Study of a non-intrusive method for measuring the hop-by-hop capacity of a path

Mathieu Goutelle and Pascale Vicat-Blanc/Primet
École Normale Supérieure de Lyon
LIP Laboratory (UMR n°5668 CNRS - ENS Lyon - UCB Lyon - INRIA), RESO team
46, allée d'Italie - 69364 Lyon FRANCE
{Mathieu.Goutelle,Pascale.Primet}@ens-lyon.fr

I. Introduction

According to the design principles of the IP protocol stack, the informations about end to end performances are not provided explicitely by the network layer. Since there is no control channel, complex mechanisms with an high participation of end hosts have to be introduced.

External means using the delay between two hosts to evaluate a *network distance* between two hosts have been proposed [1]. This metric is very easy and quick to measure (`ping`), but insufficient to estimate the time of a data transfer. As interconnected networks are more and more heterogeneous, the hop-by-hop capacity and the available bandwidth evaluation along a path will give a vision of the interconnexion that may be more useful in many cases: for example, network performance evaluation for grid computing optimisations requires such tools.

II. Proposition

We propose a new measurement methodology for discovering the topology characteristics using a packet pair dispersion analysis. This methods is split into a measurement gathering, a bins detection and finally an extraction of the capacity informations.

The method is based on the dispersion of a packet pair, that presents many advantages [2]. But, as cross-traffic taints the dispersion measurements with noise, complex analysis methodologies have to be elaborated [3]. There is two kinds of noise: the first one is typically due to cross-traffic when packets are inserted between the two probes and hence the capacity is underestimated. This noise is random and so relatively wide on small values.

The second type of problems is the modes called PNCM in [3], which can be on the contrary relatively acute and with a value greater than the capacity mode. These modes are obtained when the cross-traffic causes the first probe waits for the second one before being served by the node. The probes are again back-to-back at the node output and their final dispersion is the image of downstream capacities, after the loaded node. We argue that if we are able to gather some informations on the path topology (especially the link capacities), it is possible to eliminate this difficulty.

To know the link capacities of the path, one can use one of the principles of `traceroute`. To be more precise, if a back-to-back packet pair with a TTL value equal to n is sent, the capacity of the hop n, between the sender and the n-th equipment of the path can be evaluated. Then, by doing the same measurement with a TTL value equal to $n+1$ and by assuming the links are symmetric, the only unknown capacity is the link capacity between the nodes n and $n+1$. If this link is not a bottleneck, it can only result in a parasitic mode greater than the capacity already estimated at hop n. It this link is a bottleneck, the previous capacity value will become a parasitic mode. It is fairly easy to identify these two situations: If the distribution doesn't have a relatively acute mode below the already estimated capacity, we are in the first situation: the bottleneck (up to hop $n+1$) has already been passed and is in the previous hop. Otherwise, a mode lower than the previous capacity value is detected and the links between the nodes n and $n+1$ is the new bottleneck. Actually, the addition of a link between each measurement can only, in the worst case (loaded links), create an extra mode in the distribution.

III. Validations

A. Simulation and tests

Simulations to validate the proposition against various situations and to evaluate its accuracy, robustness and the influence of some external characteristics (utilization rate, path length), often not easily controllable in real life have been conducted in the network simulator NS.

The network model choosed for evaluations is the same as the one used in [3] for comparison purpose. The topology is a string, allowing to simulate every path between two nodes and is composed of 7 nodes. A source sends a back-to-back packet pair a thousand times for each value of TTL (from 1 to 6). Cross-traffic is generated on each link in both directions: traffic sources and sinks are set up on each extremity of the links and send traffic with a controllable rate.

1) Accuracy tests of the method: To validate the method for determining the capacity mode, we use the previous simulation to generate a measure batch with a varying utilization rate from 0 to 100% by step of 1%. The analysis module described in the previous section is used to analyse the produced data. We want to prove that our method is reliable and accurate whatever the network conditions (load, path length) may be.

This work is partially granted by the european project DataTAG (IST n°2001-32459) and the INRIA RESO team.

Hop	#1	#2	#3	#4	#5	#6
$u \leq 0.5$	0.1%	0.1%	1.1%	2.5%	4.8%	6.9%
$u \leq 0.75$	0.1%	1.4%	4.6%	7.1%	5.9%	8.3%
$u \leq 1$	0.1%	12.4%	14.9%	15.3%	11.5%	13.7%

TABLE I

ACCURACY TESTS OF THE CAPACITY EVALUATION

Actually, we can even observe that the result is often better for high utilization rate when we are one or two hops further than the bottleneck (*e.g.* hop #4). In these high load situations, the method tends to be conservative. The next steps are yet able to give the expected capacity value once the bottleneck is passed and the last given value is often correct, even for high utilization rate.

