Distributed adaptive multi-criteria load balancing: analysis and end to end simulation

Sabine Randriamasy
Alcatel - Research and Innovation
line 3: Marcoussis, France
Sabine.Randriamasy@alcatel.fr

Laurent Fournié, Dohy Hong
N2NSOFT
Paris, France
{Laurent.Fournie, Dohy.Hong}@n2nsoft.com

***Abstract*—This summary presents a 2 year joint study (2003-2005) between ALCATEL, a telecommunication equipment and solution vendor and N2NSOFT, a start-up stemming from the research center INRIA. The collaboration topic is the design and evaluation of a new Multi-Path Routing (MPR) algorithm, called Distributed Multi-criteria Load Balancing (DMLB). DMLB automatically spreads traffic flows of a "critical" link on a set of alternative paths that is provided by a load-sensitive, multi-objective path extraction algorithm.**

Keywords- distributed, load-sensitive, multi-path routing, multi-objective routing, load balancing, simulation, network dimensionng, traffc engineering.

I. INTRODUCTION

DMLB is used as an extension of the default routing, which is Shortest Path (SP). Its goal is to absorb local and temporary traffic increases. It is triggered before link congestion occurs, when its load exceeds a "critical" threshold. DMLB is distributed on the network elements involved in intra-domain IP routing and works on-line at the network operation phase. Associated to the Open Shortest Path First (OSPF) link-state routing protocol with its Traffic Engineering extensions [7] it remains low time and space consuming. Using DMLB upon Shortest Path routing improves:

- *network robustness*: in the operation phase, the network can absorb a larger traffic volume or variation while increasing end to end user sensed performances,
- *network dimensioning*: in the planning phase, network deployment costs can be reduced by downsizing link dimensioning, and link capacity upgrade is more flexible.

Starting from an initial proposition of ALCATEL, a first version of DMLB was jointly specified. N2NSOFT performed massive simulations of the consecutive versions of DMLB, with its flow-level hybrid simulation tool called Netscale [6]. DMLB was then progressively enriched according to the simulation results. The challenge is to maintain or improve both the user-sensed performances and network capacity with robust distributed Multi-Path Routing mechanisms. This required thorough simulations on existing network topologies with a realistic number of flows and sessions (several millions).

This summary focuses on the coherency of distributed multi-path routing decisions and the used simulation approach. Details on other aspects such as path selection, associated dimensioning, and more results can be found in [5].

II. PREVIOUS WORK

Traditional Shortest Path (SP) based routing aims at minimizing one single metric or cost (typically the number of hops or sum of corresponding interface usage costs). This however does not fit the traffic to the link loads and ends up in packet losses in case of link congestion. To cope with that issue, adaptive approaches such as Shortest Widest path (SWP) or Widest Shortest Path (WSP) combine both path length and available link bandwidth. SWP outperforms SP at low levels of network load, but, as it uses longer paths thus more resources than necessary, gets worse for highly loaded networks (see [2]), contrary to WSP [1], which is better suited. WSP, well-suited for QoS routing, offers to choose among feasible paths. Yet if used as a default routing, it restricts the choice to the shortest feasible paths and does not avoid using congested links. In addition, shortest path calculation with a link cost made of a scalar combination of metrics of different nature and magnitude such as length and bandwidth may be numerically instable and miss optimal solutions. Last, using by default a load sensitive routing may cause frequent and straightforward path changes and thus traffic oscillations.

