Activity Report 2011

Project-Team CEPAGE

Algorithmics for computationally intensive applications over wide scale distributed platforms

IN COLLABORATION WITH: Laboratoire Bordelais de Recherche en Informatique (LaBRI)

Table of contents

Project-Team CEPAGE

Keywords: Distributed Algorithms, Combinatorial Optimization, Cloud Computing, Information Indexing And Retrieval, Peer-to-Peer, Scheduling

1. Members

Research Scientists

Olivier Beaumont [Team Leader, Research Director, INRIA, HdR]
Lionel Eyraud-Dubois [Research Associate (CR), INRIA]
Nicolas Hanusse [Research Director (DR), CNRS, HdR]
David Ilcinkas [Research Associate (CR), CNRS]
Ralf Klasing [Research Director (DR), CNRS, HdR]
Adrian Kosowski [Research Associate (CR), INRIA]

Faculty Members

Nicolas Bonichon [Assistant Professor (MdC), Université Bordeaux I, on partial INRIA secondment]
Cyril Gavoille [Professor (Pr), Université Bordeaux I, IUF, HdR]
Sofian Maabout [Assistant Professor (Pr), Université Bordeaux IV, on partial INRIA secondment]

External Collaborator

Philippe Duchon [Professor (Pr), Université Bordeaux I, HdR]

Technical Staff

Ludovic Courtès [Engineer INRIA SED (September 2010 – August 2011)]
Lucas Verdonk [Engineer, Euler European grant (October 2011 –)]

PhD Students

Hejer Rejeb [Université Bordeaux I, MENRT grant, (2007 – 2011)]
Emilie Diot [Université Bordeaux I, MENRT grant, (September 2008 – December 2011)]
Florent Foucaud [Université Bordeaux I, MENRT grant, (since September 2009)]
Przemyslaw Uznanski [INRIA / CRA (since October 2010)]
Halftermeyer Pierre [MESR grant (since October 2010)]
Quentin Godfroy [University Bordeaux I, ANR Aladdin grant (since October 2009)]
Christian Glacet [INRIA / FP7 EULER Grant (since November 2010)]
Ahmed Mouhamadou Wade [Université Bordeaux I, MENRT grant (since February 2011)]
Dominik Pajak [INRIA (since October 2011)]

Post-Doctoral Fellows

Hubert Larchevêque [INRIA, USS-Simgrid ANR grant (September 2011 – September 2012)]
Young Won [INRIA ANR USS-SimGrid (Feb. 2010 – Feb. 2011)]

Administrative Assistant

Marie Sanchez [INRIA]

2. Overall Objectives

2.1. General Objectives

The development of interconnection networks has led to the emergence of new types of computing platforms. These platforms are characterized by heterogeneity of both processing and communication resources, geographical dispersion, and instability in terms of the number and performance of participating resources. These characteristics restrict the nature of the applications that can perform well on these platforms. Due to middleware and application deployment times, applications must be long-running and involve large amounts of data; also, only loosely-coupled applications may currently be executed on unstable platforms.

The new algorithmic challenges associated with these platforms have been approached from two different directions. On the one hand, the parallel algorithms community has largely concentrated on the problems associated with heterogeneity and large amounts of data. On the other hand, the distributed systems community has focused on scalability and fault-tolerance issues. The success of file sharing applications demonstrates the capacity of the resulting algorithms to manage huge volumes of data and users on large unstable platforms. Algorithms developed within this context are completely distributed and based on peer-to-peer (P2P for short) communication.

The goal of our project is to establish a link between these two directions, by gathering researchers from the distributed algorithms and data structures, parallel and randomized algorithms communities. Indeed, the change in scale for distributed applications raises several new questions.

The first set of questions is related to distributed computations, where the Internet is the underlying network. Since the topology of the underlying network is unknown, the use of logical networks (overlays) is required. In turn, the choice of the overlay will have an impact on the complexity of the algorithms. In this context, only the performance of the whole chain is meaningful, which requires to collect raw data and then to propose network models and algorithms based on these models, such that the performance of the resulting algorithm is good on raw data. This also requires studying the influence of the topology of the overlay network (the underlying graph) on the complexity of fundamental questions, such as graph exploration or black-hole search.

The second set of questions is related to distributed data structures. In general, the question is related to the compromise between the size of the data structure to be stored on each node and the time to answer a request (estimate the bandwidth between two nodes, compute the closest common ancestor of two nodes in a tree) or to perform a task (route a message in a network based on the information stored at the router nodes).

In order to study these questions, our research plan is based on the following goals. Firstly, we aim both at building strong foundations for distributed algorithms (graph exploration, black-hole search,...) and distributed data structures (routing, efficient query, compact labeling...) to understand how to explore large scale networks in the context of failures and how to disseminate data so as to answer quickly to specific queries. Secondly, we aim at building simple (based on local estimations without centralized knowledge), realistic models to accurately represent resource performance and to build a realistic view of the topology of the network (based on network coordinates, geometric spanners, δ-hyperbolic spaces). Then, we aim at proving that these models are tractable by providing low complexity distributed and randomized approximation algorithms for a set of basic scheduling problems (independent tasks scheduling, broadcasting, data dissemination,...) and associated overlay networks. At last, our goal is to prove the validity of our approach through softwares dedicated to several applications (molecular dynamics simulations, continuous integration) as well as more general tools related to the model we propose (AlNEM for automatic topology discovery, SimGRID for simulations at large scale) and collections of datasets (Hubble for the continuous integration DAGs, Bedibe for latency and bandwidth measurements).

We will concentrate on the design of new services for computationally intensive applications, consisting of mostly independent tasks sharing data, with application to distributed storage, molecular dynamics and distributed continuous integration, that will be described in more details in Section 5.

Most of the research (including ours) currently carried out on these topics relies on a centralized knowledge of the whole (topology and performances) execution platform, whereas recent evolutions in computer networks technology yield a tremendous change in the scale of these networks. The solutions designed for scheduling and managing compact data structures must be adapted to these systems, characterized by a high dynamism of their entities (participants can join and leave at will), a potential instability of the large scale networks (on which concurrent applications are running), and the increasing probability of failure.

P2P systems have achieved stability and fault-tolerance, as witnessed by their wide and intensive usage, by changing the view of the networks: all communication occurs on a logical network (fixed even though resources change over time), thus abstracting the actual performance of the underlying physical network. Nevertheless, disconnecting physical and logical networks leads to low performance and a waste of resources. Moreover, due to their original use (file exchange), those systems are particularly well suited to exact search

using Distributed Hash Tables (DHT's) and are based on fixed regular virtual topologies (Hypercubes, De Bruijn graphs...). In the context of the applications we consider, more complex queries and services will be required (finding the set of edges used for content distribution, finding a set of replicas covering the whole database) and, in order to reach efficiency, unstructured virtual topologies must be considered.

In this context, the main scientific challenges of our project are:

- **Models:**
  - At a low level, to understand the underlying physical topology and to obtain both realistic and instanciable models. This requires expertise in graph theory (all the members of the project) and platform modelling (Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois and Ralf Klasing). The obtained results will be used to focus the algorithms designed in Section 6.1 and Section 6.2.
  - At a higher level, to derive models of the dynamism of targeted platforms, both in terms of participating resources and resource performances (Olivier Beaumont, Philippe Duchon). Our goal is to derive suitable tools to analyze and prove algorithm performances in dynamic conditions rather than to propose stochastic modeling of evolutions (see Section 2.2).

- **Overlays and distributed algorithms:**
  - To understand how to augment the logical topology in order to achieve the good properties of P2P systems. This requires knowledge in P2P systems and small-world networks (Olivier Beaumont, Nicolas Bonichon, Philippe Duchon, Nicolas Hanusse, Cyril Gavoille). The obtained results will be used for developing the algorithms designed in Sections 6.2 and 6.3.
  - To build overlays dedicated to specific applications and services that achieve good performances (Olivier Beaumont, Nicolas Bonichon, Philippe Duchon, Lionel Eyraud-Dubois, Ralf Klasing, Adrian Kosowski). The set of applications and services we target will be described in more details in Section 4.1.
  - To understand how to dynamically adapt scheduling algorithms (in particular collective communication schemes) to changes in network performance and topology, using randomized algorithms (Olivier Beaumont, Nicolas Bonichon, Nicolas Hanusse, Philippe Duchon, Adrian Kosowski, Ralf Klasing) (see Section 6.2).
  - To study the computational power of the mobile agent systems under various assumptions on few classical distributed computing problems (exploration, mapping problem, exploration of the network in spite of harmful hosts. The goal is to enlarge the knowledge on the foundations of mobile agent computing. This will be done by developing new efficient algorithms for mobile agent systems and by proving impossibility results. This will also allow us to compare the different models (David Ilcinkas, Ralf Klasing, Adrian Kosowski, Evangelos Bampas) (see Section 6.2).

- **Compact and distributed data structures:**
  - To understand how to dynamically adapt compact data structures to changes in network performance and topology (Nicolas Hanusse, Cyril Gavoille) (Section 6.3)
  - To design sophisticated labeling schemes in order to answer complex predicates using local labels only (Nicolas Hanusse, Cyril Gavoille) (Section 6.3)

We will detail in Section 5. how the various expertises in the team will be employed for the considered applications.

We therefore tackle several problems related to two priorities that INRIA identified in its strategic plan (2008-2012): "Modeling, Simulation and Optimization of Complex Dynamic Systems" and "Information, Computation and Communication Everywhere "

2.2. Goal and Context

2.2.1. General Context

The recent evolutions in computer networks technology, as well as their diversification, yield a tremendous change in the use of these networks: applications and systems can now be designed at a much larger scale than before. This scaling evolution is dealing with the amount of data, the number of computers, the number of users, and the geographical diversity of these users. This race towards *large scale* computing has two major implications. First, new opportunities are offered to the applications, in particular as far as scientific computing, data bases, and file sharing are concerned. Second, a large number of parallel or distributed algorithms developed for average size systems cannot be run on large scale systems without a significant degradation of their performances. In fact, one must probably relax the constraints that the system should satisfy in order to run at a larger scale. In particular the coherence protocols designed for distributed applications are too demanding in terms of both message and time complexity, and must therefore be adapted for running at a larger scale.

The aggregation of information stored at each node can be seen as a distributed data structure. Whenever an application senda a request in the network, the result corresponds to a combination of individual queries performed by some nodes. It is important, in terms of resource consumption and query time, that the least number of nodes are requested.

Moreover, most distributed systems deployed nowadays are characterized by a high dynamism of their entities (participants can join and leave at will), a potential instability of the large scale networks (on which concurrent applications are running), and an increasing individual probability of failure. Therefore, as the size of the system increases, it becomes necessary that it adapts automatically to the changes of its components, requiring self-organization of the system to deal with the arrival and departure of participants, data, or resources. As it is not reasonnable that each node get an exact knowledge of the network, oracles are promising local and light data structures that sum up different metrics about the network: approximate distance, latencies, ...

As a consequence, it becomes crucial to be able to understand and model the behavior of large scale systems, to efficiently exploit these infrastructures, in particular w.r.t. designing dedicated algorithms handling a large amount of users and/or data.

2.2.2. Limitations of Parallel Processing Solutions

In the case of parallel computation solutions, some strategies have been developed in order to cope with the intrinsic difficulty induced by resource heterogeneity. It has been proved that changing the metric (from makespan minimization to throughput maximization) simplifies most scheduling problems, both for collective communications and parallel processing. This restricts the use of target platforms to simple and regular applications, but due to the time needed to develop and deploy applications on large scale distributed platforms, the risk of failures, the intrinsic dynamism of resources, it is unrealistic to consider tightly coupled applications involving many tight synchronizations. Nevertheless, (1) it is unclear how the current models can be adapted to large scale systems, and (2) the current methodology requires the use of (at least partially) centralized subroutines that cannot be run on large scale systems. In particular, these subroutines assume the ability to gather all the information regarding the network at a single node (topology, resource performance, etc.). This assumption is unrealistic in a general purpose large size platform, in which the nodes are unstable, and whose resource characteristics can vary abruptly over time. Moreover, the proposed solutions for small to average size, stable, and dedicated environments do not satisfy the minimal requirements for self-organization and fault-tolerance, two properties that are unavoidable in a large scale context. Therefore, there is a strong need to design efficient and decentralized algorithms. This requires in particular defining new metrics adapted to large scale dynamic platforms in order to analyze the performance of the proposed algorithms.