2) Robustness tests of the method: This time, we will generate a topology with random characteristics in terms of links capacity and utilization rate. We have done hundred simulations and extracted two informations: first, the squarred correlation factor R^2 between the measure and the capacity of the bottleneck and second, the average relative error (ARE) between these two quantities versus the network load. The results are presented in table II for a short string (the previous one) and for a longer one.

	R^2	ARE	R^2, $u \leq 0.5$	ARE, $u \leq 0.5$
7 hops path	0.58	0.28	0.82	0.14
11 hops path	0.62	0.37	0.88	0.16

TABLE II

ROBUSTNESS TEST OF THE CAPACITY EVALUATION METHOD

First, we can see that the correlation is strong between the measure and the real value. This correlation is much stronger if we restrict the measurements to the one with an utilization rate lower than 50%. We show that the relative error grows logically with the utilization rate. But, it remains low for an utilization rate lower than 50%. The path length seems to have little influence on the result quality.

B. Experimental validation

Simulations give us a validation of the analysis method. The implementation in Linux of the measure module and of the tool `tracerate` in Linux needs validations too, for example to study the influence of OS mechanisms (invisible queue [2], interrupt coalescing, timing accuracy, *etc.*) which can disturb dispersion measurements.

We have done this experiment thanks to the European project DataTAG (`http://www.datatag.org`). We have conducted the following tests to validate `tracerate` in a high-performance environment and to compare it with `pathchar`, the only other tool proposing hop-by-hop measurements, and `pathrate` because it uses the same packet pair technique. The test consists in doing a measurement between a machine at CERN and a machine in Chicago. The path is a 3 hop path with 1 Gbit/s links. `tracerate` gives delay and loss figures too during the measurements but these figures are not given here.

The results are presented in table III. It gives the measured capacity with `tracerate`, `pathchar` and `pathrate` and on different load conditions. All capacities are in Mbit/s. The N/A mention indicates that the measure didn't succeed or that this information isn't available with this particular tool. Values with an asterisk indicates that the tool has given an information message about the validity of this value. The last row gives an estimation of the measurement duration. These results show that the two first tools suffer from incoherent measures (capacity doesn't decrease) due probably to a ICMP rate limitation in the first hops. The fact that these two tools aren't affected exactly the same way comes from the filters configuration.

	`tracerate`			`pathchar`			`pathrate`		
	0%	25%	50%	0%	25%	50%	0%	25%	50%
Hop #1	165	170	165	92	92	93	N/A	N/A	N/A
Hop #2	*162	*165	*162	996	977	832	N/A	N/A	N/A
Hop #3	933	862	862	N/A	N/A	N/A	981-986	760-776	927-947
Duration	2'40	2'40	2'40	N/A	N/A	N/A	25"	5'30	5'40

TABLE III

CAPACITY MEASUREMENTS ON THE DATATAG PLATFORM

On the other way, this experiment shows `tracerate` ability to perform well in a high-performance environment. Nevertheless, the ICMP rate limitation on the first measurement disadvantage a little `tracerate` because, in a normal condition, the given result would be always equal to 933 Mbit/s. Besides, an information message warns the user about this rate limitation problem. Finally, we can remark that `pathchar` doesn't manage to give a result on the last hop, again due to ICMP *Port Unreachable* rate limitation on the Linux receiver.

IV. FUTURE WORK AND CONCLUSION

We have presented an analysis of rate evaluation method, then a proposition of a new method of capacity evaluation and topology discovery between two nodes, using a packet pair technique. We have shown that this method is relatively non-intrusive, robust, relatively accurate and reliable and keep these qualities under bad network conditions (high load, long path, *etc.*). We have show that our tool work up to 1 Gbit/s.

We have validate the Linux implementation and demonstrate that its provides usable results in real life, without the participation of the receiving computer. Many perspectives are still open for this kind of methods: performances evaluation of an end-to-end path or utilization as an OS service or directly in a transport protocol.

REFERENCES

[1] T. E. Ng and H. Zhang, “Predicting Internet Network Distance with Coordinates-Based Approaches,” in *Proceedings of IEEE INFOCOM'02*, New York, NY, USA, June 2002, pp. 170–179.

[2] R. S. Prasad, C. Dovrolis, and B. A. Mah, “The effect of layer-2 store-and-forward devices,” in *Proceedings of INFOCOM '03*, San Fransisco, CA, Apr. 2003.

[3] C. Dovrolis, P. Ramanathan, and D. Moore, “What do packet dispersion techniques measure?” in *Proceedings of INFOCOM'01*, 2001, pp. 905–914. [Online]. Available: citeseer.nj.nec.com/479183.html