Multi-path routing has been investigated for several years. "Equal Cost Multi Path" (ECMP) was proposed with the OSPF standard [12]. ECMP splits the traffic evenly among equal cost paths to a given destination by distributing the packets in a round and robin fashion among them. The drawback is that link loads are not considered and packets arriving out of order can deteriorate TCP performances. Optimized Multi-path (OMP) [4], is a thorough investigation that attempts to better fit the traffic distribution by splitting the traffic unevenly among alternative paths of "comparable costs". Traffic of paths containing a "critically loaded" segment is shifted away to alternative paths containing none. "Adaptive Multi-path" Routing (AMP) [8], uses the same path selection and flow distribution techniques than OMP, but with link state information restricted to a local scope and thus lighter multi-path data structures and signaling. However, in these approaches, alternative paths are not selected w.r.t. their load or

This work was funded in part by the European IS&T project NOBEL

available bandwidth. The latter is only considered afterwards, at the stage of load adjustment, limiting thus the chances for an optimal traffic distribution. Last, few fully distributed solutions with several routers in multi-path mode have been presented [10].

III. OUTLINES OF DMLB

While the default routing is shortest path, DMLB uses a link load limit *ThLoad* (80% to 90%) above which a link becomes "critical" and the additional incoming traffic shared among alternative paths. Multi-path routing is triggered at the origin I of any directed critical link (I,J), and used for all destinations for which J is the next hop.

DMLB involves the available link bandwidth, defined as the minimum available bandwidth on its links, at the path set selection stage. In order to limit the use of additional network resources the path metrics also include: hop count, transit delay and administrative cost. Unlike many approaches, the link cost is a vector and all the Pareto-optimal solutions are extracted and kept until a later scalar path cost function ranks them [3], providing thus the largest possible choice of alternative paths. The path cost depends on the ratio, for each metric, between the path value and the best one observed in the set of solutions, to maintain numerical stability.

A. *Unequal Flow Distribution*

Each alternative path gets a "target traffic share", representing the maximal proportion of traffic flows to be shifted to it and inversely proportional to its cost. Flow shifting is controlled and progressive, to avoid traffic oscillations:

- a hashing function *H* (typically CRC16) is applied on the flow attributes: *IP source, IP destination, source port, destination port, Protocol ID.* The *H* values are then mapped on *M* "flow bins" (100 by default), flow *f* being assigned to bin *H(f) modulo M*;
- the flow bins are assigned to path interfaces independently, by picking them randomly, to ensure equity and avoid always impacting the same flows/users.
- at every shifting iteration, a fixed number of flow bins (typically 5 or 10) are shifted away from the critical path. The amount of these bins assigned to each alternative path is proportional to its pre-computed target traffic share.

The default interval between two shifting iterations is 10 seconds and corresponds to the minimal interval between 2 forwarding updates, for purposes of traffic stability, in most of the implementations of OSPF. Flow shifting is stopped once the overloaded link load goes below *ThLoad*. Routing stays in multi-path mode until the link load goes below a smaller threshold *ThLoadBack* (65 to 75%) and stays there during a certain time. In that case, the shifting process is reversed to progressively go back to the initial mono-path routing.

B. *Coordination of distributed Multi-Path Routing decisions*

Triggering multi-path routing (MPR) in a distributed mode, at several places and dates jeopardizes network stability. Loops may appear when a packet crosses several routers in multi-path mode. In addition, when distributed routing involves available path bandwidth, loops may appear if routing coherency is not carefully managed: routers can have different views of the network state (a highly loaded outgoing link occults any downstream low loaded link) and their path selection can thus differ. Decision coherency is maintained by the mechanisms presented below.

1) *Flow deviation advertisement*

A router that has triggered DMLB and selected a set of alternative paths to an "area-egress" router D becomes a "Load Balancing Advertiser" (LBA). Prior to multi-path forwarding update of the packets, an LBA sends "Flow Deviation Advertisements" (FDA) containing an alternative path, the set of flow bins to be deviated to this path and the LBA ID. Neither packet marking nor "ingress" router information is used to forward the flows. The FDA are sent along the alternative paths to all the downstream routers until router D. The resulting flow bin deviation orders, (see below) are executed by all the downstream routers. They can thus affect flows that have never crossed the advertising LBA but that are mapped into the corresponding bins. Contrary to what one could fear, this "downstream cumulative flow deviation", while satisfying the traffic demand, has a negligible impact on network performances according to simulation results.

