Pierre Fraigniaud (Ed.)

LNCS 3724

Distributed Computing

19th International Conference, DISC 2005
Cracow, Poland, September 2005
Proceedings

Springer

Pierre Fraigniaud (Ed.)

Distributed Computing

19th International Conference, DISC 2005
Cracow, Poland, September 26-29, 2005
Proceedings

Springer

Volume Editor

Pierre Fraigniaud
Université Paris-Sud
CNRS, LRI
91405 Orsay Cedex, France
E-mail: Pierre.Fraigniaud@lri.fr

Library of Congress Control Number: 2005932827

CR Subject Classification (1998): C.2.4, C.2.2, F.2.2, D.1.3, F.1.1, D.4.4-5

ISSN 0302-9743
ISBN-10 3-540-29163-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-29163-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11561927 06/3142 5 4 3 2 1 0

Preface

DISC, the International Symposium on Distributed Computing, is an annual forum for presentation of research on all facets of distributed computing, including the theory, design, analysis, implementation, and application of distributed systems and networks. The nineteenth edition of DISC was held on September 26–29, 2005, in Cracow, Poland.

There were 162 fifteen-page-long (in LNCS format) extended abstracts submitted to DISC this year, and this volume contains the 32 contributions selected by the Program Committee among these 162 submissions. All submitted papers were read and evaluated by at least three Program Committee members, assisted by external reviewers. The final decision regarding every paper was taken during the Program Committee meeting, which took place in Paris, July 1–2, 2005.

The Best Student Award was split and given to two papers: the paper "General Compact Labeling Schemes for Dynamic Trees", authored by Amos Korman, and the paper "Space and Step Complexity Efficient Adaptive Collect", co-authored by Yaron De Levie and Yehuda Afek.

The proceedings also include 14 two-page-long brief announcements (BA). These BAs are presentations of ongoing works for which full papers are not ready yet, or of recent results whose full description will be soon or has been recently presented in other conferences. Researchers use the brief announcement track to quickly draw the attention of the community to their experiences, insights and results from ongoing distributed computing research and projects. The BAs included in this proceedings were selected among 30 BA submissions.

DISC 2005 was organized in cooperation with Warsaw University and Jagiellonian University. The support of the University of Liverpool, INRIA, CNRS, and the University of Paris Sud (LRI) is also gratefully acknowledged. The review process and the preparation of this volume were done using CyberChairPRO.

July 2005

Pierre Fraigniaud
DISC 2005 Program Chair

Organization

DISC, the International Symposium on Distributed Computing, is an annual forum for research presentations on all facets of distributed computing. The symposium was called the International Workshop on Distributed Algorithms (WDAG) from 1985 to 1997. DISC 2005 was organized in cooperation with the European Association for Theoretical Computer Science (EATCS).

Program Chair:
Pierre Fraigniaud, CNRS and University of Paris Sud

Organizing Chair:
Dariusz Kowalski, Warsaw University and University of Liverpool

Steering Committee Chair:
Alex Shvartsman, University of Connecticut

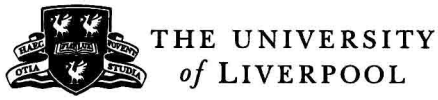
Organizing Committee:
Krzysztof Diks, Warsaw University
Kazimierz Grygiel, Warsaw University
Dariusz Kowalski, Warsaw University and University of Liverpool (Chair)
Krzysztof Szafran, Warsaw University
Marek Zaionc, Jagiellonian University

Steering Committee:
Alex Shvartsman, University of Connecticut (Chair)
Paul Vitanyi, CWI and University of Amsterdam (Vice-Chair)
Hagit Attiya, Technion
Faith Fich, University of Toronto
Pierre Fraigniaud, CNRS and University of Paris Sud
Rachid Guerraoui, EPFL
Roger Wattenhofer, ETH Zurich

Program Committee

Lenore Cowen	Tufts University
Panagiota Fatourou	University of Ioannina
Hugues Fauconnier	University of Paris VII
Pierre Fraigniaud	CNRS and University of Paris Sud (Chair)
Roy Friedman	Technion
Yuh-Jzer Joung	National Taiwan University
Dariusz Kowalski	Warsaw University and University of Liverpool
Victor Luchangco	Sun Microsystems Laboratories
Maged Michael	IBM T.J. Watson Research Center
David Peleg	Weizmann Institute
Greg Plaxton	University of Texas at Austin
Sergio Rajsbaum	National Autonomous University of Mexico
Sylvia Ratnasamy	Intel Research Laboratory
Nicola Santoro	Carleton University
Sebastiano Vigna	University of Milan
Jennifer Welch	Texas A&M University

