Obtaining Performance Measures through Microbenchmarking in a Peer-to-Peer Overlay Computer*

Paolo Bertasi, Mauro Bianco†, Andrea Pietracaprina, and Geppino Pucci
Dept. of Information Engineering,
University of Padova, Padova, Italy
{bertasi,bianco1,capri,geppo}@dei.unipd.it

Abstract

We address the problem of developing a suite of microbenchmarking experiments aimed at providing the basic functionalities of a measurement tool for a P2P-based globally distributed computing platform, usually referred to as Overlay Computer. We argue that such a measuring system should take into account the communication patterns generated by the applications in order to provide useful performance insights.

1. Introduction

The impressive amount of computational resources potentially available through the Internet has stimulated several attempts at making them profitable for vast scale applications, the first and most popular being Peer-to-Peer (P2P) file-sharing applications. The opportunity to exploit idle cycles of connected computers has also led to the development of several world-distributed applications, such as SETI@Home [18], GIMPS [6], and others. While these applications are essentially based upon a simple producer-consumer paradigm, they are a witness of the potential offered by wide-area network distributed computing also for other, more general applications [13]. The parallel computing platform that would actually deploy the computing and storage resources needed for this purpose, and whose network topology is embedded within the Internet is generally referred to as an *Overlay Computer* (OC).

To be successful, an OC must provide higher performance and capabilities than the individual computing equipment available to the user. To this purpose, it is necessary to provide the user with tools for estimating the effectiveness of design choices on such a distributed computer. The tools must rely upon performance metrics aimed at providing a guestimate of the key factors impacting performance of a distributed application, such as interconnection *latency* and *bandwidth*, and available *computing power*.

The approach pursued in this paper is the use of *microbenchmarking techniques* to measure basic performance characteristics of a P2P-based OC. In turn, these techniques can be employed for improving the topology of the OC embedding onto the Internet, which yield high pay-offs in terms of performance. Furthermore, it may be useful to have a system which provides a quantitative assessment of specific OC capabilities, e.g., the ability to efficiently route prominent communications patterns. Finally, performance measurements apply to *application-side optimizations*, e.g., for efficient spawning of a given application over a suitable set of OC peers.

This paper offers two main contributions. The first is to provide an overview on performance measurement techniques for related parallel or distributed platforms such as clusters, the Internet, and P2P systems. The second is to provide a preliminary set of experimental results to serve as *proof-of-concept* for the development of a more extensive microbenchmarking tool for an OC. The proposed tests aim at identifying both measurement techniques and key performance factors that any such tool will have to embody.

The rest of the paper is structured as follows: Section 2 presents the above-mentioned overview. Section 3 presents the tests implemented as a preliminary effort towards a more complete microbenchmarking tool, together with results of performance measurements executed on a local area network running JXTA. Finally, Section 4 draws some conclusions and describes future work.

2. Techniques for Performance Measurement and Improvement: an Overview

This section is dedicated to an overview of techniques used in related systems to measure and/or improve per-

*This work was supported in part by MIUR of Italy under project MAINSTREAM and by the EC/IST Project 15964 AEOLUS.

†Mauro Bianco is currently affiliated at Computer Science Dept., Texas A&M University, College Station, TX, USA.

formance in contexts closely related to overlay computing, such as clusters of workstations, the Internet, and P2P systems.

Overview of performance measurements techniques. Several efforts have been done to provide accurate performance measurement tools for parallel platforms using MPI. The `mpptest` utility, developed in [7], consists of a suite of programs that provide reproducible performance measurements relative to low-level MPI primitives as well as to more complex communication patterns. In [7] the authors also point out the perils of badly designed tests that we tried to avoid. The problem of obtaining accurate measurements without incurring in long testing sessions is addressed in [8] where the *MPIBench* tool is developed to measure performance of MPI-based computations. The idea is based on the use of high precision timers which, however, would require substantial effort to be implemented in a OC.

The availability of high bandwidth over the Internet has made possible the development of several kinds of distributed services, many of them relying on performance metrics to achieve good performance. This has inspired numerous research projects aiming at getting performance parameters out of the Internet, with the double objective of minimizing the testing time while avoiding to overload the network itself. Typically, measures of interest for the Internet are latency and bandwidth, where latency is often referred to as *distance* and depends on the topology of the network at a given instant.