2) *Decision scheduling*

The interval timers of all involved routers are synchronized through the standard Network Time Protocol. They fire for link-state measurement and flooding, routing and forwarding update. Moreover, some delay is kept before applying the link-state and flow deviation advertisements to ensure that they have all been received by the concerned routers.

3) *Prioritizing of Multiple Bin Deviation Orders*

MPR decisions are applied to individual flow bins and called here Bin Deviation Orders (BDO). In a router, the FDA are used to update the BDO. Several LBA may want to deviate a same flow bin, see Fig.1. A rule on the BDO attributes that is common to all routers prioritizes the BDO with successively the most recent date of emission, of last alarm in the LBA, the LBA closest to the destination and "smallest" LBA ID.

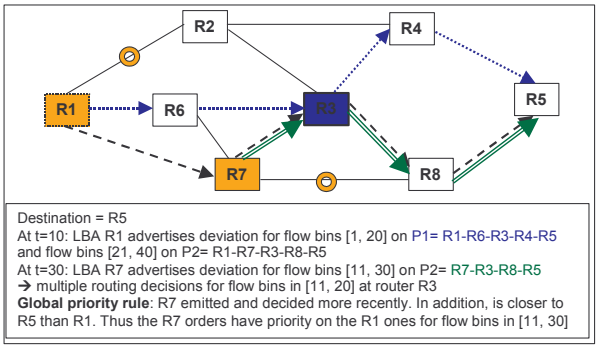

Fig.1 Prioritizing of multiple flow bin deviation orders

4) *2D forwarding based on 3D routing*

At every forwarding update the priority BDO is selected and a flow bin is thus controlled by a unique LBA. Therefore the forwarding decision on a packet only depends on its destination and flow bin ID.

5) Storing all the BDO

The non-priority BDO are all kept in the routing table in case they get the priority in a further iteration. When a link is jointly used by several LBA and one of them stops MPR, keeping track of the other LBA allows them to continue and avoids loops or undesired deviations.

6) Upstream flow deviation adjustment

Several LBAs shifting flows on a common downstream link L may congest it while the router at the source of L may not manage to unload L through MPR. A mechanism to advertise the congestion to the upstream routers is implemented: the alarm is forwarded to a selected set of upstream LBA. This set is chosen so that enough flows (but not too many) are removed from link L. The receiver LBAs then stop adding deviated flows on L and start shifting other ones from L to other links. Note that forwarding the alarm to all the upstream LBAs could lead to an over-reaction or even oscillations.

IV. DMLB EVALUATION

DMLB has been implemented and evaluated on two gigabit backbone networks using Netscale [6], a hybrid simulation tool. To predict TCP behavior together with events such as packet loss or time-outs, faced to routing updates, a massive and fine grain simulation is necessary. Netscale is based on an accurate flow-level dynamic model [9] that allows to track the interaction of hundreds of thousands of TCP flows. Improvements through DMLB have been observed from both the network side (network capacity, packet losses) and the end-to-end user side (goodput, fairness, round trip time).

The following results and demonstration are driven from simulations using the December 2001 topology map of the European research network GEANT [11]. This is a meshed topology with 19 nodes and 60 gigabit links. The simulated traffic ranges from 10 to 85 Gbits/s. This involves up to 700 000 flows and 15 millions of TCP sessions. The simulated period is 1000 seconds and requires about 3 hours of computation. The initial link loads are chosen heterogeneous. We measure the ability of DMLB to absorb a localized traffic demand increase, where half of the uniformly spread users suddenly decide to download files from one country. Results with another network and scenario are presented in [5] and the demonstration also covers a "uniform" traffic increase.