2.2.3. Limitations of P2P Strategies

As already noted, P2P file sharing applications have been successfully deployed on large scale dynamic platforms. Nevertheless, since our goal is the design of efficient algorithms in terms of actual performance

and resource consumption, we need to concentrate on specific P2P environments. Indeed, P2P protocols are mostly designed for file sharing applications, and are not optimized for scientific applications, nor are they adapted to sophisticated database applications. This is mainly due to the primitive goal of designing file sharing applications, where anonymity is crucial, exact queries only are used, and all large file communications are made at the IP level.

Unfortunately, the context strongly differs for the applications we consider in our project, and some of the constraints appear to be in contradiction with performance and resource consumption optimization. For instance, in these systems, due to anonymity, the number of neighboring nodes in the overlay network (i.e. the number of IP addresses known to each peer) is kept relatively low, much lower than what the memory constraints on the nodes actually impose. Such a constraint induces longer routes between peers, and is therefore in contradiction with performance. In those systems, with the main exception of the LAND overlay, the overlay network (induced by the connections of each peer) is kept as far as possible separate from the underlying physical network. This property is essential in order to cope with malicious attacks, i.e. to ensure that even if a geographic site is attacked and disconnected from the rest of the network, the overall network will remain connected. Again, since actual communications occur between peers connected in the overlay network, communications between two close nodes (in the physical network) may well involve many wide area messages, and therefore such a constraint is in contradiction with performance optimization. Fortunately, in the case of file sharing applications, only queries are transmitted using the overlay network, and the communication of large files is made at IP level. On the other hand, in the case of more complex communication schemes, such as broadcast or multicast, the communication of large files is done using the overlay network, due to the lack of support, at IP level, for those complex operations. In this case, in order to achieve good results, it is crucial that virtual and physical topologies be as close as possible.

2.3. Highlights

- The members of CEPAGE have been involved in the following program committees: IPDPS 2011 (Vice-Chair, Algorithm Track), PODC 2011 (General Chair), SIROCCO 2011 (Co-chair), EuroPar 2011 (Local Chair, P2P Track), STACS 2011, ESA 2011, FOMC 2011, ADHOC-NOW 2011, IWOCA 2011, IC3 2011, RENPAR 2011, ISCIS 2011, DISC 2011.

- Nicolas Hanusse has become Directeur de Recheche at CNRS. (Ralf Klasing became Directeur de Recherche in 2010 and Cyril Gavoille became junior member of the Institut Universitaire de France in 2009).

- Adrian Kosowski has become a member of the junior chapter of the Polish Academy of Sciences for the term of office 2012–2016.

3. Scientific Foundations

3.1. Modeling Platform Dynamics

Modeling the platform dynamics in a satisfying manner, in order to design and analyze efficient algorithms, is a major challenge. In distributed platforms, the performance of individual nodes (be they computing or communication resources) will fluctuate; in a fully dynamic platform, the set of available nodes will also change over time, and algorithms must take these changes into account if they are to be efficient.

There are basically two ways one can model such evolution: one can use a *stochastic process*, or some kind of *adversary model*.

In a stochastic model, the platform evolution is governed by some specific probability distribution. One obvious advantage of such a model is that it can be simulated and, in many well-studied cases, analyzed in detail. The two main disadvantages are that it can be hard to determine how much of the resulting algorithm performance comes from the specifics of the evolution process, and that estimating how realistic a given model is – none of the current project participants are metrology experts.

In an adversary model, it is assumed that these unpredictable changes are under the control of an adversary whose goal is to interfere with the algorithms efficiency. Major assumptions on the system's behavior can be included in the form of restrictions on what this adversary can do (like maintaining such or such level of connectivity). Such models are typically more general than stochastic models, in that many stochastic models can be seen as a probabilistic specialization of a nondeterministic model (at least for bounded time intervals, and up to negligible probabilities of adopting "forbidden" behaviors).

Since we aim at proving guaranteed performance for our algorithms, we want to concentrate on suitably restricted adversary models. The main challenge in this direction is thus to describe sets of restricted behaviors that both capture realistic situations and make it possible to prove such guarantees.

3.2. Models for Platform Topology and Parameter Estimation

On the other hand, in order to establish complexity and approximation results, we also need to rely on a precise theoretical model of the targeted platforms.

- At a lower level, several models have been proposed to describe interference between several simultaneous communications. In the 1-port model, a node cannot simultaneously send to (and/or receive from) more than one node. Most of the "steady state" scheduling results have been obtained using this model. On the other hand, some authors propose to model incoming and outgoing communication from a node using fictitious incoming and outgoing links, whose bandwidths are fixed. The main advantage of this model, although it might be slightly less accurate, is that it does not require strong synchronization and that many scheduling problems can be expressed as multi-commodity flow problems, for which efficient decentralized algorithms are known. Another important issue is to model the bandwidth actually allocated to each communication when several communications compete for the same long-distance link.

- At a higher level, proving good approximation ratios on general graphs may be too difficult, and it has been observed that actual platforms often exhibit a simple structure. For instance, many real life networks satisfy small-world properties, and it has been proved, for instance, that greedy routing protocols on small world networks achieve good performance. It is therefore of interest to prove that logical (given by the interactions between hosts) and physical platforms (given by the network links) exhibit some structure in order to derive efficient algorithms.

3.3. General Framework for Validation

3.3.1. *Low level modeling of communications*

In the context of large scale dynamic platforms, it is unrealistic to determine precisely the actual topology and the contention of the underlying network at application level. Indeed, existing tools such as Alnem [60] are very much based on quasi-exhaustive determination of interferences, and it takes several days to determine the actual topology of a platform made up of a few tens of nodes. Given the dynamism of the platforms we target, we need to rely on less sophisticated models, whose parameters can be evaluated at runtime.

Therefore, we propose to model each node using a small set of parameters. This is related to the theoretical notion of distance labeling [59], and corresponds to assigning labels to the nodes, so that a cheap operation on the labels of two nodes provides an estimation of the value of a given parameter (the latency or the bandwidth between two nodes, for instance). Several solutions for performance estimation on the Internet are based on this notion, under the terminology of Network Coordinate Systems. Vivaldi [57], IDES [61] and Sequoia [62] are examples of such systems for latency estimation. In the case of bandwidth estimation, fewer solutions have been proposed. We have studied the last-mile model, in which we model each node by an incoming and an outgoing bandwidth and neglect interference that appears at the core of the network (Internet), in order to concentrate on local constraints.

3.3.2. Simulation

Once low level modeling has been obtained, it is crucial to be able to test the proposed algorithms. To do this, we will first rely on simulation rather than direct experimentation. Indeed, in order to be able to compare heuristics, it is necessary to execute those heuristics on the same platform. In particular, all changes in the topology or in the resource performance should occur at the same time during the execution of the different heuristics. In order to be able to replicate the same scenario several times, we need to rely on simulations. Moreover, a metric for providing approximation results in the case of dynamic platforms necessarily requires computing the optimal solution at each time step, which can be done off-line if all traces for the different resources are stored. Using simulation rather than experiments can be justified if the simulator itself has been proven valid. Moreover, the modeling of communications, processing and their interactions may be much more complex in the simulator than in the model used to provide a theoretical approximation ratio, such as in SimGrid. In particular, sophisticated TCP models for bandwidth sharing have been implemented in SimGRID.

During the course of the USS-SimGrid ANR Arpege project, the SimGrid simulation framework has been adapted to large scale environments. Thanks to hierarchical platform description, to simpler and more scalable network models, and to the possibility to distribute the simulation of several nodes, it is now possible to perform simulations of very large platforms (of the order of 10^5 resources). This work will be continued in the ANR SONGS project, which aims at making SimGrid usable for Next Generation Systems (P2P, Grids, Clouds, HPC). In this context, simulation of exascale systems are envisioned, and we plan to develop models for platform dynamicity to allow realistic and reproducible experimentation of our algorithms.

3.3.3. Practical validation and scaling

Finally, we propose several applications that will be described in detail in Section 5. These applications cover a large set of fields (molecular dynamics, continuous integration...). All these applications will be developed and tested with an academic or industrial partner. In all these collaborations, our goal is to prove that the services that we propose in Section 4.1 can be integrated as steering tools in already developed software. Our goal is to assert the practical interest of the services we develop and then to integrate and to distribute them as a library for large scale computing.

At a lower level, in order to validate the models we propose, i.e. make sure that the predictions given by the model are close enough to the actual values, we need realistic datasets of network performance on large scale distributed platforms. Latency measurements are easiest to perform, and several datasets are available to researchers and serve as benchmarks to the community. Bandwidth datasets are more difficult to obtain, because of the measurement cost. As part of the bedibe software (see section 5.4), we have implemented a script to perform such measurements on the Planet-Lab platform [56]. We plan to make these datasets available to the community so that they can be used as benchmarks to compare the different solutions proposed.

3.4. Efficient Queries and Compact Data Structures

The optimization schemes for content distribution processes or for handling standard queries require a good knowledge of the physical topology or performance (latencies, throughput, ...) of the network. Assuming that some rough estimate of the physical topology is given, former theoretical results described in Section 6.2 show how to pre-process the network so that local computations are performed efficiently. Due to the dynamism of large distributed platforms, some requirements on the coding of local data structures and the updating mechanism are needed. This last process is done using the maintenance of light virtual networks, so-called *overlay networks* (see Section 6.2). In our approach, we focus on:

- *Compact Routing tables.*

  Routing queries and broadcasting information on large scale platforms are tasks involving many basic message communications. The maximum performance objective imposes that basic messages are routed along paths of cost as low as possible. On the other hand, local routing decisions must be fast and the algorithms and data structures involved must support a certain amount of dynamism in the platform. Since the size of the data-structure impacts negatively the query and the update time, the space of the data-structure must be of limited size.

- *Local computations.*

  Although the size of the data structures is less constrained in comparison with P2P systems (due to security reasons), however, even in our collaborative framework, it is unrealistic that each node manages a complete view of the platform with the full resource characteristic. Thus, a node has to manage data structures concerning only a fraction of the whole system. In fact, a partial view of the network will be sufficient for many tasks: for instance, in order to compute the distance between two nodes a local and limited information available at the two nodes may suffice (distance labeling).

- *Overlay and small world networks.*

  The processes we consider can be highly dynamic. The preprocessing usually assumed takes polynomial time. Hence, when a new process arrives, it must be dealt with in an *on-line* fashion, i.e., we do not want to totally re-compute, and the (partial) re-computation has to be simple.

  In order to meet these requirements, *overlay networks* are normally implemented. These are light virtual networks, i.e., they are sparse and a local change of the physical network will only lead to a small change of the corresponding virtual network. As a result, small address books are sufficient at each node.

  A specific class of overlay networks are *small-world* networks. These are efficient overlay networks for (greedy) routing tasks assuming that distance requests can be performed easily.

- *Mobile Agent Computing.*

  Mobile Agent Computing has been proposed as a powerful paradigm to study distributed systems. Our purpose is to study the computational power of the mobile agent systems under various assumptions. Indeed, many models exist but little is known about their computational power. One major parameter describing a mobile agent model is the ability of the agents to interact.

  The most natural mobile agent computing problem is the exploration or mapping problem in which one or several mobile agents have to explore or map their environment. The rendezvous problem consists for two agents to meet at some unspecified node of the network. Two other fundamental problems deal with security, which is often the main concern of actual mobile agent systems. The first one consists in exploring the network in spite of harmful hosts that destroy incoming agents. An additional goal in this context is to locate the harmful host(s) to prevent further agent losses. We already mentioned the second problem related to security, which consists for the agents in capturing an intruder.