Sponsoring Organizations



Referees

Maha Abdallah
Saurabh Agarwal
Marcos K. Aguilera
Stefan Amborg
Emmanuelle Anceaume
Filipe Araujo
Hagit Attiya
Elad Barkan
Denis Barthou
Paolo Boldi
Costas Busch
Po-An Chen
Pu-Jen Cheng
Bogdan Chlebus
Gregory Chockler
Cheng-Fu Chou
Piotr Chrzastowski
John Chuang
Bruno Codenotti
Reuven Cohen
Michele Colajanni
Massimo Coppola
Artur Czumaj
Shantanu Das
Carole Delporte
Ned Dimitrov
Shlomi Dolev
Feodor F. Dragan
Keith Duddy
Pascal Felber
Faith Ellen Fich
Matthias Fitzi
Pierre Francois
Keir Fraser
Felix C. Freiling
Eli Gafni
Jie Gao
Juan Garay
Maris Gardinariu
Leszek Gasieniec
Cyril Gavoille
Chryssis Georgiou
Bastian Pochon
Eric Goubault
Leszek Gryz
Rachid Guerraoui
Danny Hendler
Lisa Higham
Polly Huang
David Ilcinkas
Prasad Jayanti
Jehn-Ruey Jiang
Colette Johnen
Thomas Fahringer
Alexandru Jugravu
Erez Kantor
Brad Karp
Idit Keidar
David Kempe
Alex Kesselman
Gabi Kliot
Lukasz Kowalik
Piotr Krysta
Eyal Kushilevitz
Aleksandar Kuzmanovic
Mikel Larrea
Emmanuelle Lebhar
Chih-Yu Lin
Zvi Lotker
Dahlia Malkhi
Petros Maniatis
Charles Martel
Jean Mayo
Hurfin Michel
Ethan L. Miller
Neeraj Mittal
Mark Moir
Shlomo Moran
Pat Morin
Marcin Mucha
Alfredo Navarra
Robert Nikolajew
Sotiris Nikoletseas
Nicolas Nisse
Daniel Nussbaum
Alina Oprea
Katarzyna Paluch
Alessandro Panconesi
Evaggelos Papapetrou
Boaz Patt-Shamir
Andrzej Pelc
Stefan Petters
Scott Pike
Bastian Pochon
Giuseppe Prencipe
Michel Raynal
Michael Reiter
Laurent Rosaz
Sergio Ruocco
Eric Ruppert
Jared Saia
Rahul Sami
Piotr Sankowski
Elad Schiller
Michael Scott
Mordechai Shalom
Dennis Shasha
Scott Shenker
Ioannis Stamatiou
Rob van Stee
Paul Stodghill
Michal Strojnowski
Mitul Tiwari
Sebastien Tixeuil
Olga Tkachyshyn
Francisco J. Torres-Rojas
Panos Vassiliadis
Giovanni Vigna
Roman Vitenberg
Berthold Voecking
Da-Wei Wang
Hsinping Wang
Adam Warski
Mirjam Wattenhofer
Peter Widmayer
Avishai Wool
Pawel Wrzeszcz
Aleksandr Yampolskiy
Haifeng Yu
Shmuel Zaks
Xi Zhang
Aaron Zollinger

Table of Contents

Invited Talks

Regular Papers

Brief Announcements

Digital Fountains and Their Application to Informed Content Delivery over Adaptive Overlay Networks (Invited Talk)

Michael Mitzenmacher

Division of Engineering and Applied Sciences,
Harvard University,
USA

Abstract. We study how to optimize throughput of large transfers across richly connected, adaptive overlay networks, focusing on the potential of collaborative transfers between peers to supplement ongoing downloads. First, we make the case for an erasure-resilient encoding of the content, using the digital fountain paradigm. Such an approach affords reliability and a substantial degree of application-level flexibility, as it seamlessly accommodates connection migration and parallel transfers while providing resilience to packet loss. We explain the history of this paradigm, focusing on recent advances in coding that allow efficient implementations of digital fountains. We also describe our previous work showing the effectiveness of digital fountains for reliable multicast and parallel downloading.