The simplest way to measure the latency is through *ping time*, which, however, may require a long time. Since distances depend only on topology, many works try to characterize the Internet using fictitious coordinate systems whose aim is to allow the estimation of ping time without prior communication, thus avoiding time-consuming probing. Coordinates can be evaluated by measuring distances from some special hosts called *landmarks*, [14, 11], or in a totally distributed fashion, as in [17, 3]. A well known coordinate system for the Internet is Vivaldi [2], which is also employed in some actual P2P implementations to improve the quality of the overlay network [17, 3].

Measuring communication bandwidth appears more difficult than estimating latency, since the bandwidth available between two hosts depends on several (global) aspects, such as the quality of the path connecting the two nodes, the amount of competing traffic along that path, the load at the end-points, etc. Since the load of the network exhibits a certain amount of regularity in its variability, *instantaneous* measures of bandwidth could still produce results which are valid for a certain time interval. In [10] an efficient strategy is proposed to obtain the $O(N^2)$ point-to-point (bottleneck link) bandwidths on an N-node distributed system, exhibiting an overhead linear in the number of nodes and requiring limited cooperation among the nodes. However, the infrastructure to be provided is quite complex, and some of the hypotheses upon which the whole approach relies may be difficult to satisfy in an OC context.

Topology awareness techniques for performance improvement on P2P systems. P2P systems employ several techniques to provide better performance by exploiting topological characteristics of the underlying network. Recent works have pursued a quantitative approach to measure the gains produced by *Topology-awareness*. The work in [3] reports on an extensive experimentation of *DHash++*,which implements a Distributed Hash Table (DHT) based on *Chord* [21]. The authors experiment with several different lookup algorithms and evaluate the resulting performance in terms of lookup time and throughput. Both measures result to be highly affected by locality-awareness of the protocols, which can halve the average time of a request. We remark that only latency is taken into account as a measure of distance, since the data to be retrieved is as little as 8KB in the application under examination.

Earlier works also tried to characterize P2P file sharing application workloads in order to evaluate the improvements that topology-awareness would provide if included in the implementation. One such work is [9], which analyzes the traces of Kazaa (uncached) traffic within the network of the University of Washington. The results suggest that, if content were cached then about 60% of external bandwidth would be saved.

Trace analysis is also employed in [12], in the context of the *PeerMetric* project. The paper suggests that last-hop bandwidth is a major bottleneck and that latency-based optimization of the overlay embedding is somewhat a less impacting issue, especially for bandwidth intensive P2P applications.

One of the first implementations of a topology-aware P2P system is *Pastry* [17]. In Pastry, a node comes with a random ID that identifies the position of the node in the overlay topology. To allow for locality optimization, a node can choose the nearest among k potential neighbors (nodes whose ID is close to the node's ID). The proximity metric chosen by Pastry is the number of Internet hops (an approximation of latency) in the path between two nodes.

The work presented in [16] proposes a strategy to insert topology-awareness into CAN [15] (the idea is however somewhat more general). In the proposed method, the node set is partitioned into *bins*, and a new node wanting to join must find a suitable bin where to fall. Then the node probes a set of predefined unconscious landmarks (a web server, a DNS server, etc.) in order to derive a *bin identifier*, that is, the list of the landmarks sorted by increasing distance. By looking for other nodes with the same bin identifier, perfor-

mance can be improved by placing the node in the best position among the nodes within its bin. The presented results show that topology-awareness can significantly help in improving performance (measured in terms of latency stretch).

GIA [1] implements a strategy to make *unstructured* P2P systems topology-aware. Unlike *structured* P2P systems, unstructured ones allow users to retrieve objects that match partial queries, Gnutella being an example of this approach. GIA adapts itself to the underlying network and to the peer's *capacity*, i.e., the number of queries that a peer can process without being overloaded. Lookup protocols are then designed to guarantee an even utilization of the capacities of the peers involved. Results indicate that topology adaptation improves performance by orders of magnitudes with respect to the basic Gnutella system. From the point of view of this report, it is important to note that this is one of the few papers that also take into account the computational power of the peer, along with the traditional latency and bandwidth parameters.

