УДК 004.75; 004.724.2

G. V. Poryev, V. A. Poryev
National Technical University of Ukraine «Kyiv Polytechnic Institute»
Peremohy ave., 37, 02056 Kiev, Ukraine

Improved CARMA Locality Estimation Model for Peer List Reordering and Its Experimental Validation

The improvement of CARMA network model by addition of locality flavor as well as IPv6 support are concerned. The direct experimental evidence of the improvement of the efficiency of the peer-to-peer networks in terms of the network throughput using previously proposed CARMA locality awareness methods is established. The possible influences of the dynamics of the number of seeding and peering nodes in a swarm (including those directly connected to the test rig) are analyzed and taken into account. Main results indicate average 2,5 % improvement in transfer speed in comparison with clean unbiased transfer modes.

Key words: *Internet, distributed networks, peer-to-peer networks, locality, clustering.*

Introduction

Nowadays it is hard to underestimate the influence of information technologies in general, and the Internet in particular, upon our everyday life and activities. They are reaching so extreme a level of integration and mobility within our life that it could be compared to the importance of electric power. Today, both technologies are so crucial to the progress and the very survival of humankind that virtually no activity is done without either of them.

Analyzing the modern Internet and its traffic-wise composition one might notice some drastic change that passed by almost without notice since the beginning of the 21st century. Prior to that, different classes of Internet traffic were distributed more or less evenly [1, 2] — the majority of traffic consisted of HTTP, FTP, NNTP classes, as well as some minor classes such as DNS, Telnet, IRC etc.

Exponential growth of Internet users that started in about the same period and corresponding growth of the number of Internet-connected devices per average user soon displayed apparent lack of scalability and extensibility of classical «client-server architecture» in many scenarios. The situation had resulted in the new internetworking paradigm, now known as «peer-to-peer» networks, or simply P2P. Today P2P is the most dominant traffic class in every major backbone network according to CISCO VNI (Vir-

tual Networking Index).

Since the emergence of P2P networks, the range, scale and diversity of their use widened significantly, and the field of application of P2P-oriented systems was notably extended, too. Being previously considered as a means for file sharing or instant messaging, today's P2P networks serve as a basic infrastructure for a wide range of innovative application scenarios such as VoIP, multimedia on-demand, software delivery, massive multiuser environments or online games, and even text messaging, where the traditional client-server architectures are still dominating.

The original concept behind P2P systems was to relieve load stress on centralized server farms, mirroring servers and other large data arrays with random access from multiple external nodes. However, the intrinsic asymmetry of end-user broadband data links has caused ISPs to increase maintenance and upgrade cost of the «last mile» hardware in order to keep quality of service steady. Some ISPs had also introduced controversial means to detect and shape forcedly end-user bandwidth for traffic recognized as P2P, affecting file sharing networks in particular. However, these means are nowadays circumvented by protocol obfuscation, encryption and tailoring the connection issuance and acceptance queues, as well as using UDP as underlying protocol for data transport.

For this reason, researchers in the area of P2P systems are aiming to optimize P2P traffic and consider the inherently clustered nature of the Internet as a potential leverage mechanism. The general idea is to maximize network throughput inside the particular network clusters while minimizing the traffic usage between such clusters.

In this paper we have two major goals: a) to improve the previously used CARMA (see below) model by implementing IPv6 support and additional locality flavor into the network segment model and b) to test experimentally selective connectivity approach based on the so-called locality information inferred from CARMA model and determine whether it has any impact on the typical P2P filesharing scenario.

Previous work

Our previous works [3–5] suggested a locally computed approach for topological distance estimation that does not rely on third-party non-guaranteed external infrastructures or unreliable active traffic probing and considered its usage in P2P networks. This approach was previously named CARMA which stands for Combined Affinity Reconnaissance Metric Architecture.

In short, we have devised CARMA as three-layered, with the first layer being the locality awareness expressed in flavors [3], the second layer that uses additional traffic but does not involve actual P2P communication, and the third layer that requires active communication to remote parties over a compatible protocol.

As described in [5], CARMA works by initially preloading structural information from publicly accessible services called Regional Internet Registries (RIRs) and converting it into an internal graphlike data structure. Once loaded, CARMA builds a model to approximate the Internet topology with some simplifications, resulting in 4 common structural layers (CSLs) as follows: a) IPv4 ranges are divided into b) subranges but at the same time they also belong to c) Autonomous Systems (ASs), which are joined into sets called d) AS-SETs or ASSETs.