In the setting of collaborative transfers on overlay networks, there is an additional consideration since sets of encoded symbols acquired by peers during downloads may overlap substantially. We describe a collection of useful algorithmic tools for efficient estimation, summarization, and approximate reconciliation of sets of symbols between pairs of collaborating peers, all of which keep messaging complexity and computation to a minimum. Through simulations and experiments on a prototype implementation, we demonstrate the performance benefits of our informed content delivery mechanisms.

P. Fraigniaud (Ed.): DISC 2005, LNCS 3724, p. 1, 2005.

Securing the Net: Challenges, Failures and Directions (Invited Talk)

Amir Herzberg

Department of Computer Science,
Bar Ilan University,
Israel

Abstract. The Internet is infamously insecure (fraudulent and spoofed sites, phishing and spam e-mail, viruses and Trojans, Denial of Service attacks, etc.) in spite of extensive efforts, standards, tools, and research. We will discuss the problems and the pitfalls, and outline solutions and directions for future applied and analytical research.

P. Fraigniaud (Ed.): DISC 2005, LNCS 3724, p. 2, 2005.

Coterie Availability in Sites

Flavio Junqueira and Keith Marzullo

Department of Computer Science and Engineering,
University of California, San Diego
{flavio, marzullo}@cs.ucsd.edu

Abstract. In this paper, we explore new failure models for multi-site systems, which are systems characterized by a collection of sites spread across a wide area network, each site formed by a set of computing nodes running processes. In particular, we introduce two failure models that allow sites to fail, and we use them to derive coteries. We argue that these coteries have better availability than quorums formed by a majority of processes, which are known for having best availability when process failures are independent and identically distributed. To motivate introducing site failures explicitly into a failure model, we present availability data from a production multi-site system, showing that sites are frequently unavailable. We then discuss the implementability of our abstract models, showing possibilities for obtaining these models in practice. Finally, we present evaluation results from running an implementation of the Paxos algorithm on PlanetLab using different quorum constructions. The results show that our constructions have substantially better availability and response time compared to majority coteries.

1 Introduction

There has been a proliferation of large distributed systems that support a diverse set of applications such as sensor nets, data grids, and large simulations. Such systems consist of multiple sites connected by a wide area network, where a site is a collection of computing nodes running one or more processes. The sites are often managed by different organizations, and the systems are large enough that site and process failures are common facts of life rather than rare events.

Critical services in such systems can be made highly available using replication. In data grids, for example, data sets are the most important assets, and having them available under failures of sites is very desirable. To improve availability, the well-known *quorum update* technique can be used. This technique consists of implementing a mutual exclusion mechanism by reading and writing to sets of processes that intersect (*quorums*) [7]. As another example, the Paxos protocol [16] enables the implementation of fault-tolerant state machines for asynchronous systems. Paxos is a popular choice because of its ability to produce results when a majority of replicas survive, for its feature of not producing erroneous results when failures of more than a majority (indeed, up to a complete failure) occur, and its very weak assumptions about the environment. Underlying Paxos (and other similar protocols) is the same quorum update technique.

This paper considers quorum constructions for multi-site systems. The problem area of quorums for multi-site systems is large and not well studied. We address a set of

P. Fraigniaud (Ed.): DISC 2005, LNCS 3724, pp. 3–17, 2005.

problems from this area as an early foray. We first give a failure model for multi-site systems that is simple and has intuitive appeal, and then give a second failure model that has less intuitive appeal but theoretical and practical interest. Because sites can fail, the failures of processes are not independent, and so an IID (independent, identically distributed) model is not appropriate. We define a new metric for availability that is suitable to non-IID failures, and give optimal quorum constructions for both models. We discuss the implementability of the two failure models, and discuss an experiment of running Paxos on PlanetLab [21] that gives some validation of our results.

Related work. Quorum systems have been studied for over two decades. The first algorithms based on quorums use voting [8]. Garcia-Molina and Barbara generalized the notion of voting mechanisms, and proposed the use of minimal collections of intersecting sets, or *coteries* [7]. Most of the following work (such as [15,18,20]) has concentrated on how quorums can be constructed to give good availability, load and capacity assuming relatively simple system properties (such as identical processes and independent failures) [2,3,19]. Only recently the problem of choosing quorums according to properties of the system (such as location) has attracted some attention [9,14]. Of particular interest to our current work are the constructions of [15] and [6]. In [15], Kumar proposed, to the best of our knowledge, the first hierarchical quorum construction, and showed that by doing so one can have smaller quorums. The analysis in [15], however, assumes IID failures. The work by Busca *et al.* assumes a multi-site system similar to what we assume here, and their quorum construction [6] is very similar to our *Qsite* construction. Their focus, however, was on performance. If one considers the distribution of response times from a quorum system, performance is often measured using the average or median, while availability is a property of the tail of the distribution. Thus, high performance does not necessarily imply high availability. Availability in quorum systems has been studied before [2,3,19], but we argue here that the previous metrics are not suitable for multi-site systems. A notable exception is the work by Amir and Wool [1], which evaluates several existing quorum constructions in the context of a small, real network.