  The goal is to enlarge the knowledge on the foundations of mobile agent computing. This will be done by developing new efficient algorithms for mobile agent systems and by proving impossibility results. This will also allow to compare the different models.

Of course, the main difficulty is to adapt the maintenance of local data structures to the dynamism of the network.

4. Application Domains

4.1. New Services for Scheduling on Large-Scale Platforms

As mentioned in Section 2.1, solutions provided by the parallel algorithm community are dedicated to stable platforms whose resource performances can be gathered at a single node that is responsible for computing the optimal solution. On the other hand, P2P systems are fully distributed but the set of available queries in these systems is much too poor for computationally intensive applications. Therefore, actual solutions for large scale distributed platforms such as BOINC [1], WCG [2] or XTremWeb mostly rely on a client-server model, where no direct communication between peers is allowed. The objective of our project is to extend the application field that can be executed on large scale distributed platforms, by providing, for example, tools for:

---

[1] http://boinc.berkeley.edu/
[2] http://www.worldcommunitygrid.org/

- Requests and Task scheduling on large scale platforms;
- New services for processing on large scale platforms.

4.2. Routing in the Internet

The localized data-structures developed in the project, in particular the near-shortest path representation of a graph, can be naturally applied to routing in the Internet. More, precisely it can be applied to compact routing information while preserving routing along near-shortest routes. The current solution, based on BGP, uses routing tables of size linear in the number of T1-routers. The increase of the number of these routers, and so of the resulting routing tables, negatively impacts the routing time to forward each packet in the routers.

One of the goals of the project (see also Section 8.3) is to provide routing tables for the Internet network with sublinear size while achieving near-shortest path routing.

4.3. Decision Support Systems

Datacubes are an intuitive interface allowing users to mine their data by selecting subsets of dimensions, drilling down and rolling up through dimensions hierarchies and by constraining some dimensions values. In order to optimize this navigation, the best solution would be to precompute all possible query results. However, this is unfeasible in practice. Hence one has to select the "most beneficial" part to precompute and materialize. Traditionally, this problem has been modeled as a query execution time minimization under storage space hard constraint. We have proposed to revisit this problem by considering query performance as being the hard constraint whilst minimizing the storage space. Due to the hardness of this problem, we proposed approximate solutions. For this purpose, we used, among others, the concept of *border* which is encountered in several data mining applications. We developed a general parallel algorithm for computing such borders with performance guarantees which is quite easily adaptable to the case where the data are distributed. This algorithm has been implemented and the extensive experiments confirmed the theoretical results.

5. Software

5.1. Hubble

Participants: Ludovic Courtès [correspondant], Nicolas Bonichon.

Hubble is implemented in Scheme, using GNU Guile version 2. Details of the simulation, such as keeping track of processor occupation and network usage, are taken care of by SimGrid, a toolkit for the simulation of distributed applications in heterogeneous distributed environments.

The input to Hubble is an XML description of the DAG of build tasks. For each task, a build duration and the size in bytes of the build output are specified. For our evaluation purposes, we collected this data on a production system, the http://hydra.nixos.org/ build farm hosted at the Technical University of Delft. The DAG itself is the snapshot of the Nix Package Collection (Nixpkgs) corresponding to this data. Hubble has its own in-memory representation of the DAG in the form of a purely functional data structure.

The Nixpkgs DAG contains fixed-output nodes, i.e., nodes whose output is known in advance and does not require any computation. These nodes are typically downloads of source code from external web sites. The raw data collected on http://hydra.nixos.org/ specifies a non-zero duration for these nodes, which represents the time it took to perform the download. This duration info is irrelevant in our context, since they don't require any computation, and Hubble views these nodes as instantaneous.

See also the web page http://hubble.gforge.inria.fr/.

Software assessment : A-2, SO-3, SM-2, EM-1, SDL-2.

5.2. namdP2P

Participants: Nicolas Bonichon [correspondant], Olivier Beaumont.

NamdP2P is a distributed implementation of ABF method using NAMD. It is worth noting that NAMD is designed to run on high-end parallel platforms or clusters, but not to run efficiently on instable and distributed platforms.

Software assessment : A-1, SO-2, SM-1, EM-1, SDL-1.

5.3. Malleable minWCT

Participants: Nicolas Bonichon [correspondant], Olivier Beaumont, Lionel Eyraud-Dubois.

This applet considers the scheduling of malleable tasks with bounded amount of processing resources. The goal is to compute schedules that minimize the weighted completion time of tasks. The applet generates all possible greedy schedules for a given instance and displays only the best ones.

This applet illustrates the complexity of finding an optimal order of tasks [31].

See also the web page http://www.labri.fr/perso/bonichon/malleable

Software assessment : A-2, SO-1, SM-2, EM-1, SDL-2.

5.4. Bedibe

Participants: Lionel Eyraud-Dubois [correspondant], Przemyslaw Uznanski.

Bedibe (Benchmarking Distributed Bandwidth Estimation) is a software to compare different models for bandwidth estimation on the Internet, and their associated instantiation algorithms. The goal is to ease the development of new models and algorithms, and the comparison with existing solutions. The development of this software is just starting.

See also the web page http://bedibe.gforge.inria.fr/.

Software assessment : A-1-up2, SO-3, SM-1-up2, EM-2, SDL-1-up2.

5.5. MineWithRounds

Participants: Sofian Maabout [correspondant], Nicolas Hanusse.

This software extracts the Maximal Frequent Itemsets from a transaction data base. It is designed in `C++` using OpenMP Library to take full advantage of multicore, multi-cpu machines.

Software assessment: A:1, SO:4, SM:2, EM:2, SDL:1.

6. New Results

6.1. Models

6.1.1. *Using the Last-mile Model as a Distributed Scheme for Available Bandwidth Prediction*
Participants: Olivier Beaumont, Lionel Eyraud-Dubois, Young Won.

Several Network Coordinate Systems have been proposed to predict unknown network distances between a large number of Internet nodes by using only a small number of measurements. These systems focus on predicting latency, and they are not adapted to the prediction of available bandwidth. But end-to-end path available bandwidth is an important metric for the performance optimisation in many high throughput distributed applications, such as video streaming and le sharing networks. In [34], we propose to perform available bandwidth prediction with the last-mile model, in which each node is characterised by its incoming and outgoing capacities. This model has been used in several theoretical works for distributed applications. We design decentralised heuristics to compute the capacities of each node so as to minimise the prediction error. We show that our algorithms can achieve a competitive accuracy even with asymmetric and erroneous end-to-end measurement datasets. A comparison with existing models (Vivaldi, Sequoia, PathGuru, DMF) is provided. Simulation results also show that our heuristics can provide good quality predictions even when using a very small number of measurements.

6.1.2. *Divisible Load Scheduling*

Participants: Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois.

Malleable tasks are jobs that can be scheduled with preemptions on a varying number of resources. In [31] We focus on the special case of work-preserving malleable tasks, for which the area of the allocated resources does not depend on the allocation and is equal to the sequential processing time. Moreover, we assume that the number of resources allocated to each task at each time instant is bounded. We consider both the clairvoyant and non-clairvoyant cases, and we focus on minimizing the weighted sum of completion times. In the weighted non-clairvoyant case, we propose an approximation algorithm whose ratio (2) is the same as in the unweighted non-clairvoyant case. In the clairvoyant case, we provide a normal form for the schedule of such malleable tasks, and prove that any valid schedule can be turned into this normal form, based only on the completion times of the tasks. We show that in these normal form schedules, the number of preemptions per task is bounded by 3 on average. At last, we analyze the performance of greedy schedules, and prove that optimal schedules are greedy for a special case of homogeneous instances. We conjecture that there exists an optimal greedy schedule for all instances, which would greatly simplify the study of this problem. Finally, we explore the complexity of the problem restricted to homogeneous instances, which is still open despite its very simple expression. (Join work with Loris Marchal from ENS Lyon)

6.1.3. *Modeling and Practical Evaluation of a Service Location Problem in Large Scale Networks*

Participants: Olivier Beaumont, Nicolas Bonichon, Hubert Larchevêque.

In [33], we consider a generalization of a classical optimization problem related to server and replica location problems in networks. More precisely, we suppose that a set of users distributed over a network wish to have access to a particular service proposed by a set of providers. The aim is then to distinguish a set of service providers able to offer a sufficient amount of resources in order to satisfy the requests of the clients. Moreover, a quality of service following some requirements in terms of latencies is desirable. A smart repartition of the servers in the network may also ensure good fault tolerance properties. We model this problem as a variant of Bin Packing, namely Bin Packing under Distance Constraint (BPDC) where the goal is to build a minimal number of bins (i.e. to choose a minimal number of servers) so that (i) each client is associated to exactly one server, (ii) the capacity of the server is large enough to satisfy the requests of its clients and (iii) the distance between two clients associated to the same server is minimized. We prove that this problem is hard to approximate even when using resource augmentation techniques : we compare the number of obtained bins when using polynomial time algorithms allowed to build bins of diameter at most βd_{max}, for $\beta > 1$, to the optimal number of bins of diameter at most d_{max}. On the one hand, we prove that (i) if $\beta = (2 - \epsilon)$, BPDC is hard to approximate within any constant approximation ratio, for any $\epsilon > 0$; and that (ii) BPDC is hard to approximate at a ratio lower than $\frac{3}{2}$ even if resource augmentation is used. On the other hand, if $\beta = 2$, we propose a polynomial time approximation algorithm for BPDC with approximation ratio $\frac{7}{3}$ in the general case. We show how to turn an approximation algorithm for BPDC into an approximation algorithm for the non-uniform capacitated K-center problem and vice-versa. Then, we present a comparison of the quality of results for BPDC in the context of several Internet latency embedding tools such as Sequoia and Vivaldi, using datasets based on PlanetLab latency measurements.

6.1.4. *Use of Internet Embedding Tools for Heterogeneous Resources Aggregation*

Participants: Olivier Beaumont, Nicolas Bonichon, Philippe Duchon, Hubert Larchevêque.

In [28], we are interested in large scale distributed platforms like BOINC, consisting of heterogeneous resources and using the Internet as underlying communication network. In this context, we study a resource clustering problem, where the goal is to build clusters having at least a given capacity and such that any two participants to the same cluster are not too far from each other. In this context, the distance between two participants corresponds to the latency of a communication between them. Our goal is to provide algorithms with provable approximation ratios. In such large scale networks, it is not realistic to assume that the whole latency matrix (that gives the latency between any two participants) is known, and we need to rely on

embedding tools such as Vivaldi or Sequoia. These tools enable to work on compact descriptions and well described metric spaces in which the distance between two points can be obtained directly from a small amount of information available at each node. We present the Bin Covering under Distance Constraint problem (BCDC for short), and propose dedicated algorithms for this problem for each metric space induced by each of the embedding tools. Then, we propose a comparison of these algorithms based on actual latency measures, that enables to decide which algorithm/embedding tool pair offers in practice for realistic datasets the best balancing between distance prediction and approximation ratios for the resource clustering problem.

6.1.5. *Broadcasting on Large Scale Heterogeneous Platforms with Connectivity Artifacts under the Bounded Multi-Port Model*

Participants: Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois, Przemyslaw Uznanski.

In [32], we consider the classical problem of broadcasting a large message at an optimal rate in a large scale distributed network. The main novelty of our approach is that we consider that the set of participating nodes can be split into two parts: "green" nodes that stay in the open-Internet and "red" nodes that lie behind firewalls or NATs. Two red nodes cannot communicate directly, but rather need to use a green node as a gateway for transmitting a message. In this context, we are interested in both maximizing the throughput (i.e. the rate at which nodes receive the message) and minimizing the degree at the participating nodes, i.e. the number of TCP connections they must handle simultaneously. We consider both cyclic and acyclic solutions for the flow graph. In the cyclic case, our main contributions are a closed form formula for the optimal cyclic throughput and the proof that the optimal solution may require arbitrarily large degrees. In the acyclic case, we propose an algorithm to achieve the optimal throughput with low degree. Then, we prove a worst case ratio between the optimal acyclic and cyclic throughput and show through simulations that this ratio is on average very close to 1, which makes acyclic solutions efficient both in terms of throughput and of number of connections.