From these basic entities, CARMA is then able to define and determine the relative topological locality of two arbitrary nodes by sequentially find the lowest CSL.

As an estimation result, CARMA produces flavor index ranging from 0 to 6 according to the lowest found CSL, namely: «Subrange», «Range», «AS», «ASSET», «ASSET-Link», «Backbone» and «Distant». The latter is returned when all tests for the CSLs have failed to produce one.

Improvement of the structural model

As of 2011, the long-anticipated IPv4 address space exhaustion took place officially. In February 2011 the IANA allocated the last /8 (formerly Class-A) IPv4 ranges to the RIRs, whereupon the corresponding local address space exhaustion is anticipated at the level of LIRs (Local Internet Registries) within several months.

The architectural drawback of IPv4 addressing started to become obvious long before the address space was nearing its exhaustion, as soon as the number of Internet-capable devices started to grow exponentially and long after the suboptimal decisions to reserve Class-A networks for unroutable ranges were implemented.

Since then, numerous measures to amend low address availability were designed and gained widespread use, among those are NAT, proxy servers and dynamic IP address pool.

Meanwhile the efforts to create new IP architecture that would overcome such problems were started and eventually resulted in the design and standardization of the IPv6.

Shortly, IPv6 theoretically allows 2^{128} addresses, i.e. 2^{96} times as many as IPv4. Of course, as with IPv4, not every possible address is routable, there are reserved ranges and conventions. Still, the address range capacity is far larger. Even though the recommended practice is to implement IPv6 routing aggregation (i.e. sub-range delegation) as /64, the remaining 64 bits of address theoretically allow every human being on Earth to have more than billion devices connected through his private sub-network.

Structurally IPv6 introduces insignificant changes in the protocol itself, removing unnecessary fields such as header checksum and adding flow labeling and IPSec support.

The reach of connectivity and points of presence of IPv6 are still underdeveloped, so that no more than 1 % of Internet users are IPv6-capable today.

Since CARMA model was conceived prior to IPv4 address exhaustion, we initially did not include any specific support for IPv6 modeling. Now so far as the demand for IPv6 is expected to grow faster, it is reasonable to investigate what might be done to include it in the CARMA model.

Our research into the database structure of IPv6 indicates the following:

— address range delegation as recorded in *delegated* database is done primarily through allocating the /32 segments (i.e. 2^{96} addresses per segment) from RIRs to the LIRs, although various delegation lengths are seen such as /19, /23, /24, /27, /30 and others;

— *db-inetnum* database has its IPv6 equivalent, thereby facilitating sub-range lookup at the typical scale of /48 network;

— *db-route* database has its IPv6 equivalent, thereby allowing the linkage definition between IPv6 ASes and allocated ranges and sub-ranges.

The existence and general similarity of the IPv6 data to that of the IPv4 makes possible to upgrade CARMA network model to support IPv6 with insignificant modifications.

The apparent lack of dedicated IPv6 specific *asset* database is explained by the ASSET linkage being done in IPv4 version for both the IPv6 and the IPv4 ASes.

To sum up, provided all relevant databases are accessible, CARMA implementation is able to build network structural model for IPv6 with the same level of detail as it was done in IPv4.

Our recent investigation on the peering practices and the apparent ambiguity between ASSET membership and actual routing path had led to the necessity to add one additional flavor to the CARMA flavor set.

Since national carriers sometimes are unwilling to participate either in traffic exchange points or establish mutually accountable peering partnerships with other local ISPs, the actual route path may cross several borders even for the address pair belonging to the same country and to the model entities formally registered under exchange point or related ASSET.

For situations like this, we deem it necessary to add one additional flavor of locality between «Backbone» and «Distant» flavors, named «Country», making the total number of possible flavors equal to 8.

This flavor should be returned by CARMA implementation in cases where IP pair test yielded negative results for «Backbone» case, i.e. where there are no intermediate ASSET declared to contain corresponding ASSETs of input IP addresses, but when both of them belong to the IP ranges declared under the same ISO country code in the *delegated* database.

Consequently, if ISO codes are still different, the «Distant» flavor is returned.

This flavor will potentially allow leveraging advantages posed by the topological locality in cases of complicated ASSET linkage and peering conditions but where there are significant differences between routing path lengths and timings in transoceanic links in comparison with local or neighboring country links.

Experimental validation

In the scope of this paper we also aim to focus on the core of CARMA technique, determining the peer query order of neighboring nodes for cluster selection while the second and third layers are intended to fine-tune the first layer. If the first layer works, they all will work, too; otherwise it does not matter significantly.