A *network partition* is a failure event that leads to one set of non-faulty processes being unable to communicate with another set of non-faulty processes (and, often, vice versa). Quorum systems are asynchronous, and so a network partition is treated identically to slow-to-respond processes. Long-lasting network partitions can make it impossible to obtain a quorum. A recent paper by Yu presents a probabilistic construction that does increase availability in the face of partitions, but it assumes a uniform distribution of servers across the network [24]. In comparison, our constructions are deterministic and make no assumption about distributions of sites.

2 System Model

We consider a system of a set P of processes. The processes are partitioned into *sites* $\mathcal{B} = \{B_1, B_2, \ldots, B_k\}$, and between each pair of processes there is a bidirectional communication channel. Processes can fail by crashing, and a crashed process can recover. Similarly, a site can fail and recover. A site failure represents the loss of a key resource used by the processes in the site (such as network, power, or a storage server)

or some event that causes physical damage to the equipment on the site (such as loss of A/C); the processes in the site are all effectively crashed while the site is faulty.

Let $\mathcal{E}$ represent the executions of the system. Each execution $E \in \mathcal{E}$ is a sequence of system states. Each state $s \in E$ of an execution has an associated *failure pattern* $F(s, E) \subseteq P$, which is the set of processes that are faulty in s. If site B_i is faulty in s, then all of the processes in B_i are in $F(s, E)$. We use $NF(E, s) = P \setminus F(E, s)$ to denote the set of non-faulty processes in s. We say that a failure pattern f is valid iff $\exists E \in \mathcal{E} : \exists s \in E : f = F(E, s)$.

We use survivor sets to express valid failure patterns. Survivor sets were introduced in [11] to provide a more expressive model of process failures. Informally, a survivor set is a minimal subset of non-faulty processes. There are different ways to define survivor sets more formally: we have used probabilities [11] and have used the complement of maximal failure patterns [13]. We use the second one here. This definition does not rely on probabilities directly, although failure probabilities can be used to determine survivor sets; we discuss this point later in this paper. The definition is:

Definition 1. *Given a set of processes P, a set S is a survivor set if and only if:*[1]

$$\begin{array}{l} \bigwedge S \subseteq P \\ \bigwedge \exists E \in \mathcal{E} : \exists s \in E : S = \mathrm{NF}(E, s) \\ \bigwedge \forall p \in S : \forall E \in \mathcal{E} : \forall s \in E : S \setminus p \neq \mathrm{NF}(E, s) \end{array}$$

We use $\mathcal{S}_P$ to denote the set of survivor sets of P, and we call a pair $\langle P, \mathcal{S}_P \rangle$ a *system profile*.

We now repeat a few definitions that have appeared elsewhere and that we use in this paper. A coterie $\mathcal{Q}$ is a set of subsets of P that satisfies the following two properties [7]: 1) $\forall Q_i, Q_j \in \mathcal{Q} : Q_i \cap Q_j \neq \emptyset$; 2) $\forall Q_i, Q_j \in \mathcal{Q}, Q_i \neq Q_j : Q_i \not\subset Q_j \wedge Q_j \not\subset Q_i$. The first property is called 2-Intersection [13], and it says that quorums in a coterie pairwise intersect. This property guarantees mutual exclusion when executing operations on quorums, such as reads and writes, as every pair of quorums must have at least one process in common. The second property states that all quorums are minimal. A coterie $\mathcal{Q}$ is *dominated* if there is a coterie $\mathcal{Q}'$ such that: 1) $\mathcal{Q} \neq \mathcal{Q}'$; 2) $\forall Q \in \mathcal{Q} : \exists Q' \in \mathcal{Q}' : Q' \subseteq Q$. If no coterie dominates a coterie $\mathcal{Q}$, then we say that $\mathcal{Q}$ in *non-dominated.*

A *transversal* of a coterie is a subset of processes that intersects every quorum in the coterie. We use $\mathcal{T}(\mathcal{Q})$ to denote the set of transversals of the coterie $\mathcal{Q}$. Transversals are useful for defining the availability of a coterie: a coterie $\mathcal{Q}$ is available in a step s of some execution E if and only if $F(s, E) \notin \mathcal{T}(\mathcal{Q})$.