6.2. Overlays and distributed algorithms

6.2.1. *Locally Fair Graph Exploration Strategies*

Participants: David Ilcinkas, Ralf Klasing, Adrian Kosowski.

In [16], we considered the problem of exploring an anonymous undirected graph using an oblivious robot. The studied exploration strategies are designed so that the next edge in the robot's walk is chosen using only local information, and so that some local equity (fairness) criterion is satisfied for the adjacent undirected edges. Such strategies can be seen as an attempt to derandomize random walks, and are natural counterparts for undirected graphs of the rotor-router model for symmetric directed graphs. The first of the studied strategies, known as Oldest-First (**OF**), always chooses the neighboring edge for which the most time has elapsed since its last traversal. Unlike in the case of symmetric directed graphs, we show that such a strategy in some cases leads to exponential cover time. We then consider another strategy called Least-Used-First (**LUF**) which always uses adjacent edges which have been traversed the smallest number of times. We show that any Least-Used-First exploration covers a graph $G = (V, E)$ of diameter D within time $O(D|E|)$, and in the long run traverses all edges of G with the same frequency.

6.2.2. *Black Hole Search in Directed Graphs*

Participant: Adrian Kosowski.

In [21] we considered a team of agents which has to explore a graph G where some nodes can be harmful. Robots are initially located at the so called home base node. The dangerous nodes are the so called black hole nodes, and once a robot enters in one of them, it is destroyed. The goal is to find a strategy in order to explore G in such a way that the minimum number of robots is wasted. The exploration ends if there is at least one surviving robot which knows all the edges leading to the black holes. As many variations of the problem have been considered so far, the solution and its measure heavily depend on the initial knowledge and the capabilities of the robots. We assume that G is a directed graph, the robots are associated with unique identifiers, they know the number of nodes n of G (or at least an upper bound on n), and they know the number

of edges Δ leading to the black holes. Each node is associated with a white board where robots can read and write information in a mutual exclusive way. A recently posed question (Czyzowicz et al. 2009) is whether some number of robots, expressed as a function of parameter Delta only, is sufficient to detect black holes in directed graphs of arbitrarily large order n. We give a positive answer to this question for the synchronous case, i.e., when the robots share a common clock, showing that $O(\Delta 2^{\Delta})$ robots are sufficient to solve the problem. This bound is nearly tight, since it is known that at least $2^{D}elta$ robots are required for some instances. Quite surprisingly, we also show that unlike in the case of undirected graphs, for the directed version of the problem, synchronization can sometimes make a difference: for $\Delta = 2$, in the synchronous case 4 robots are always sufficient, whereas in the asynchronous case at least 5 robots are sometimes required.

6.2.3. *Rendezvous for Location-Aware Agents*
Participant: Adrian Kosowski.

In [35] we studied rendezvous of two anonymous agents, where each agent knows its own initial position in the environment. Their task is to meet each other as quickly as possible. The time of the rendezvous is measured by the number of synchronous rounds that agents need to use in the worst case in order to meet. In each round, an agent may make a simple move or it may stay motionless. We consider two types of environments, finite or infinite graphs and Euclidean spaces. A simple move traverses a single edge (in a graph) or at most a unit distance (in Euclidean space). The rendezvous consists in visiting by both agents the same point of the environment simultaneously (in the same round). In [35], we propose several asymptotically optimal rendezvous algorithms. In particular, we show that in the line and trees as well as in multi-dimensional Euclidean spaces and grids the agents can rendezvous in time $O(d)$, where d is the distance between the initial positions of the agents. The problem of location-aware rendezvous was studied before in the asynchronous model for Euclidean spaces and multi-dimensional grids, where the emphasis was on the length of the adopted rendezvous trajectory. We point out that, contrary to the asynchronous case, where the cost of rendezvous is dominated by the size of potentially large neighborhoods, the agents are able to meet in all graphs of at most n nodes in time almost linear in d, namely, $O(d \log^2 n)$. We also determine an infinite family of graphs in which synchronized rendezvous takes time $\Omega(d)$.

6.2.4. *Boundary Patrolling by Mobile Agents*
Participant: Adrian Kosowski.

In the boundary patrolling problem, a set of k mobile agents are placed on the boundary of a simply connected planar object represented by a cycle of unit length. Each agent has its own predefined maximal speed, and is capable of moving around this boundary without exceeding its maximal speed. The agents are required to protect the boundary from an intruder which attempts to penetrate to the interior of the object through a point of the boundary, unknown to the agents. The intruder needs some time interval of length τ to accomplish the intrusion. Will the intruder be able to penetrate into the object, or is there an algorithm allowing the agents to move perpetually along the boundary, so that no point of the boundary remains unprotected for a time period τ? Such a problem may be solved by designing an algorithm which defines the motion of agents so as to minimize the idle time I, i.e., the longest time interval during which any fixed boundary point remains unvisited by some agent, with the obvious goal of achieving $I < \tau$. Depending on the type of the environment, this problem is known as either boundary patrolling or fence patrolling in the robotics literature. The most common heuristics adopted in the past include the cyclic strategy, where agents move in one direction around the cycle covering the environment, and the partition strategy, in which the environment is partitioned into sections patrolled separately by individual agents. We have obtained, to our knowledge, the first study of the fundamental problem of boundary patrolling by agents with distinct maximal speeds. In this scenario, we give special attention to the performance of the cyclic strategy and the partition strategy. In [36], we propose general bounds and methods for analyzing these strategies, obtaining exact results for cases with 2, 3, and 4 agents. We show that there are cases when the cyclic strategy is optimal, cases when the partition strategy is optimal and, perhaps more surprisingly, novel, alternative methods have to be used to achieve optimality.

6.2.5. *Rendezvous in trees*
Participant: Adrian Kosowski.

In the rendezvous problem in trees, two identical (anonymous) mobile agents start from arbitrary nodes of an unknown tree and have to meet at some node. Agents move in synchronous rounds: in each round an agent can either stay at the current node or move to one of its neighbors. We consider deterministic algorithms for this rendezvous task. In [51] we have presented a tight trade-off between the optimal time of completing rendezvous and the size of memory of the agents. For agents with k memory bits, we show that optimal rendezvous time is $\Theta(n + n^2/k)$ in n-node trees. More precisely, if $k \geq c \log n$, for some constant c, we design agents accomplishing rendezvous in arbitrary trees of unknown size n in time $O(n + n^2/k)$, starting with arbitrary delay. We also show that no pair of agents can accomplish rendezvous in time $o(n + n^2/k)$, even in the class of lines of known length and even with simultaneous start. Finally, we prove that at least logarithmic memory is necessary for rendezvous, even for agents starting simultaneously in a n-node line.

6.2.6. *How many oblivious robots can explore a line*

Participant: David Ilcinkas.

In [20] we consider the problem of exploring an anonymous line by a team of k identical, oblivious, asynchronous deterministic mobile robots that can view the environment but cannot communicate. We completely characterize sizes of teams of robots capable of exploring a n-node line. For $k < n$, exploration by k robots turns out to be possible, if and only if either $k = 3$, or $k \geq 5$, or $k = 4$ and n is odd. For all values of k for which exploration is possible, we give an exploration algorithm. For all others, we prove an impossibility result.

6.2.7. *Asynchronous deterministic rendezvous in bounded terrains*

Participant: David Ilcinkas.

Two mobile agents (robots) have to meet in an a priori unknown bounded terrain modeled as a polygon, possibly with polygonal obstacles. Robots are modeled as points, and each of them is equipped with a compass. Compasses of robots may be incoherent. Robots construct their routes, but the actual walk of each robot is decided by the adversary that may, e.g., speed up or slow down the robot. In [18], we consider several scenarios, depending on three factors: (1) obstacles in the terrain are present, or not, (2) compasses of both robots agree, or not, (3) robots have or do not have a map of the terrain with their positions marked. The cost of a rendezvous algorithm is the worst-case sum of lengths of the robots' trajectories until they meet. For each scenario we design a deterministic rendezvous algorithm and analyze its cost. We also prove lower bounds on the cost of any deterministic rendezvous algorithm in each case. For all scenarios these bounds are tight.

6.2.8. *On the Power of Waiting when Exploring Public Transportation Systems*

Participants: David Ilcinkas, Ahmed Mouhamadou Wade.

We study the problem of exploration by a mobile entity (agent) of a class of dynamic networks, namely the periodically-varying graphs (the PV-graphs, modeling public transportation systems, among others). These are defined by a set of carriers following infinitely their prescribed route along the stations of the network. Flocchini, Mans, and Santoro [58] (ISAAC 2009) studied this problem in the case when the agent must always travel on the carriers and thus cannot wait on a station. They described the necessary and sufficient conditions for the problem to be solvable and proved that the optimal number of steps (and thus of moves) to explore a n-node PV-graph of k carriers and maximal period p is in $\Theta(k \cdot p^2)$ in the general case.

In [46], we study the impact of the ability to wait at the stations. We exhibit the necessary and sufficient conditions for the problem to be solvable in this context, and we prove that waiting at the stations allows the agent to reduce the worst-case optimal number of moves by a multiplicative factor of at least $\Theta(p)$, while the time complexity is reduced to $\Theta(n \cdot p)$. (In any connected PV-graph, we have $n \leq k \cdot p$.) We also show some complementary optimal results in specific cases (same period for all carriers, highly connected PV-graphs). Finally this new ability allows the agent to completely map the PV-graph, in addition to just explore it.

6.2.9. *The impact of edge deletions on the number of errors in networks*

Participants: Christian Glacet, Nicolas Hanusse, David Ilcinkas.

In [41], we deal with an error model in distributed networks. For a target t, every node is assumed to give an *advice*, i.e. to point to a neighbor that take closer to the destination. Any node giving a bad advice is called *a liar*. Starting from a situation without any liar, we study the impact of topology changes on the number of liars.

More precisely, we establish a relationship between the number of liars and the number of distance changes after one edge deletion. Whenever ℓ deleted edges are chosen uniformly at random, for any graph with n nodes, m edges and diameter D, we prove that the expected number of liars and distance changes is $O(\frac{\ell^2 Dn}{m})$ in the resulting graph. The result is tight for $\ell = 1$. For some specific topologies, we give more precise bounds.

6.2.10. Computations in interconnection networks with a shared whiteboard

Participant: Adrian Kosowski.

In [52], We study the computational power of graph-based models of distributed computing in which each node additionally has access to a global whiteboard. A node can read the contents of the whiteboard and, when activated, can write one message of $O(\log n)$ bits on it. A message is only based on the local knowledge of the node and the current content of the whiteboard. When the protocol terminates, each node computes the output based on the final contents of the whiteboard in order to answer some question on the network's topology. We propose a framework to formally define several scenarios modelling how nodes access the whiteboard, in a synchronous way or not. This extends the work of Becker *et al.* where nodes were imposed to create their messages only based on their local knowledge (i.e., with the whiteboard empty). We prove that the four models studied have increasing power of computation: any problem that can be solved in the weakest one can be solved in the second, and so on. Moreover, we exhibit problems that *separate* models, i.e., that can be solved in one model but not in a weaker one. These problems are related to Maximal Independent Set and detection of cycles. Finally we investigate problems related to connectivity as the construction of spanning- or BFS-tree in our different models.

6.2.11. Network Verification

Participant: Ralf Klasing.

In [27], we address the problem of verifying the accuracy of a map of a network by making as few measurements as possible on the nodes of the network. In the past, this task has been formalized as an optimization problem that, given a graph $G = (V, E)$, asks for finding a minimum-size subset Q of vertices of G such that the information returned by the *queries* on Q uniquely identifies G. Previously, two *global* query models have been studied. In [27], we propose a query model that uses only *local* knowledge about the network. Quite naturally, we assume that a query at a given node q returns the associated routing table, namely a set of entries which provides, for each destination node, a corresponding (set of) first-hop node(s) along an underlying shortest path. First, we show that any network of n nodes needs $\Omega(\log \log n)$ queries to be verified. Then, we prove that there is no $o(\log n)$ approximation algorithm for the problem, unless $\mathsf{P} = \mathsf{NP}$, even for graphs with diameter 2. On the positive side, we provide an $O(\log n)$-approximation algorithm to verify a network of diameter 2, and we give exact polynomial-time algorithms for paths, trees and cycles of even length.