The idea to organize the network with emphasis on high-capacity nodes is also adopted in [20]. The paper describes a method to build a hierarchical overlay topology where nodes are classified yet again in terms of their capacity. The nodes with highest capacity form a *backbone* structure, while lower capacity nodes refer hierarchically to higher capacity ones. The resulting topology is tree-like, reminiscent of a *fat-tree*. Results indicate that substantial bandwidth savings can be attained with respect to a random topology.

Another aspect impacting the overall performance of a P2P system is *churn*, that is, the rate at which peers join and leave the system, since maintenance overhead increases with churn. The work in [20] includes a description of Bamboo, a DHT that explicitly addresses the problem of routing performance under heavy churn. Results indicate that topology awareness attains a lower latency under high churn.

Few works try to face the problem of topology-awareness for P2P computing (rather than file-sharing) systems, since there are not yet many such platforms. *Zorilla* [4] is one such prototype P2P supercomputing platform. An algorithm similar to the one implemented in Gnutella is used by Zorilla for discovering peers when allocating jobs. The algorithm has been modified to be locality aware in a way similar to the strategy employed in [1]. The result is that the reached nodes are close to the node that initiates the job submission, rather that being randomly distributed as they are in Gnutella. As stated by the authors, this feature speeds-up the initial phase of moving input files from the submitting node to the workers. While it may be reasonable, it looks like a limitation to force the selected peers to be closer to the origin rather than simply to one another (consider, e.g., the case of very long computations not featuring heavy I/O).

3. Design and implementation of initial tests

This section describes a preliminary suite of experiments aimed at identifying the main characteristics that impact performance in a P2P system based on JXTA [5], a P2P API built over JAVA which is becoming popular as a *de facto* standard for P2P platforms development. The tests have been implemented using the *JXTASocket* interface, which provides reliable bi-directional communications. Lower level interfaces (e.g., *Pipes*), although faster, have not been used because of their unreliability. Our objective is to measure key performance quantities at the user level, such as *latency*, *bandwidth*, and *computing power*. The machines employed during the experiments, that are connected through a 100Mbit/s switched Ethernet, are identified as *fast* (>1.5GHz processors) or *slow* (<1GHz processors). This distinction is necessary to evaluate the impact of the software layers on performance.

To measure latency, the core of the experiment is a simple *ping-test*, where a peer sends a small packet (8 bytes) to a selected peer, which then replies as soon as it receives the message. To filter out noise this process is iterated several times. The initiating peer then computes the round-trip time by averaging over the iterations. Since JXTA is based on Java, it is important to quantify the software overhead. For this reason we have measured ping times between fast and slow computers and also compared them against the times obtained using the ICMP protocol. In Table 1 the latency measurements are showed. The table reports two ping times, the first is the JXTA level ping messages, the second time is the time obtained by running the ping command between pairs of machines. It possible to note that the JXTA-level ping is up to three orders of magnitude slower than the ICMP one, because of the software overhead introduced by JXTA, hence the fastest times are obviously those among fast computers. It is also possible to note that the JXTA ping time from slow to fast computers is lower than the one from fast to slow. A similar phenomenon is visible also in the uni-directional bandwidth measurement presented in Figure 1.

Bandwidth is measured through several tests. Each peer involved in the benchmark measures the time for receiving and/or sending the amount of data assigned to it. Measuring communication times requires some attention. Since receives are posted before the actual data arrives, receiving time is measured from the reception of the first bytes till the end of communication, which are clearly identified instants. Since the IP protocol stack in fact does not allow the buffering of a large amount of data before actual transmission, a send can be considered to be *blocking*. Hence, sending times are measured by timing the beginning and the end of the application-level send operations. Communication time is the time a peer spends executing sends and/or receives.

Bandwidth is then computed as the sum of the outgoing and incoming bytes divided by the measured communication time. This bandwidth measure captures both the capability of the peer, the state of congestion of the network, and provides a uniform measure of bandwidth and a clear way to compare results of different communication patterns.