In our previous paper [4] we have conducted a series of experiments aimed to establish a provable relation between CARMA metric expressed in flavor indices and traditional distance metric still in use today, namely standard *traceroute* method. Although *traceroute* may show up to 20–30 intermediate hop nodes while the previous version of CARMA only has 7 flavors of no literal interpretation in terms of physical links or interfaces, the experiments have conclusively indicated that there is indeed a relation, and by means of mathematical statistics it was estimated as a significant one using ANOVA method, the value of significance was calculated at 8,6 % and 31,4 % for nationally-

centered and international swarm samples, respectively. Though these values may seem low, it should be noted that *traceroute* method is also greatly affected by the network switchovers, bandwidth conditions throughout every tracing path etc.

Setup

Experimental results obtained in [3, 4] are indicative of good correlation between CARMA metric and standard *traceroute* metric. This is important, but only if we were to deploy CARMA as a replacement in an already existing solution requiring network metric. In such scenario it would have been enough to comply with the interfacing specifications, such as taking a remote address as a parameter and giving flavor index as a result. Such an approach would still offer an advantage despite the apparently lower resolution in terms of intermediate hop nodes, because unlike *traceroute* it not only lacks the need to use actively measurement traffic but, most important, are capable of estimating locality between two arbitrary Internet nodes while not running on either of them. Those results, however, are none the less not to be taken as the proof of the overall efficiency of the proposed method, since no direct measurements of either increased throughput or decreased network latency were obtained in the previous works.

There is a more general goal we now envision for CARMA locality estimation methods and its possible future derivatives, namely to drive the optimization of the overlay networks on larger scales, to leverage the inherent nature of P2P networks in that they are more robust, stress-resistant and reliable if consisting of more nodes. This somewhat paradoxical property of P2P has always been the cornerstone of the technology, allowing it to provide economic benefits where other network architectures demand larger expenses or fail altogether.

It is therefore important to demonstrate experimentally that utilizing CARMA mechanisms one can actually improve performance of deployed solutions by an appreciable quantity. In this context, we define performance as suggested in [6] by the throughput relative to the specific communication cost. Since by the nature of P2P and their overlay networks we are implying that the specific communication cost remains constant, it is necessary to demonstrate the increase in the average network throughput in a system employing CARMA mechanisms. It is important, however, to consider the increase in the network throughput not only as simply floating-window average of transfer speed (as this is unlikely to increase during bandwidth saturation) but to view it as a complex parameter depending on, particularly, the apparent acceleration of transfer in terms of time it takes to reach bandwidth saturation.

Design of the experiment

Prior to the actual measurements, some specific conditions pertinent to the setup needed to be clarified.

The experiments should be carried out live, on an already deployed, roughly equally distributed overlay network that uses Internet as the transport infrastructure. This condition is the consequence to the fact that large scale overlay networks with all underlying Internet segments and their constantly changing specifics are very hard to be adequately modeled inside an environment since too many external factors would have been needed to be taken into account. Those include traffic switchovers, bandwidth

variations, noise and good approximation of other traffic classes passing by involved nodes etc.

Among many P2P implementations we have to select those with performance expressed as relevant to the throughput. By this constraint, the file-sharing P2P networks are considered the best fit, since transfer speeds and overhead traffic tolerance are directly influencing both completion time and network latency. Of all file-sharing P2P currently in widespread use, we chose BitTorrent as more likely to demonstrate bandwidth saturation and reasonable plugin-driven extensibility of its most popular client software uTorrent.

Considering the two previous constraints, we may not push our experimental setup to all participating nodes in the BitTorrent overlay. And even if this was somehow possible, we cannot gain any knowledge of present overlay clustering state, including all the circuits, links and groupings, except from those peer nodes directly connected to our experimental rig. Therefore to prove directly the effectiveness of the proposed CARMA method we aim to demonstrate that in the presence of the sufficiently large overlay, deploying CARMA even on a single node in it improves its local performance in terms defined above.

It is necessary to exclude throughput patterns observed during initial phase of the transfer, where not all reported peer nodes from the list are queried and the peer lists are not yet obtained through DHT and PEX mechanisms, if enabled.