6.3. Compact and distributed data structures

6.3.1. Query optimization in databases

Participants: Nicolas Hanusse, Sofian Maabout.

Datacubes are data structures designed to query optimization in databases. In [42] we provide some algorithmic solutions in a user centric setting: the request time is guaranteed and the amount of memory space is minimized.

6.3.2. Parallel computations of Borders

Participants: Nicolas Hanusse, Sofian Maabout.

Borders are fundamental building blocks in data mining. They are used to find frequent patterns, dependencies between attributes, ... In[43] we provide an algorithm that computes borders with a speedup of p (under reasonable hypothesis) for p cores.

6.3.3. *Node-disjoint multipath spanners and their relationship with fault-tolerant spanners*
Participants: Cyril Gavoille, Quentin Godfroy.

Motivated by multipath routing, we introduce a multi-connected variant of spanners. For that purpose we introduce in [40] the p-multipath cost between two nodes u and v as the minimum weight of a collection of p internally vertex-disjoint paths between u and v. Given a weighted graph G, a subgraph H is a p-multipath s-spanner if for all u, v, the p-multipath cost between u and v in H is at most s times the p-multipath cost in G. The s factor is called the stretch.

Building upon recent results on fault-tolerant spanners, we show how to build p-multipath spanners of constant stretch and of $O(n^{1+1/k})$ edges, for fixed parameters p and k, n being the number of nodes of the graph. Such spanners can be constructed by a distributed algorithm running in $O(k)$ rounds.

Additionally, we give an improved construction for the case $p = k = 2$. Our spanner H has $O(n^{3/2})$ edges and the p-multipath cost in H between any two node is at most twice the corresponding one in G plus $O(W)$, W being the maximum edge weight.

6.3.4. *On approximate distance labels and routing schemes with affine stretch*
Participant: Cyril Gavoille.

For every integral parameter $k > 1$, given an unweighted graph G, we construct in polynomial time in [25], for each vertex u, a distance label $L(u)$ of size $O(n^{2/(2k-1)})$. For any $u, v \in G$, given $L(u), L(v)$ we can return in time $O(k)$ an *affine* approximation $\eth(u, v)$ on the distance $d(u, v)$ between u and v in G such that $d(u, v) \leq \eth(u, v) \leq (2k - 2)d(u, v) + 1$. Hence we say that our distance label scheme has affine stretch of $(2k - 2)d + 1$. For $k = 2$ our construction is comparable to the $O(n^{5/3})$ size, $2d + 1$ affine stretch of the distance oracle of Pătraşcu and Roditty (FOCS '10), it incurs a $o(\log n)$ storage overhead while providing the benefits of a distance label.

For any $k > 1$, given a restriction of $o(n^{1+1/(k-1)})$ on the total size of the data structure, our construction provides distance labels with affine stretch of $(2k - 2)d + 1$ which is better than the stretch $(2k - 1)d$ scheme of Thorup and Zwick (J. ACM '05).

Our second contribution is a compact routing scheme with poly-logarithmic addresses that provides affine stretch guarantees. With $O(n^{3/(3k-2)})$-bit routing tables we obtain affine stretch of $(4k - 6)d + 1$, for any $k > 1$.

Given a restriction of $o(n^{1/(k-1)})$ on the table size, our routing scheme provides affine stretch which is better than the stretch $(4k - 5)d$ routing scheme of Thorup and Zwick (SPAA '01).

6.3.5. *Sparse spanners vs. compact routing*
Participant: Cyril Gavoille.

Routing with *multiplicative* stretch 3 (which means that the path used by the routing scheme can be up to three times longer than a shortest path) can be done with routing tables of $\Theta(\sqrt{n})$ bits[3] per node. The space lower bound is due to the existence of dense graphs with large girth. Dense graphs can be sparsified to subgraphs, called *spanners*, with various stretch guarantees. There are spanners with *additive* stretch guarantees (some even have constant additive stretch) but only very few additive routing schemes are known.

---

[3]Tilde-big-O notation is similar to big-O notation up to factors poly-logarithmic in n.

In [39], we give reasons why routing in unweighted graphs with *additive* stretch is difficult in the form of space lower bounds for general graphs and for planar graphs. We prove that any routing scheme using routing tables of size M bits per node and addresses of poly-logarithmic length has additive stretch $\Omega(\sqrt{n/M})$ for general graphs, and $\Omega(\sqrt{n}/M)$ for planar graphs. Routing with tables of size $O(n^{1/3})$ thus requires a polynomial additive stretch of $\Omega(n^{1/3})$, whereas spanners with average degree $O(n^{1/3})$ and *constant* additive stretch exist for all graphs. Spanners, however sparse they are, do not tell us how to route. These bounds provide the first separation of sparse spanner problems and compact routing problems.

On the positive side, we give an almost tight upper bound: we present the first non-trivial compact routing scheme with $o(\lg^2 n)$-bit addresses, *additive* stretch $O(n^{1/3})$, and table size $O(n^{1/3})$ bits for planar graphs. Actually, our scheme applies to general graphs, and the above bound of $n^{1/3}$ on the stretch and the routing table size holds for all graphs of linear local tree-width, a class of graphs including bounded-genus graphs and apex-minor-free graphs.

7. Contracts and Grants with Industry

7.1. 4SH

Participants: Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois, Ludovic Courtès.

We started an informal collaboration with Xavier Hanin (4SH) who has developed Xooctory and who has initiated the project Ivy which is now a project of the Apache Software Foundation. This collaboration is supported by INRIA who delegated Ludovic Courtès (INRIA SED Engineer) to work in Cepage for one year from July 2009 on a distributed version of Xooctory.

We have tested and implemented in Xooctory different scheduling algorithms that distribute the build process. This lead in particular to the conception of Hubble (see Section 5.1) software.

7.2. Alcatel Luccent

Participants: Cyril Gavoille, Nicolas Hanusse, David Ilcinkas.

In this collaboration, we mainly focus on the scalability properties that a new routing protocol should guarantee (see Section 8.3). The main measures are the size of the local routing tables, and the time (or message complexity) to update or to generate such tables. The design of schemes achieving sub-linear space per routers, say in $\sqrt{n}$ where n is the number of AS routers, is the main challenge. The target networks are AS-network like with more than 100,000 nodes. This projet, in colloboration with the MASCOTE INRIA-project in Nice Sophia-Antipolis, makes the use of simulation, developped at both sites

7.3. IBM

Participants: Olivier Beaumont, Lionel Eyraud-Dubois, Nicolas Bonichon.

We have started an informal collaboration with IBM Montpellier and IBM Haifa in the context of ANR "SONGS". This collaboration aims to produce accurate and scalable simulations for Cloud Computing and HPC with SimGrid.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. *Projet Région Aquitaine : Performance modeling for heterogeneous platforms*
Participants: Lionel Eyraud-Dubois, Olivier Beaumont, Nicolas Bonichon.

This project in coordination with the INRIA project RUNTIME aims at designing models for communication times on heterogeneous platforms of two types: high-scale platforms for volunteer computing, and high performance NUMA machines. The goal is to reach a compromise between precision and algorithmic tractability.

8.2. National Initiatives

8.2.1. ANR "programme blanc" Aladdin (2007-2011)

Participants: Olivier Beaumont, Nicolas Bonichon, Philippe Duchon, Cyril Gavoille, Nicolas Hanusse, David Ilcinkas, Ralf Klasing.

The scientific objectives of ALADDIN are to solve what are identified as the most challenging problems in the theory of interaction networks. The ALADDIN project is thus an opportunity to create a full continuum from fundamental research to applications in coordination with both INRIA projects CEPAGE and GANG.

8.2.2. ANR programme "défis": project IDEA (2009-2012)

Participant: Ralf Klasing.

The goal of this ANR is the study of identifying codes in evolving graphs. Ralf Klasing is the overall leader of the project.

8.2.3. ANR "USS SimGrid" (2008-2011)

Participants: Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois.

The objectives of USS SimGrid is to create a simulation framework that will answer (i) the need for simulation scalability arising in the HPC community; (ii) the need for simulation accuracy arising in distributed computing. The Cepage team will be involved in the development of tools to provide realistic model instantiations.

The project involves the following INRIA and CNRS teams: AlGorille, ASAP, Cepage, Graal, MESCAL, SysCom, CC IN2P3.

8.2.4. ANR "programme blanc" Displexity (2012-2015)

Participants: Cyril Gavoille, David Ilcinkas, Ralf Klasing, Adrian Kosowski.

The main goal of DISPLEXITY (for DIStributed computing: computability and ComPLEXITY) is to establish the scientific foundations for building up a consistent theory of computability and complexity for distributed computing. The other partners are from IRISA (Rennes) and LIAFA (Paris).

8.2.5. ANR "SONGS" (2012-2016)

Participants: Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois.

SONGS (Simulation of Next Generation Systems) is a follow-up to the USS-SimGrid project. Its objective is to design a unified and open simulation framework for performance evaluation of next generation systems: Grids, Peer-to-Peer systems, Clouds and HPC systems. Cepage will be involved in the Peer-to-peer and Cloud use cases by designing and testing efficient allocation policies. Cepage will also take part in the design of efficient and realistic models and their validation.

The project involves the following INRIA and CNRS teams: AlGorille, ASCOLA, AVALON, CEPAGE, HiePACS, ICPS, MASCOTTE, MODALIS, MESCAL, RUNTIME, CC IN2P3.

8.3. European Initiatives

8.3.1. FP7 Projet

8.3.1.1. EULER

Title: EULER (Experimental UpdateLess Evolutive Routing)

Type: COOPERATION (ICT)

Defi: Future Internet Experimental Facility and Experimentally-driven Research

Instrument: Specific Targeted Research Project (STREP)

Duration: October 2010 - September 2013

Coordinator: ALCATEL-LUCENT (Belgium)

Others partners:

- Alcatel-Lucent Bell, Antwerpen, Belgium
- 3 projects from INRIA: CEPAGE, GANG and MASCOTTE, France
- Interdisciplinary Institute for Broadband Technology (IBBT),Belgium
- Laboratoire d'Informatique de Paris 6 (LIP6), Université Pierre Marie Curie (UPMC), France
- Department of Mathematical Engineering (INMA) Université Catholique de Louvain, Belgium
- RACTI, Research Academic Computer Technology Institute University of Patras, Greece
- CAT, Catalan Consortium: Universitat Politècnica de Catalunya, Barcelona and University of
- Girona, Spain

See also: http://www-sop.inria.fr/mascotte/EULER/wiki/

Abstract: The title of this study is "Dynamic Compact Routing Scheme". The aim of this projet is to develop new routing schemes achieving better performances than current BGP protocols. The problems faced by the inter-domain routing protocol of the Internet are numerous:

1. The underlying network is dynamic: many observations of bad configurations show the instability of BGP;

2. BGP does not scale well: the convergence time toward a legal configuration is too long, the size of routing tables is proportional to the number of nodes of network (the network size is multiplied by 1.25 each year);

3. The impact of the policies is so important that the many packets can oscillated between two Autonomous Systems.

In this collaboration, we mainly focus on the scalability properties that a new routing protocol should guarantee. The main measures are the size of the local routing tables, and the time (or message complexity) to update or to generate such tables. The design of schemes achieving sub-linear space per routers, say in n where n is the number of AS routers, is the main challenge. The target networks are AS-network like with more than 100,000 nodes. This project, in collaboration with the MASCOTE INRIA-project in Nice Sophia-Antipolis, makes the use of simulation, developed at both sites.

8.3.1.2. Royal Society Grant with the University of Liverpool

Participants: Nicolas Hanusse, David Ilcinkas, Ralf Klasing, Adrian Kosowski.