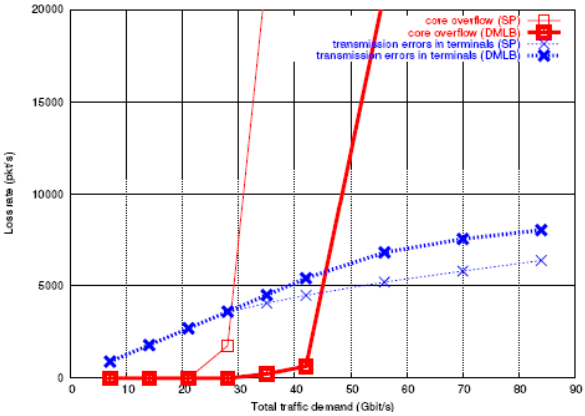

Fig.2 : packet losses at the core and terminal node levels.

Packet loss rate vs. traffic demand is drawn in Fig.2 for congestions in the core network (red squares), and transmission errors in terminals (blues crosses) that serves as a reference for an acceptable level of losses. Thin lines are for shortest path (SP) routing and large lines for DMLB. The first losses with SP routing appear at a demand level of around 28 Gbits/s and grow drastically for larger loads. With DMLB, the network can absorb 53% more traffic (up to 43 Gbit/s) before losses in the core network. Losses due to route changing (e.g. because of packet inversion) are not significant.

V. EXTENSION: DOWNSIZED LINK DIMENSIONING

The ability of DMLB to increase the network capacity can be turned into the possibility of downsizing link bandwidth and thus reconsidering network dimensioning in a more economical way. Whereas traditional SP routing uses link capacity over-provisioning to prevent losses in case of important traffic variations, DMLB deviates traffic bursts on alternate paths and balances traffic on the network. The resources of alternative paths can be jointly dimensioned: they are virtually grouped, as if the associated traffic was multiplexed in a common resource, allowing to reduce the corresponding bandwidth allocation. This complies with many operator strategies who wish to minimize over-provisioning expenses. This new dimensioning approach, presented in [5] , allows:

- *deployment cost reduction*: reduce the network deployment cost for a same or higher quality of service,
- *flexibility*: in case of economical or physical constraints such as bandwidth granularity, a link capacity increase can be distributed on several links.

VI. REFERENCES

[1] IETF RFC 2676 , "QoS routing mechanisms and OSPF Extensions"

[2] K. Kowalik and M. Collier, "Should QoS routing algorithms prefer shortest paths ?", proc. IEEE Int. Conf. on Communications, May 2003, Anchorage, USA, pp 213-217.

[3] X. Gandibleux, F. Beugnies, S. Randriamasy, "Martin's algorithm revisited for multi-objective shortest path problems with a MaxMin cost function", 4OR: A Quarterly Journal of Operations Research, Springer Verlag, Volume 4, Number 1, pp 47-59, March 2006.

[4] C. Villamizar, "OSPF-OMP optimized multipath", IETF draft, draft-ietf-ospf-omp-03, January 2002.

[5] L. Fournié, D. Hong and S. Randriamasy, "Distributed multi-path and multi-objective routing for network operation and dimensioning", 2nd conf. On Next Generation Internet design and engineering, 2006, in press.

[6] NETSCALE, www.n2nsoft.com

[7] IETF RFC 3630, "Traffic engineering (TE) extensions to OSPF Version 2", September 2003.

[8] I. Gojmerac, T. Ziegler, F. Ricciato, P. Reichl, "Adaptive multipath routing for dynamic traffic engineering" , IEEE GLOBECOM 2003.

[9] F. Baccelli, D. Hong, « Flow level simulation of large IP networks », proc. of f INFOCOM 2003, San Francisco.

[10] T. Güven, C. Kommareddy, R.J. La, M.A. Shayman and B. Bhattacharjee, "Measurement based optimal multi-path routing", proc. INFOCOM 2004

[11] GEANT web-site: http://www.geant.net/server/show/nav.128

[12] "*OSPF Version 2*", J. MOY, IETF RFC2328, STD54, April 1998.