Therefore the design of the experiment is as followes:

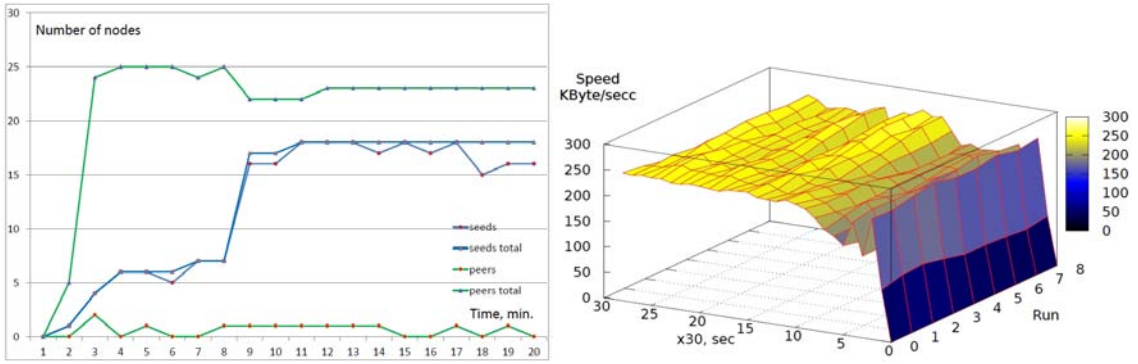
— **Phase A** provides the determination of the repeated pattern of the dynamics of peer and seed node numbers with respect to the channel saturation given that no other traffic is running through the experimental node at the time;

— **Phase B** establishes the observable performance impact of using selective peer query order according to CARMA metric by the first querying the nodes with the closest distance flavor, compared to the normal throughput dynamics without CARMA.

Experiments and data analysis

The experiments were conducted using a single EM64T-powered computer operating Microsoft Windows 7, running x64 version of uTorrent 3.1, connected to the Internet through end-user ADSL-link with declared downlink and uplink speeds of 4000 and 1000 KBits/sec, respectively. To exclude the influence of the link congestion and decrease the probability of choking (in terms of BitTorrent protocol), uTorrent was set up to shape bandwidth to 2500 kBits/sec on the average. This provided for the saturation by the connected peers only. The allocated IPv4 addresses belonged to the Ukrainian Internet Exchange Point UAIX (AS15645), the IPv6 address was obtained through 6to4 tunnel broker operated by NETASSIST.UA, which generally provided for diversity of CARMA flavors observed.

The experiments consisted of 20 runs, each run involved reconnecting to the Internet to obtain different IPv4 and IPv6 addresses and clean-up possible noise traffic from previous experiments; initiating clean download of the same 300 MBytes test file from BitTorrent swarm consisting of 45 nodes on the average. During the first 10 runs Phase A tests were conducted to establish the typical behavior of the numbers of total and of connected peers and seeds (Fig. 1,*a*).



a) Node number dynamics b) Transfer speed

Fig. 1. Averaged dynamics of number of seeds and peers in the experimental swarm

These results indicated that although the number of seeds and peers after stabilization differed only about 20 %, a new node in the swarm is less likely to connect to seeder node because of unequal distribution of «interest» depending on the availability of data, resulting in the seeder node's queue is likely to be overcrowded by regular peers. Note that in this context the node is considered as a «peer» when it has less than 100 % of published content, regardless of the actual percentage. Since our experimental node has no parts of the content at the beginning of the download, for it all other nodes are equal in «interest», therefore the peer list and query order is quickly filled by the peers. This is observed to happen persistently within the first 3 minutes of download. However, the presence of the peer in the query order alone does not guarantee immediate commence of file transfer, and in fact for almost all the duration of the run the number of actually connected pure «peers» remained low.

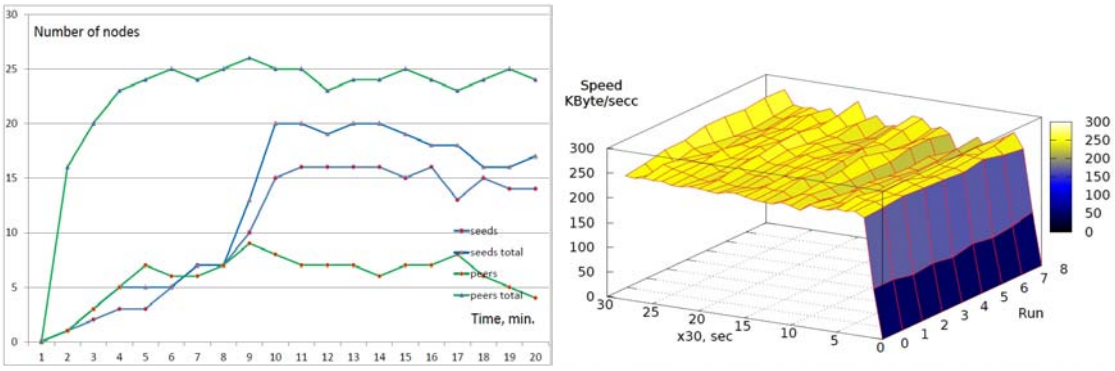
Total and connected seeds numbers were observed to match closely within the duration of run, slowly opening download slots for the first 9 minutes before reaching the preset bandwidth saturation limit, after which the client software decided not to seek additional download sources, remaining at or slightly below 18 connected seeds.