International Joint Project, 2011-2013, entitled "SEarch, RENdezvous and Explore (SERENE)", on foundations of mobile agent computing, in collaboration with the Department of Computer Science, University of Liverpool. Funded by the Royal Society, U.K. Principal investigator on the UK side: Leszek Gasieniec. Ralf Klasing is the principal investigator on the French side.

8.3.1.3. European COST Action ComplexHPC (2009-2012)
Participants: Olivier Beaumont, Nicolas Bonichon, Lionel Eyraud-Dubois.

The goal of ComplexHPC is to coordinate European groups working on the use of heterogeneous and hierarchical systems for HPC as well as the development of collaborative activities among the involved research groups, to tackle the problem at every level (from cores to large-scale environments) and to provide new integrated solutions for large-scale computing for future platforms (see http://complexhpc.org/index.php).

8.3.1.4. Emergent Project with the University of Perugia
Participants: David Ilcinkas, Ralf Klasing, Adrian Kosowski.

International Joint Project, 2011, on foundations of mobile agent computing, in collaboration with the Department of Computer Science, University of Perugia, Italy. Principal investigator on the Italian side: Alfredo Navarra. Ralf Klasing is the principal investigator on the French side.

8.3.2. *Major European Organizations with which you have followed Collaborations*

Marcin Markiewicz, University of Gdansk (Poland)

Quantum distributed computing models and simulation of quantum correlations using classical information channels.

Ashley Deflumere and Alexey Lastovetsky, University College Dublin (Ireland)

Design of efficient distribution scheme for linear algebra kernels on modern heterogeneous architectures

Gabriele Di Stefano, University of L'Aquila (Italy)

Alfredo Navarra, University of Perudia (Italy)

Mobile agent coordination in distributed computing.

Miroslaw Korzeniowski, Technical University of Wroclaw (Poland)

Design of distributed and randomized algorithms for P2P networks.

Leszek Gasieniec, University of Liverpool (UK)

Design of distributed algorithms for mobile agents in exploration and patrolling tasks.

Guido Proetti, University of L'Aquila (Italy)

Davide Bilo, University of Sassari (Italy)

Network discovery and verification. ation techniques for chosen task scheduling problems.

Tobias Mömke, KTH Royal Institute of Technology, Stockholm (Sweden)

Centralized approximation techniques for chosen task scheduling problems.

Thomas Sauerwald, Max-Planck-Institut für Informatik, Saarbrücken (Germany)

Propp machine, Multiple random walks.

8.4. International Initiatives

8.4.1. Visits of International Scientists

8.4.1.1. Visits to Cepage Members

- Ashley Deflumere, University College Dublin, 04/12 - 17/12/2011
- Alfredo Navarra, University of Perugia, Italy, 11/12-16/12/2011
- Ljubomir Perkovic, De Paul University Chicago, (September 2011–)
- Miroslaw Korzeniowski, Technical University of Wroclaw, (23/08- 28/08/2011)
- Marcin Markiewicz, University of Gdansk, 10/05-17/05/2011
- Miroslaw Korzeniowski, Technical University of Wroclaw, (06/07- 22/07/2011)
- Tobias Mömke, KTH Royal Institute of Technology, Stockholm, Sweden, 17/07 - 31/07/2011
- Leszek Gasieniec, University of Liverpool, UK, 10/09-17/09/2011
- Alfredo Navarra, University of Perugia, Italy, 10/09-17/09/2011
- Gabriele Di Stefano, University of L'Aquila, Italy, 10/09-17/09/2011
- Davide Bilo, University of Sassari, Italy, 27/11-07/12/2011

8.4.1.2. Visits of Cepage Members

- Microsoft Research Montainview, CA, invited research visit by I. Abhram (C. Gavoille, 10 days, April 2011)
- Weizmann Institute, research visit with D. Peleg (Q. Godfroy, one week, November 2011)
- Universidad Adolfo Ibanez, Chile, research visit as part of joint grant (A. Kosowski, 16/01-03/02/2011)
- University of Gdansk, Poland, research visit (A. Kosowski, 07/02-14/02/2011)
- University of Liverpool, UK, research visit as part of joint grant (A. Kosowski, 15/02-27/02/2011)
- Carleton University, Canada, invited research visit (A. Kosowski, 08/11-20/11/2011)

8.4.1.3. Internship

8.4.1.4. Major Non-European Organizations with which you have followed Collaborations

- Weizmann Institute (IL), David Peleg
- MIT (USA), Christian Sommer
- MicroSoft Research, Montainview (USA), Ittai Abraham

8.4.2. Participation In International Programs

8.4.2.1. Foreign partner of ANILLO grant at Universidad Adolfo Ibanez
Participant: Adrian Kosowski.

Foreign partner of the project entitled "Mathematical modeling for industrial and management science" funded by the Government of Chile through its CONICYT program (ANILLO for Science and Technology).

This grant involves research into mathematical programming models, network dynamics and graph models, stochastic models, as well as other interdisciplinary projects. The joint work performed during the research collaboration lead to new results on the computational power of interconnection networks in distributed computing, and to new algorithms for compact routing in special graph classes.

8.4.2.2. Foreign partner of Comunidad de Madrid grant
Participants: Olivier Beaumont, Lionel Eyraud-Dubois.

Foreign partner of the project entitled "CLOUDS: Cloud Computing para Servicios Escalables, Confiables y Ubicuos" (2010-2013) funded by the Comunidad de Madrid.

9. Dissemination

9.1. Animation of the scientific community

9.1.1. Editorial Work

Ralf Klasing is a member of the Editorial Board of *Theoretical Computer Science*, *Discrete Applied Mathematics*, *Wireless Networks*, *Networks*, *Journal of Interconnection Networks*, *Parallel Processing Letters*, *Algorithmic Operations Research*, *Fundamenta Informaticae*, and *Computing and Informatics*.

Adrian Kosowski is a guest editor for *Theoretical Computer Science* and *Computing and Informatics*.

Olivier Beaumont is Associate Editor for *IEEE Transactions on Parallel and Distributed Systems (IEEE TPDS)*.

9.1.1.1. Conference Chair

- Olivier Beaumont was the Vice-Chair (Algoritm Track) of *IPDPS 2011*, the 25th IEEE International Parallel and Distributed Processing Symposium, May 16-20, 2011 Anchorage (Alaska) USA

- Nicolas Hanusse is the Conference Chair of *14th French Conference of Communications in Network (Algotel 2012)*

- Nicolas Hanusse is the Conference Chair of *International Conference on Algorithm (FUN 2012)*

- Ralf Klasing was the Conference Chair of the *Bordeaux Workshop on Identifying Codes (BWIC 2011)*, Bordeaux, France, 21/11-25/11/2011.

- Ralf Klasing is the Conference Chair of the *11th International Symposium on Experimental Algorithms (SEA 2012)*, Bordeaux, France, June 7-9, 2012.

- Adrian Kosowski was the Conference Organization & Proceedings Co-Chair of the *18th International Colloquium on Structural Information and Communication Complexity (SIROCCO 2011)*, Gdańsk, Poland, June 26-29, 2011.

9.1.1.2. Steering Committees

- Cyril Gavoille is a member of the Steering Committee of the *ACM Symposium on Principles of Distributed Computing (PODC)*.

- Ralf Klasing is a member of the Steering Committee of the *International Colloquium on Structural Information and Communication Complexity (SIROCCO)*.

9.1.1.3. Program Committees

- ESA 2011 (Sep. 5-9, Saarbrücken, Germany) Annual European Symposium on Algorithms (C. Gavoille)

- DISC 2012 (Oct. 16-18, Salvador, Brazil), International Symposium on Distributed Computing (C. Gavoille)

- ISAAC 2012 (Dec. 19-21, Taipei, Taiwan), International Symposium on Algorithms and Computation (C. Gavoille)

- ICDCN 2012 (Jan. 3-6, Hong-Kong), International Conference on Distributed Computing and Networking (C. Gavoille)

- SEA 2012, 11th International Symposium on Experimental Algorithms, Bordeaux, France, June 7-9, 2012 (R. Klasing, A. Kosowski)

- IWOCA 2012, 23rd International Workshop on Combinatorial Algorithms, July 19-21, 2012, Kalasalingam University, Anand Nagar, Krishnankoil, Tamil Nadu, India (R. Klasing)

- SSS 2012, 14th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Toronto, September 24-27, 2012 (A. Kosowski).

- ICDCN 2012, 13th International Conference on Distributed Computing and Networking, January 3-6, 2012, Hong Kong, China (A. Kosowski)

- ICPP 2012, 41st International Conference on Parallel Processing, Pittsburgh, PA, USA, September 10-13, 2012 (O. Beaumont)

- IPDPS 2012, 26th IEEE International Parallel & Distributed Processing Symposium, May 21-25, 2012, Shanghai, China (O. Beaumont)

- HCW 2012, 20th International Heterogeneity in Computing Workshop, May 21-25, 2012, Shanghai, China (O. Beaumont)

- HIPC 2011, 18th annual IEEE International Conference on High Performance Computing (HiPC 2011), Bengaluru (Bangalore), India, December 18-21, 2011. (O. Beaumont)

- IEEE CloudCom 2011, 3rd IEEE International Conference on Cloud Computing Technology and Science (IEEE CloudCom 2011), Athens, Greece, Nov 29 - Dec 1, 2011 (O. Beaumont)

- ISCIS 2011, 26th International Symposium on Computer and Information Sciences, 26-28 September 2011, London, UK (O. Beaumont)

- EUROPAR 2011, local chair (P2P systems track), 17th International European Conference on Parallel and Distributed Computing Euro-Par 2011, Bordeaux, France, August 29-September 2, 2011 (O. Beaumont)

- IC3, The Fourth International Conference on Contemporary Computing (IC3-2011), NOIDA, (outskirts of New Delhi), India, August 8-10, 2011 (O. Beaumont)

- OPODIS 2011, 15th International Conference On Principles Of Distributed Systems, December 13-16, 2011, Toulouse, France (O. Beaumont)

- IPDPS 2011, Vice-Chair (Algorithms) 25th IEEE International Parallel & Distributed Processing Symposium, May 16-20, 2011 Anchorage (Alaska) USA (O. Beaumont)

- SIROCCO 2011, 18th International Colloquium on Structural Information and Communication Complexity, June 26-29, 2011, Gdansk, Poland (R. Klasing, A. Kosowski)

- DISC 2011, 25th International Symposium on DIStributed Computing, September 20-22, 2011, Rome, Italy (D. Ilcinkas, A. Kosowski)

- FOMC 2011, 7th ACM SIGACT/SIGMOBILE International Workshop on Foundations of Mobile Computing (formerly known as DIALM-POMC), June 9, 2011, San Jose, California, USA (R. Klasing)

- ADHOC-NOW 2011, 10th International Conference on Ad Hoc Networks and Wireless, July 18-20, 2011, Paderborn, Germany (R. Klasing)

- IWOCA 2011, 22nd International Workshop on Combinatorial Algorithms, June 20-22, 2011, University of Victoria, Canada (R. Klasing)

- ALGOSENSORS - Track B 2011, 7th International Symposium on Algorithms for Sensor Systems, Wireless Ad Hoc Networks and Autonomous Mobile Entities (D. Ilcinkas)

- AlgoTel 2011, 13èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (D. Ilcinkas)

- CSA 2011, 3rd FTRA International Conference on Computer Science and its Applications, December 12-15, 2011, Jeju, Korea (L. Eyraud-Dubois)

9.1.2. Organizing Committees

- Cyril Gavoille was the General Chair of *PODC 2011*, the 30th Annual ACM Symposium on Principles of Distributed Computing, June 6-8, San Jose, California.
- Ralf Klasing was the Conference Chair of the *Bordeaux Workshop on Identifying Codes (BWIC 2011)*, Bordeaux, France, 21/11-25/11/2011.
- Ralf Klasing is the Conference Chair of the *11th International Symposium on Experimental Algorithms (SEA 2012)*, Bordeaux, France, June 7-9, 2012.
- Adrian Kosowski was the Conference Organization Chair of the *18th International Colloquium on Structural Information and Communication Complexity (SIROCCO 2011)*, Gdańsk, Poland, June 26-29, 2011.
- David Ilcinkas was member of the organization committee of the second TERANET workshop (***T***oward ***E***volutive ***R***outing ***A***lgorithms *for scale-free/internet-like **NET**works*), September 19, 2011.