Table 1. Application level vs ICMP ping times

Sender	Receiver	Time (ms)	ICMP (ms)
Fast	Fast	17.2	0.24
Fast	Slow	70.4	0.16
Slow	Fast	57.6	0.14
Slow	Slow	85.5	0.22

When measuring bandwidth, a simple clock alignment algorithm is employed to ensure that peers start the measurement roughly at the same time to stress the network. To perform clock alignment, a selected peer measures latencies from itself to all other peers involved in the test, and then sends them the time they have to wait before starting the experiment. Although quite naive, this simple algorithm has been deemed sufficient in our preliminary experiments to provide a lightweight synchronization of the peers.

The first measurement concerns **point-to-point** bandwidth and is implemented by letting one peer send a given amount of data to another peer and wait for the acknowledgement from the receiver. Several message sizes have been employed to identify the point where the bandwidth saturates. Results are depicted in Figure 1, that shows a high dependency on the underlying architecture. When employing fast computers, the bandwidth is quite close to the peak bandwidth of LAN (i.e., 12.5MB/s) while slow computers dramatically affect communication speed. We note that fast-to-slow communication is faster than slow-to-fast. This phenomenon, also visible in latency measures, will be investigated in future research.

We also measure the execution times of **gather** (i.e., all-to-one) and **scatter** (i.e., one-to-all) communication patterns, to evaluate the OC communication capabilities with respect to typical patterns arising in distributed applications. To measure the execution time of the gather pattern, the alignment algorithm is used to make all peers send the data to the *collector* roughly at the same time. The collector uses a number of receiving threads equal to the number of peers sending data to it. Similar considerations are valid for scatter: a *distributor* sends data to a number of involved peers and measures the elapsed time after the last send terminates. The distributor employs a number of threads equal to the number of receiving peers. All the bandwidth measurements have been carried out with different configurations involving fast and slow machines to provide insights over the software overheads.

Results for scatter are shown in Figures 2, as the size of the sent messages vary. Figure 2.a shows the results of a scatter performed from a slow sender to the others, while Figure 2.b depicts the scatter from a fast sender. Both the graphs exhibit the same trend. The difference in the graphs between the bars of the sender and the receivers is justified by the fact that the sender sends much more data than the data received by the receivers. On the other hand, the bandwidths for scattering 1MB of data are all comparable. This is due to the fact that the packet size is too short to create congestion in the network. Hence, on our testbed platform, any direct measure of bandwidth must require at least 10MB of data. The same conclusion can be reached by looking at the unidirectional bandwidth experiments, noting that the bandwidth saturates when the messages reach a size of 10MB. Clearly these observations are only valid in the context of our toy experiments, where the underlying interconnection is a homogeneous LAN. Different thresholds (but likely similar behaviors) are to be expected when performing the experiments on a wide-area network. Gather measurements involve similar considerations as scatter.

A global measure of bandwidth can be provided by running an **all-to-all** communication pattern, where every peer sends the same amount of data to all the other peers. We remark that applications heavily employing this pattern may prove less attractive for execution on an OC, since, most likely, the availability of large network bandwidth would be required for the OC execution to be competitive. Figure 3 shows the result for an all-to-all pattern among 5 peers. As it can be clearly seen, all individual bandwidths, as well as the average bandwidth, are flattened near the one of the slowest computer. This makes an application using this pattern hardly suitable for execution on an OC.

To measure the computing power of a remote host in an uncooperative P2P setting we use a quiz-like approach [13].

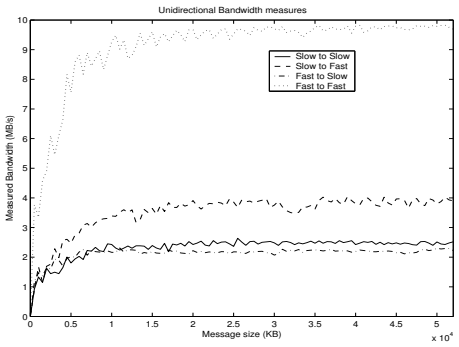
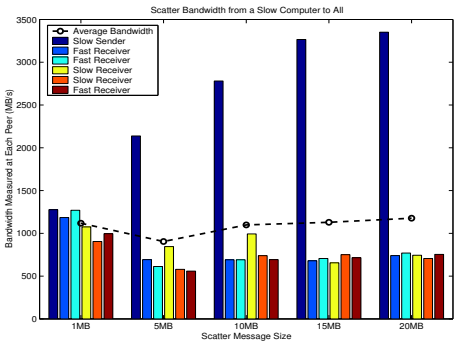
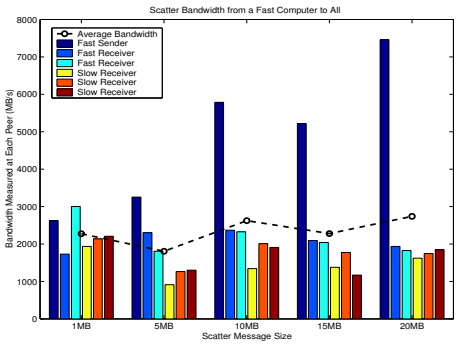