In general, Phase A experiment indicated that due to the nature of overlay network functioning the average download run does not reach bandwidth saturation limit in the first 8 minutes of download, if peer selection process remains random. Fig. 1,*b* shows typical transfer speeds with noticeable gap within the first 7 minutes which is consistent with the behavior of node numbers observed earlier.

For the next 10 runs of Phase B, the behavior of uTorrent was augmented through BTAPPS interface to the external specialized minifirewall in such a way as to block connections to the nodes of higher flavor indices until all nodes of lower indics are queried. This was necessary as BTAPPS interface lacked the ability to establish directly or re-sort the querying order or prohibiting the connection request internally. Fig. 2,*b* shows typical transfer speeds with externally augmented peer query order, whereas Fig. 2,*a* indicates the corresponding node number dynamics.

Apparently, the bandwidth saturation limit is now reached within the first 2 minutes of download session. We assume that this effect was partially leveraged by larger number of non-seeder nodes at the very first minutes of download, as shown in Fig. 2,*a*.

At the speed of 2500 KBits/sec the observed difference of about 40 % in transfer speeds for 6 minutes could lead to potential gain of B_L = (2500*0,4*6*60)/8 = 7500 KBytes which in this demonstrational case was 2,5 % of the total publication size or transfer time decrease.



a) Node number dynamics b) Transfer speed

Fig. 2. Averaged dynamics of number of seeds and peers in the experimental swarm

Conclusion

Taking into account possible discrepancies between formal ASSET linkage registrations and actual routing path we have modified previously developed CARMA model by adding additional «Country» flavor to help facilitate prioritization where ASSET tests fail.

Taking into consideration the IPv4 address space exhaustion we have modified CARMA model construction algorithms and implementations to include IPv6 entities in the network model.

We have demonstrated that using the locality metric such as CARMA as a factor in constructing peer query order in the file-sharing P2P applications could result in performance gain with no additional effort on QoS, hardware modifications or channel reservations along the overlay link path. Although in this case the gain of 2,5 % was relatively low, it may be explained by the singularity of the augmented node among it swarm neighbors.

Strictly speaking, the purpose of the paper was to prove the viability of the concept in the real Internet. We believe that the application potential of such technique is not limited to file-sharing in particular nor in the P2P in general. Any distributed system built in the framework of overlay network using the Internet as a transport layer could ultimately benefit from utilizing CARMA to facilitate automatic overlay clustering, lowering interdomain traffic and decreasing intradomain latency.

1. *Frazer K.D.* NSFNET: a Partnership for High-Speed Networking: Final Report, 1987–1995. — Merit Network, 1996.

2. *Claffy K.* Traffic Characteristics of the T1 NSFNET Backbone / Claffy K., Polyzos G.C., Braun H.W. — In INFOCOM93, 1993. — P. 885–892.

3. *Poryev G.* CARMA-Based MST Approximation for Multicast Provision in P2P Networks / Poryev G., Schloss H., Oechsle R. — In Proceedings of the Sixth International Conf. on Networking and Services (ICNS 2010). — IEEE: Cancun (Mexico). — P. 123–128.

4. *Poryev G.* CARMA: A Distance Estimation Method for Internet Nodes and its Usage in P2P Networks / G. Poryev, H. Schloss, R. Oechsle // Internation. J. on Advances in Networks and Services. — IARIA. — 2010. — Vol. 4, N 3,4. — P. 114–128.

5. *Poryev G.V.* Using the Internet Registries to Construct Structural Model for Locality Estimation in the Overlay Networks / G.V. Poryev // Реєстрація, зберігання і оброб. даних. — 2011. — Т. 13, № 1. — С. 78–86.

6. *Алишев Н.И.* Развитые методы взаимодействия ресурсов в распределенных системах / Н.И. Алишев. — К.: Сталь, 2009. — 448 с.

Received 12.12.2011