9.1.3. Research Administration

- Olivier Beaumont is the "Délégué Scientifique" and the head of the Project Committee at INRIA Bordeaux Sud-Ouest.
- Ralf Klasing is the Head of the "Combinatorics and Algorithms" team of the LaBRI.
- Ralf Klasing is a member of the *Conseil Scientique* of the LaBRI.
- Ralf Klasing is a member of the *Commission Consultative* of the LaBRI.
- Ralf Klasing is responsible for the International Relations of the LaBRI.
- Ralf Klasing is a member of the Evaluation Committee of the programme "ANR DEFIS 2009".

9.1.4. PhD Thesis Refereeing Activities

- Olivier Beaumont was external reviewer of The-Thung NGuyen (LAAS Toulouse), November 2011
- Olivier Beaumont was external reviewer of Mathieu Valero (LIP6, Paris), December 2011
- Nicolas Hanusse was reviewer of Thomas Aynaud (Université Paris 6), Novembre 2011
- Nicolas Hanusse was reviewer of Mauricio Soto (Université Paris 7), December 2011
- Ralf Klasing was reviewer and opponent in the Ph.D. defense of Ville Junnila (University of Turku, Finland), June 2011.
- Cyril Gavoille was external reviewer of Mathieu Chapelle (Université d'Orléans, LIFO), December 2011.

9.1.5. Invited Talks

- Ralf Klasing was an Invited Speaker at the *2nd Nordic Workshop on System and Network Optimization for Wireless (SNOW 2011)*, Sälen, Sweden, 24 - 26 March 2011. (Title of the talk: Cost Minimization in Multi-Interface Networks.)
- Cyril Gavoille was an Invited Speaker at the First International Workshop on Dynamic Systems (DYNAM), Toulouse, December 2011, joint with OPODIS. (Title: Dynamic algorithms via forbidden-set labeling).
- Cyril Gavoille was an Invited Speaker at the "10 ans du séminaire MaMux - Mathématiques, musique et relations avec d'autres disciplines", Paris, May 2011. (Title: Oracles pour les arbres et les graphes).

9.2. Teaching

Licence : The members of CEPAGE are heavily involved in teaching activities at undergraduate level (Licence 1, 2 and 3, Master 1 and 2, Engineering Schools ENSEIRB). The teaching is carried out by members of the University as part of their teaching duties, and for CNRS/INRIA (at master 2 level) as extra work. It represents more than 500 hours per year.

Master : Communication and Routing (last year of engineering school ENSEIRB, 2011) O. Beaumont, L. Eyraud, N. Hanusse, R. Klasing, A. Kosowski (16h)

Master : Communication Algorithms in Networks (2nd year MASTER "Algorithms and Formal Methods", University of Bordeaux, 2011) R. Klasing (24h)

Master : Structure of Web Search Engines, ENSEIRB, O. Beaumont (16h)

Doctorat : titre du cours, nombre d'heures en équivalent temps plein, université, pays

Master : Distributed Algorithms (2nd year MASTER Informatique, University of Bordeaux, 2011) Cyril Gavoille (32h)

10. Bibliography

Major publications by the team in recent years

[1] O. BEAUMONT, H. REJEB. *On the Importance of Bandwidth Control Mechanisms for Scheduling on Large Scale Heterogeneous Platforms*, in "24th IEEE International Parallel and Distributed Processing Symposium (IPDPS 2010)", Atlanta, United States, April 2010, http://hal.inria.fr/inria-00444585/en.

[2] N. BONICHON, C. GAVOILLE, N. HANUSSE, L. PERKOVIC. *Plane Spanners of Maximum Degree Six*, in "Automata, Languages and Programming 37th International Colloquium, ICALP 2010, Bordeaux, France, July 6-10, 2010, Proceedings, Part I", France, 2010, vol. 6198, p. 19-30 [*DOI* : 10.1007/978-3-642-14165-2_3], http://hal.archives-ouvertes.fr/hal-00534212/en/.

[3] C. COOPER, D. ILCINKAS, R. KLASING, A. KOSOWSKI. *Derandomizing random walks in undirected graphs using locally fair exploration strategies*, in "Distributed Computing", 2011, vol. 24, n⁰ 2, p. 91-99, See paper for details [*DOI* : 10.1007/s00446-011-0138-4], http://hal.archives-ouvertes.fr/hal-00638229/en/.

[4] J. CZYZOWICZ, A. KOSOWSKI, A. PELC. *How to meet when you forget: log-space rendezvous in arbitrary graphs*, in "29th Annual ACM Symposium on Principles of Distributed Computing (PODC)", ACM Press, July 2010, p. 450-459.

[5] P. DUCHON, N. HANUSSE, E. LEBHAR. *Towards small world emergence*, in "Proceedings of the eighteenth annual ACM symposium on Parallelism in algorithms and architectures", ACM Press New York, NY, USA, 2006, p. 225–232.

[6] P. DUCHON, N. HANUSSE, N. SAHEB, A. ZEMMARI. *Broadcast in the rendezvous model*, in "Information and Computation", 2006, vol. 204, n⁰ 5, p. 697–712.

[7] L. EYRAUD-DUBOIS, G. MOUNIÉ, D. TRYSTRAM. *Analysis of Scheduling Algorithms with Reservations*, in "IEEE International Parallel and Distributed Processing Symposium, IPDPS 2007", Long Beach, California, USA, IEEE, 2007, p. 1-8, http://dx.doi.org/10.1109/IPDPS.2007.370304.

[8] P. FLOCCHINI, D. ILCINKAS, N. SANTORO. *Ping Pong in Dangerous Graphs: Optimal Black Hole Search with Pure Tokens*, in "Proceedings of the 22nd International Symposium on Distributed Computing", France, G. TAUBENFELD (editor), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, September 2008, vol. 5218, p. 227-241, This work was done during the stay of David Ilcinkas at the University of Ottawa, as a postdoctoral fellow. Paola Flocchini was partially supported by the University Research Chair of the University of Ottawa. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada under Discovery grants [*DOI* : 10.1007/978-3-540-87779-0_16], http://hal.archives-ouvertes.fr/hal-00341523/en/.

[9] P. FRAIGNIAUD, C. GAVOILLE, D. ILCINKAS, A. PELC. *Distributed Computing with Advice: Information Sensitivity of Graph Coloring*, in "ICALP", 2007, p. 231-242.

[10] P. FRAIGNIAUD, C. GAVOILLE, A. KOSOWSKI, E. LEBHAR, Z. LOTKER. *Universal augmentation schemes for network navigability: overcoming the $\sqrt{n}$-barrier*, in "Proceedings of the nineteenth annual ACM symposium on parallelism and architectures", San Diego, California, États-Unis, ACM, 2007, p. 1-7, COST Action 295 "Dynamo", ACI project "Fragile", INRIA project "Grand Large", http://hal.archives-ouvertes.fr/hal-00155186/en/.

[11] E. GARNAUD, S. MAABOUT, M. MOSBAH. *Using Functional Dependencies for Reducing the Size of a Data Cube*, in "Proceedings of the 7th Int. Symp. on Foundations of Information and Knowledge Systems (FoIKS)", LNCS, Springer, 2012, 20.

[12] L. GASIENIEC, R. KLASING, M. RUSSELL, A. NAVARRA, X. ZHANG. *Fast Periodic Graph Exploration with Constant Memory*, in "Journal of Computer and System Sciences", 2008, vol. 74, n⁰ 5, p. 808–822 [*DOI : 10.1016/J.JCSS.2007.09.004*], http://hal.archives-ouvertes.fr/hal-00306995/en/.

[13] J. HROMKOVIČ, P. KANAREK, R. KLASING, K. LORYS, W. UNGER, H. WAGENER. *On the Size of Permutation Networks and Consequences for Efficient Simulation of Hypercube Algorithms on Bounded-Degree Networks*, in "SIAM Journal on Discrete Mathematics", 2009, vol. 23, n⁰ 3, p. 1612–1645, http://hal.archives-ouvertes.fr/hal-00402764/en/.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[14] E. DIOT. *Etude des graphes k chemins séparables*, Université Sciences et Technologies - Bordeaux I, December 2011.

[15] H. REJEB. *Étude des problèmes d'ordonnancement sur des plates-formes hétérogènes en modèle multi-port*, Université Sciences et Technologies - Bordeaux I, August 2011, http://hal.inria.fr/tel-00651988/en.

Articles in International Peer-Reviewed Journal

[16] C. COOPER, D. ILCINKAS, R. KLASING, A. KOSOWSKI. *Derandomizing random walks in undirected graphs using locally fair exploration strategies*, in "Distributed Computing", 2011, vol. 24, n⁰ 2, p. 91-99 [*DOI : 10.1007/S00446-011-0138-4*], http://hal.inria.fr/hal-00638229/en.

[17] B. COURCELLE, C. GAVOILLE, M. M. KANTÉ. *Compact Labelings For Efficient First-Order Model-Checking*, in "Journal of Combinatorial Optimization", January 2011, vol. 21, n⁰ 1, p. 19–46, http://hal.inria.fr/hal-00342668/en.

[18] J. CZYZOWICZ, D. ILCINKAS, A. LABOUREL, A. PELC. *Asynchronous deterministic rendezvous in bounded terrains*, in "Theoretical Computer Science", November 2011, vol. 412, n⁰ 50, p. 6926-6937 [*DOI : 10.1016/J.TCS.2011.09.002*], http://hal.inria.fr/hal-00643686/en.

[19] Y. DIENG, C. GAVOILLE. *Routage compact optimal dans les (k, r)-constellations*, in "Technique et Science Informatiques (TSI)", May 2011, vol. 30, n⁰ 05/2011, p. 485-513 [*DOI : 10.3166/TSI.30.485-513*], http://hal.inria.fr/hal-00651841/en.

[20] P. FLOCCHINI, D. ILCINKAS, A. PELC, N. SANTORO. *How many oblivious robots can explore a line*, in "Information Processing Letters", October 2011, vol. 111, n° 20, p. 1027-1031 [*DOI :* 10.1016/J.IPL.2011.07.018], http://hal.inria.fr/hal-00643668/en.

[21] A. KOSOWSKI, A. NAVARRA, C. PINOTTI. *Synchronous Black Hole Search in Directed Graphs*, in "Theoretical Computer Science", June 2011, vol. 412, n° 41, p. 5752-5759 [*DOI :* 10.1016/J.TCS.2011.05.054], http://hal.inria.fr/inria-00614476/en.

[22] J.-F. MARCKERT, N. BONICHON. *Asymptotics of geometrical navigation on a random set of points of the plane*, in "Advances in Applied Probability", 2011, vol. 43, p. 1–44, http://hal.inria.fr/hal-00649901/en.

Invited Conferences

[23] C. GAVOILLE. *Dynamic Algorithms via Forbidden-Set Labeling*, in "First International Workshop on Dynamic Systems (DYNAM)", LASS, Toulouse, France, December 2011, http://hal.inria.fr/hal-00656916/en.

[24] C. GAVOILLE. *Oracles pour les arbres et les graphes*, in "10 ans du séminaire MaMux - Mathématiques, musique et relations avec d'autres disciplines", IRCAM, Paris, France, May 2011, http://hal.inria.fr/hal-00651845/en.

International Conferences with Proceedings

[25] I. ABRAHAM, C. GAVOILLE. *On approximate distance labels and routing schemes with affine stretch*, in "25th International Symposium on Distributed Computing (DISC)", Rome, Italy, Lecture Notes in Computer Science (ARCoSS), Springer, 2011, vol. 6950, p. 404-415 [*DOI :* 10.1007/978-3-642-24100-0_39], http://hal.inria.fr/hal-00651833/en.

[26] I. ATSONIOS, O. BEAUMONT, N. HANUSSE, Y. KIM. *On Power-Law Distributed Balls in Bins and its Applications to View Size Estimation*, in "ISAAC", Yokohama, Japan, December 2011, http://hal.inria.fr/inria-00618785/en.