3 Computing Availability

The availability of coteries can be computed in various ways. One metric is *node vulnerability* which is the minimum number of nodes that, if removed, make it impossible

[1] We use the "bulleted conjunction" and the "bulleted disjunction" notation list invented in TLA$^+$ [17]. In Definition 1, the list corresponds to the conjunction of the statements to the right of the "$\bigwedge$" marks.

to obtain a quorum [3]. A similar metric, *edge vulnerability*, counts the minimum number of channels whose removal makes it impossible to obtain a quorum (no connected component contains a quorum). Both of these metrics are appropriate when failures are independent and identically distributed (IID) because they measure the minimum number of failures necessary to halt the system. They are not necessarily good metrics for multi-site systems. Consider the following three-site system in which a survivor set is the union of majorities of processes in a site for some majority of sites:[2]

$$\begin{aligned} P &= \{a_1, a_2, a_3, b_1, b_2, b_3, c_1, c_2, c_3\} \\ \mathcal{B} &= \{a_1a_2a_3, b_1b_2b_3, c_1c_2c_3\} \\ \mathcal{S}_P &= \{a_ia_jb_lb_m : i,j,l,m \in \{1,2,3\} \wedge\ i \neq j \wedge l \neq m\} \\ &\quad \cup \{a_ia_jc_lc_m : i,j,l,m \in \{1,2,3\} \wedge\ i \neq j \wedge l \neq m\} \\ &\quad \cup \{b_ib_jc_lc_m : i,j,l,m \in \{1,2,3\} \wedge\ i \neq j \wedge l \neq m\} \end{aligned}$$

From our system model, processes are pairwise connected. According to the results in [3], the best strategy for both node and edge vulnerability is then to use quorums formed of majorities, which for this system is any subset of five processes. By definition, for every $S \in \mathcal{S}_P$, there is some step s of some execution $E \in \mathcal{E}$ such that $S = F(s, E)$, where $\mathcal{E}$ is the set of executions of $\langle P, \mathcal{S}_P \rangle$. As P contains nine processes and every $S \in \mathcal{S}_P$ contains four processes, there are five faulty processes in such a step, and hence no majority quorum can be obtained. If one uses $\mathcal{S}_P$ as a coterie, however, then there is one quorum available in every step, by construction. $\mathcal{S}_P$ has therefore better availability than the majority construction.

An alternative to node and edge vulnerability is, given probabilities of failures, to directly compute the probability of the most likely failure patterns that make it impossible to obtain a quorum. Probability models, however, can become quite complex when failures are not IID. To avoid such complexity, we use a different counting metric: the number of survivor sets that allow a quorum to be obtained. More carefully,

Definition 2. *Let $\langle P, \mathcal{S}_P \rangle$ be a system profile and $\mathcal{Q}$ be a coterie over P. The availability of $\mathcal{Q}$ is given by: $\mathcal{A}(\mathcal{Q}) = |\{S : S \in \mathcal{S}_P \wedge S \notin \mathcal{T}(\mathcal{Q})\}|$*

A coterie $\mathcal{Q}$ *covers* a survivor set S if there is a quorum $Q \in \mathcal{Q}$ such that $Q \subseteq S$. By the definition, $\mathcal{A}(\mathcal{Q})$ is hence the number of survivor sets that $\mathcal{Q}$ covers.

This is a good metric because in every step s of an execution E, there is at least one survivor set in $\mathcal{S}_P$ that does not intersect $F(E, s)$. If a coterie allows a quorum to be obtained for more survivor sets, then this coterie is available during more steps. As node vulnerability and edge vulnerability, $\mathcal{A}()$ is a deterministic metric and as such has a similar limitation with respect to probabilities. If we assign probabilities of failure to subsets of processes, then our metric may lead to wrong conclusions, as there might be higher available coteries that include discarded survivor sets. For the constructions and examples we discuss in this paper, however, using this metric gives us coteries with optimal availability.

If a coterie $\mathcal{Q}$ is dominated, then by definition there is some other coterie $\mathcal{Q}'$ that dominates $\mathcal{Q}$. Under reasonable assumptions, the availability of $\mathcal{Q}'$ is at least as high as

[2] We use $x_1x_2 \ldots x_n$ as a short notation for the set $\{x_1, x_2, \ldots, x_n\}$.