Identifier Separation Protocol (LISP) provides a tunnel based approach for separation of IP addresses, Endpoint Identifiers (EIDs) and Routing Locators (RLOCs) [19]. Routing Locators are in the routable Internet and under BGP routing. LISP can be therefore considered a tunneling based load balancing system. LISP tunnel end points communicate over LISP signaling messages to set up the LISP tunnels. The egress tunnel point can provide alternative Routing Locators with different priorities. This facilitates load balancing between the ingress and egress tunnel points. Multi-path support is not explicit objective of LISP but with proper use of the priorities associated with RLOC of the egress tunnel point multiple parallel paths can be established.

Because the tunnels may traverse multiple intermediate domains with varying performance levels, tunneling based solutions can not ensure similar TE support as in the intradomain environment. Interdomain tunneling solutions need to include efficient performance monitoring and probing tools to measure the actual performance.

6 Conclusions

In intradomain environment making the weight settings for OSPF routing more responsive to traffic load is a promising approach to increase the network efficiency beyond the traditional ECMP. This approach could even react to a biased traffic flow distribution if the weights are set accordingly to actual traffic load. However, applying tunneling based approaches manageability of the traffic engineering and multi-pathing is improved in the intra domain settings. A particular traffic flow, traffic from certain sources or destinations can directed to a given multipath route dynamically. Allowing traffic to route over slightly longer paths can improve the overall through put as well as the load balancing of the network. Naturally tunneling based approaches add the transport overhead that should be taken into account. Tunneling based approaches can be applied also in the interdomain environment if they include a proper performance monitoring tools. Naturally this adds to the complexity of these approach.

In the context of multi-access support for WiFi access to mobile network (Enhanced Packet Core) as studied in FI3-D1.2.1, "Study on Access Selection Steering Mechanisms", traffic from

a host using WiFi access is tunneled to the packet core network of the mobile operator through a gateway that is located between the boundary of the packet core. This tunneling fits well with the tunnel based approached studied in this document if the tunnel end points are controlled with multipath routing. If the traffic steering for the multiple paths is not co-located with the WiFi gateways (e.g. Mobile Access Gateway and Local Mobility Anchor) of 3GPP WLAN interworking case, there is an issue with on-path routers. These intermediate routers are not able to identify flows originated by a certain host because the traffic is aggregated between the mobility gateways. The intermediate routers are not able to peak into the tunneled traffic because the tunnels as secured (IPsec and TLS). This limits the per subscription traffic engineering if the multipath support is not part of the mobility gateways (e.g. Mobile Access Gateway and Local Mobility Anchor).

[1] Akella, A., Maggs, B., Seshan, S., Shaikh, A., Sitaraman, R., A Measurement-Based Analysis of Multihoming, SIGCOMM 2003, Karlsruhe, Germany.

[2] Thaler, D., Hopps, C., Multipath Issues in Unicast and Multicast Next-Hop Selection, RFC 2991, Nov 2000

[3] Fortz, B., Thorup, M., Internet Traffic Engineering by Optimizing OSPF weights. INFOCOM 2000.

[4] Hopps, C., Analysis of an Equal-Cost Multi-Path Algorithm, RFC 2992, November 2000.

[5] Villamizar C., OSPF Optimized Multipath (OSPF-OMP), ietf draft 1999. http://tools.ietf.org/html/draft-ietf-ospf-omp-02

[6] Van Der Linden, S., Detal, G.,Bonaventure, O., Revisiting Next-hop Selection in Multipath Networks, 420-421. In Proc of ACM SIGCOMM, 2011

[7] Katz D., Kompella K., Yeung D., Traffic Engineering (TE) Extensions to OSPF Version 2, RFC 3630, September 2003.

[8] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M. and J. McManus, Requirements for Traffic Engineering Over MPLS, RFC 2702, September 1999.

[9] Tam, A., Xi, K., Chao, J., Trimming the Multipath for Efficient Dynamic Routing, Technical report, presented at CoRR, 2011.

[10] Kandula, S., Katabi, D., Davie, B., Charny A., Walking the tightrope: Responsive yet stable traffic engineering, Proc. ACM SIGCOMM, 2005

[11] Muscarello, L., Perino D., Modeling multi-path routing and congestion control under FIFO and Fair Queuing, LCN 2009: 360-363

[12] Akella A., Pang J., Maggs B., Seshan, S., Shaikh, A,. A Comparison of Overlay Routing and Multihoming Route Control ACM SIGCOMM 2004, Portland, OR.

[13] Dhamdhere, A., Dovrolis, C., ISP and Egress Path Selection for Multihomed Networks. INFOCOM 2006

[14] Yong Zhu, Dovrolis C., Ammar M., Combining Multihoming with Overlay Routing, INFOCOM 2007. 26th IEEE International Conference on Computer Communications. IEEE

[15] Akella, A., Maggs, B., Seshan, S., Shaikh, A., On the Performance Benefits of Multihoming Route Control, Networking, IEEE/ACM Transactions on, Feb. 2008, Vol 16, Issue:1 pages 91 – 104

[16] Walton, D., Chen, E., Retana, A., and J. Scudder, Advertisement of Multiple Paths in BGP,draft-ietf-idr-add-paths-06 (work in progress), September 2011.

[17] Raszuk, R., Fernando, R., Patel, K., McPherson D., Kumaki K., draft-ietf-grow-diverse-bgp-path-dist-06 (work in progress), November 17, 2011

[18] Xu, W., Rexford, J., MIRO: multi-path interdomain routing, SIGCOMM '06, Pisa, Itlay, September 11-15, 2006.

[19] Farinacci, D., Fuller, V., Meyer, D., Lewis, D., Locator/ID Separation Protocol (LISP), Internet-Draft, Work in progress, draft-ietf-lisp-22