[27] E. BAMPAS, D. BILÒ, G. DROVANDI, L. GUALÀ, R. KLASING, G. PROIETTI. *Network Verification via Routing Table Queries*, in "SIROCCO 2011", Poland, Lecture Notes in Computer Science, 2011, vol. 6796, p. 270–281 [*DOI :* 10.1007/978-3-642-22212-2_24], http://hal.inria.fr/hal-00647362/en.

[28] O. BEAUMONT, N. BONICHON, P. DUCHON, H. LARCHEVÊQUE. *Use of Internet Embedding Tools for Heterogeneous Resources Aggregation*, in "Heterogeneity in Computing Workshop", Anchorage, United States, 2011, 10, http://hal.inria.fr/inria-00562655/en.

[29] O. BEAUMONT, N. BONICHON, P. DUCHON, H. LARCHEVÊQUE. *Use of Internet Embedding Tools for Heterogeneous Resources Aggregation*, in "Heterogeneity in Computing Workshop (HCW) – in IPDPS 2011", Anchorage, United States, IEEE, 2011, http://hal.inria.fr/inria-00588650/en.

[30] O. BEAUMONT, N. BONICHON, P. DUCHON, H. LARCHEVÊQUE. *Utilisation d'outils de plongement d'Internet pour l'agrégation de ressources hétérogènes*, in "Conférence : 13es Rencontres Francophones sur les Aspects Algorithmiques de Télécommunications (AlgoTel)", Cap Estérel, France, B. DUCOURTHIAL, P. FELBER (editors), 2011, http://hal.inria.fr/inria-00585254/en.

[31] O. BEAUMONT, N. BONICHON, L. EYRAUD-DUBOIS, L. MARCHAL. *Minimizing Weighted Mean Completion Time for Malleable Tasks Scheduling*, in "IPDPS 2012, 26th IEEE International Parallel & Distributed Processing Symposium", Shangai, China, IEEE, 2012, http://hal.inria.fr/inria-00564056/en.

[32] O. BEAUMONT, N. BONICHON, L. EYRAUD-DUBOIS, P. UZNANSKI. *Broadcasting on Large Scale Heterogeneous Platforms with connectivity artifacts under the Bounded Multi-Port Model*, in "ICPADS 2011 – 17th Internation Conference on Parallel and Distributed Systems", Taïnan, Taiwan, Province Of China, IEEE, December 2011, vol. Proceeding of ICPADS 2011, http://hal.inria.fr/inria-00618784/en.

[33] O. BEAUMONT, N. BONICHON, H. LARCHEVÊQUE. *Modeling and Practical Evaluation of a Service Location Problem in Large Scale Networks*, in "Internation Conference on Parallel Computing 2011", Taipei, Taiwan, Province Of China, March 2011, 10, http://hal.inria.fr/inria-00580969/en.

[34] O. BEAUMONT, L. EYRAUD-DUBOIS, Y. WON. *Using the Last-mile Model as a Distributed Scheme for Available Bandwidth Prediction*, in "EuroPar 2011", Bordeaux, France, Springer, 2011, http://hal.inria.fr/inria-00588651/en.

[35] A. COLLINS, J. CZYZOWICZ, L. GASIENIEC, A. KOSOWSKI, M. RUSSELL. *Synchronous Rendezvous for Location-Aware Agents*, in "DISC 2011 - 25th International Symposium on Distributed Computing", Rome, Italy, D. PELEG (editor), Lecture Notes in Computer Science, Springer, September 2011, vol. 6950, p. 447-459 [*DOI :* 10.1007/978-3-642-24100-0_42], http://hal.inria.fr/hal-00646907/en.

[36] J. CZYZOWICZ, L. GASIENIEC, A. KOSOWSKI, E. KRANAKIS. *Boundary Patrolling by Mobile Agents with Distinct Maximal Speeds*, in "ESA 2011 - 19th Annual European Symposium on Algorithms", Saarbruecken, Germany, C. DEMETRESCU, M. M. HALLDÓRSSON (editors), Lecture Notes in Computer Science, Springer, September 2011, vol. 6942, p. 701-712 [*DOI :* 10.1007/978-3-642-23719-5_59], http://hal.inria.fr/hal-00646910/en.

[37] E. DIOT, C. GAVOILLE, P. OCHEM. *Sur la difficulté de séparer un graphe par des plus courts chemins*, in "13es Rencontres Francophones sur les Aspects Algorithmiques de Télécommunications (AlgoTel)", Cap Estérel, France, B. DUCOURTHIAL, P. FELBER (editors), 2011, http://hal.inria.fr/inria-00588312/en.

[38] P. DUCHON. *Random generation of combinatorial structures: Boltzmann samplers and beyond*, in "Winter Simulation Conference", Phoenix, United States, S. JAIN, R. CREASEY, J. HIMMELSPACH, K. WHITE, M. FU (editors), December 2011, http://hal.inria.fr/hal-00654267/en.

[39] C. GAVOILLE, S. CHRISITAN. *Sparse Spanners vs. Compact Routing*, in "23rd Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)", San Jose, United States, ACM, June 2011, p. 225-234 [*DOI :* 10.1145/1989493.1989526], http://hal.inria.fr/hal-00651836/en.

[40] C. GAVOILLE, Q. GODFROY, L. VIENNOT. *Node-Disjoint Multipath Spanners and their Relationship with Fault-Tolerant Spanners*, in "15th International Conference on Principles of Distributed Systems (OPODIS)", France, Lecture Notes in Computer Science, Springer, 2011, vol. 7109, p. 143-158, http://hal.inria.fr/hal-00651825/en.

[41] C. GLACET, N. HANUSSE, D. ILCINKAS. *The Impact of Edge Deletions on the Number of Errors in Networks*, in "OPODIS 2011", Toulouse, France, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 7109, p. 378-391 [*DOI :* 10.1007/978-3-642-25873-2_26], http://hal.inria.fr/hal-00652051/en.

[42] N. HANUSSE, S. MAABOUT. *A parallel algorithm for computing borders*, in "CIKM", France, November 2011, p. 1639-1648, http://hal.inria.fr/hal-00652055/en.

[43] N. HANUSSE, S. MAABOUT. *A parallel Algorithm for Computing Borders*, in "International Conference on Information and Knowledge Managment (CIKM)", Glasgow, United Kingdom, ACM, October 2011, p. 1639-1648, http://hal.inria.fr/hal-00644517/en.

[44] N. HANUSSE, S. MAABOUT. *Revisiting the Partial Data Cube Materialization*, in "ADBIS", Austria, 2011, p. 70-83, http://hal.inria.fr/hal-00652059/en.

[45] N. HANUSSE, S. MAABOUT, R. TOFAN. *Revisiting the Partial Data Cube Materialization*, in "Conference on Advances in Data Bases and Information Systems (ADBIS)", Vienne, Austria, LNCS, Springer, September 2011, p. 70-83, http://hal.inria.fr/hal-00644526/en.

[46] D. ILCINKAS, A. WADE. *On the Power of Waiting when Exploring Public Transportation Systems*, in "OPODIS 2011", Toulouse, France, Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 7109, p. 451-464 [*DOI :* 10.1007/978-3-642-25873-2_31], http://hal.inria.fr/hal-00656607/en.

Scientific Books (or Scientific Book chapters)

[47] N. BONICHON, C. GAVOILLE, N. HANUSSE. *An Information-Theoretic Upper Bound on Planar Graphs Using Well-Orderly Maps*, in "Towards an Information Theory of Complex Networks", M. DEHMER, F. EMMERT-STREIB, A. MEHLER (editors), Birkhäuser Boston, 2011, p. 17-46, http://hal.inria.fr/hal-00649898/en.

Books or Proceedings Editing

[48] A. KOSOWSKI, M. YAMASHITA (editors). *Structural Information and Communication Complexity. 18th International Colloquium, SIROCCO 2011, Gdańsk, Poland, June 26-29, 2011*, Lecture Notes in Computer Science, Springer, June 2011, vol. 6796 [*DOI :* 10.1007/978-3-642-22212-2], http://hal.inria.fr/hal-00646905/en.

Research Reports

[49] O. BEAUMONT, L. BOBELIN, H. CASANOVA, P.-N. CLAUSS, B. DONASSOLO, L. EYRAUD-DUBOIS, S. GENAUD, S. HUNOLD, A. LEGRAND, M. QUINSON, C. ROSA, L. SCHNORR, M. STILLWELL, F. SUTER, C. THIERY, P. VELHO, J.-M. VINCENT, Y. WON. *Towards Scalable, Accurate, and Usable Simulations of Distributed Applications and Systems*, INRIA, October 2011, n[o] RR-7761, http://hal.inria.fr/inria-00631141/en.

[50] A. BENOIT, H. LARCHEVÊQUE, P. RENAUD-GOUD. *Optimal algorithms and approximation algorithms for replica placement with distance constraints in tree networks*, INRIA, September 2011, n[o] RR-7750, http://hal.inria.fr/inria-00630292/en.

[51] J. CZYZOWICZ, A. KOSOWSKI, A. PELC. *Time vs. space trade-offs for rendezvous in trees*, INRIA, December 2011, http://hal.inria.fr/hal-00646912/en.

[52] B. FLORENT, A. KOSOWSKI, N. NISSE, I. RAPAPORT, K. SUCHAN. *Interconnection network with a shared whiteboard: Impact of (a)synchronicity on computing power*, INRIA, 2011, n[o] RR-7746, http://hal.inria.fr/inria-00627910/en.

[53] C. GAVOILLE, Q. GODFROY, L. VIENNOT. *Node-Disjoint Multipath Spanners and their Relationship with Fault-Tolerant Spanners*, INRIA, September 2011, 15 pages, http://hal.inria.fr/hal-00622915/en.

Other Publications

[54] J.-C. AVAL, P. DUCHON. *Half-turn symmetric FPLs with rare couplings and tilings of hexagons*, 2011, http://hal.inria.fr/hal-00618319/en.

[55] J.-C. BERMOND, R. KLASING, N. MORALES, S. PÉRENNES, P. REYES. *Gathering radio messages in the path*, 2011, http://hal.inria.fr/hal-00618974/en.

References in notes

[56] B. CHUN, D. CULLER, T. ROSCOE, A. BAVIER, L. PETERSON, M. WAWRZONIAK, M. BOWMAN. *PlanetLab: an overlay testbed for broad-coverage services*, in "ACM SIGCOMM Comput. Commun. Rev.", July 2003, vol. 33, p. 3–12, http://doi.acm.org/10.1145/956993.956995.

[57] F. DABEK, R. COX, F. KAASHOEK, R. MORRIS. *Vivaldi: a decentralized network coordinate system*, in "SIGCOMM '04: Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications", New York, NY, USA, ACM, 2004, p. 15–26, http://doi.acm.org/10.1145/1015467.1015471.

[58] P. FLOCCHINI, B. MANS, N. SANTORO. *Exploration of Periodically Varying Graphs*, in "ISAAC", 2009, p. 534-543.

[59] C. GAVOILLE, D. PELEG, S. PÉRENNES, R. RAZ. *Distance labeling in graphs*, in "Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms", Society for Industrial and Applied Mathematics, 2001, 219.

[60] A. LEGRAND, F. MAZOIT, M. QUINSON. *An Application-level network mapper*, LIP, ENS Lyon, France, feb 2003, nº RR-2003-09.

[61] Y. MAO, L. SAUL, J. SMITH. *IDES: An Internet Distance Estimation Service for Large Networks*, in "IEEE JSAC", Dec. 2006, vol. 24, nº 12, p. 2273 -2284.

[62] V. RAMASUBRAMANIAN, D. MALKHI, F. KUHN, M. BALAKRISHNAN, A. GUPTA, A. AKELLA. *On the treeness of internet latency and bandwidth*, in "SIGMETRICS '09: Proceedings of the eleventh international joint conference on Measurement and modeling of computer systems", New York, NY, USA, ACM, 2009, p. 61–72.