Figure 1. Unidirectional bandwidth exhibits a strong dependency on the computing power of the hosts.



(a)



(b)

Figure 2. Scatter measures from a slow sender to the others (Figure (a)), and from a fast sender to the others (Figure (b)). The plots show measured bandwidth at each peer along with the average value, as the size of the message received by each peer varies.

The remote peer is required to perform a given computation and the execution time is measured by the inquiring peer (that needs to filter out communication time). In order to be effective, the computation needs to exhibit the following characteristics: 1) requiring both a small input and a small output; and 2) requiring a given amount of computation that cannot be avoided. The first requirement is necessary since we do not want the network to become a bottleneck, the second allows for a trusted measure, since the inquired peer cannot employ faster algorithms to provide the right answer. A computation that matches these requirements is given by a random number generator where the input is the seed and the output is the number generated after either a given number of iterations or a given amount of time. A fairly simple generator has been chosen, whose main iteration is `seed=MOD(8121*seed+28411, 134456)`.

Figure 4 shows the results of CPU performance measurements. The plot shows the outcome of four tests: two of them fix the number of iterations to be performed, respectively on a slow and fast CPU; while the other two fix the computing time, respectively on a slow and fast CPU (in this latter case data are again plotted against the number of iterations performed in the allotted time). Even though the measurement is influenced by the fluctuations of the computational load of the inquired peer, the results indicate that such a test may be employed if a sufficiently large number of iterations are executed by the remote peer to deal with the clock resolution. For instance, from the plot we can say that 4 million iterations are sufficient to get a reasonable measure of computing power, which is equivalent to about 300ms of fast CPU time and 800ms of slow CPU time.

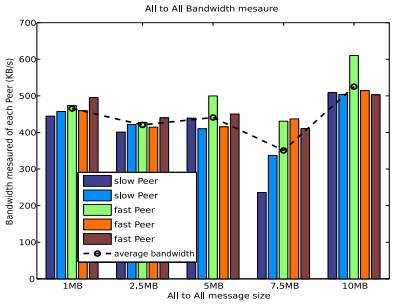


Figure 3. All-to-All communication pattern among 5 peers. As it may be seen, bandwidth is dominated by the slow computer.

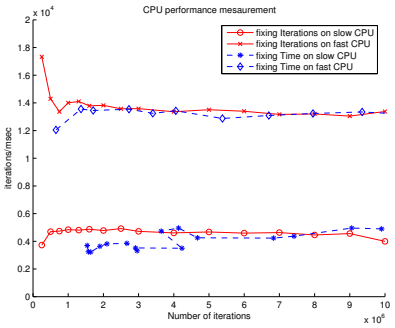


Figure 4. CPU power measurements obtained fixing by the number of iterations and the total execution time on a slow and fast CPU, respectively.

4. Conclusions and future work

In this paper we dealt with the issue of measuring performance in an P2P-based OC, focusing on parameters such

as the computing power of participating hosts, latency, and bandwidth for communication patterns arising in typical applications. After reviewing some relevant state-of-the-art measurement approaches employed in globally distributed system, we developed a suite of microbenchmarking experiments for measuring performance an OC built over JXTA. Preliminary results have shown that different patterns exhibit highly different behaviors, suggesting that the efficient execution of an application requires a careful choice of the executing nodes. Measuring systems should also provide adeguate countermeasures against selfishness and free riders, especially for what concerns estimating the computing power of a given node.

Future work will aim at extending the microbenchmarking suite to produce a complete measurement toolkit to be employed in an OC and to extend the experiments to large heterogeneous testbeds (such as PlanetLab) to fully assess the effectiveness of the proposed approach.

References

[1] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, S. Shenker. Making Gnutella-like P2P Systems Scalable. In *Proc. SIGCOMM*, 407–418, 2003.

[2] F. Dabek, R. Cox, F. Kaashoek, R. Morris. Vivaldi: A decentralized Network Coordinate System. *Proc. SIGCOMM*, 15–26, 2004.

[3] F. Dabek, J. Li, E. Sit, J. Robertson, F. Kaashoek, R. Morris. Designing a DHT for Low Latency and High Throughput. Proc. of the 1st Symposium on Networked System Design and Implementation (NSDI '04), 2004.

[4] N. Drost, R. Nieuwpoort, H. E. Bal. Simple Locality-Aware Co-allocation in Peer-to-Peer Supercomputing. In *Proc. IEEE CCGRID*, 14, 2006

[5] J.D. Gradecki *Mastering JXTA: Building Java Peer-to-Peer Applications*. Wiley, 2002.

[6] GIMPS: Great Internet Mersenne Prime Search. `www.mersenne.org`

[7] W. Gropp, E.L. Lusk. Reproducible Measurements of MPI Performance Characteristics. In *Proc. 6th European PVM/MPI Users' Group Meeting*, LNCS 1697, 11–18, 1999

[8] D. Grove, P. Coddington. Precise MPI performance measurement using MPIBench. In *Proc. HPC Asia*, 2001.

[9] K.P. Gummadi, R.J. Dunn, S. Saroiu, S.D. Gribble, H.M. Levy, J. Zahorjan. Measurements, Modeling, and Analysis of a Peer-to-Peer File-Sharing Workload. In *Proc. 18th ACM SOPS*, 314–329, 2003.

[10] N. Hu, P. Steenkiste. Exploiting Internet Route Sharing for Large Scale Available Bandwidth Estimation. In *Proc. of Internet Measurement Conf.*, 187–192, 2005

[11] M. Kleis, X. Zhou. A Placement Scheme for Peer-to-Peer Networks Based on Principles from Geometry. In *Proc. 4th IEEE Intl. Conf. on P2P Computing*, 134–141, Aug. 2004.

[12] K. Lakshminarayanan, V.N. Padmanabhan. Network Performance of Broadband Hosts. Microsoft Research, Tech. Rep. MSR-TR-2003-15, 2003.

[13] V. Lo, D. Zhou, D. Zappala, Y. Liu, and S. Zhao. *Cluster Computingon the Fly*: P2P Scheduling of Idle Cycles in the Internet. In *Proc. 3rd Int.l Workshop on P2P Systems*, 227–236, 2004.

[14] E. Ng, H. Zhang. Predicting Interne Network Distance with Coordinates-Based Approach. In *Proc. IEEE INFOCOM*, 170–179, 2002.

[15] S. Ratnasamy, P. Francis, M. Handley, R. Karp, S. Shenker. A Scalable Content-Addressable Network. In *Proc. ACM SIGCOMM*, 161–172, 2001.

[16] S. Ratnasamy, M. Handley, R. Karp, S. Shenker. Topologically Aware Overlay Construction and Server Selection. In *Proc. IEEE INFOCOM*, 1190–1199, 2002.

[17] A. Rowstron, P. Druschel. Pastry: Scalable, Distributed Object Location and Routing for Large Scale Peer-to-Peer Systems. In *Proc. IFIP/ACM Int.l Conf. on Distributed System Platforms (Middleware)*, 329–350, 2001.

[18] SETI@Home Project. `setiathome.berkeley.edu`

[19] S. Sodhi, J. Subhlok. Automatic Construction and Evaluation of Performance Skeletons. In *Proc. 19th IEEE IPDPS*, 88–97, 2005.

[20] M. Srivatsa, B. Gedik, L. Liu. Scaling Unstructured Peer-to-Peer Networks With Multi-Tier Capacity-Aware Overlay Topoloies. In *Proc. 10th Int.l Conference on Parallel and Distributed Systems*, 17–24, 2004.

[21] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, H. Balakrishnam. Chord: a Scalable Peer-to-Peer Lookup Service for Internet Applications. In *Proc. ACM SIGCOMM*, 149–160, 2001.