Distributed Computing

Proposal for a MAP-I optional curricular unit on Theory and Foundations

Francisco Moura Rui Oliveira Paulo Almeida
Carlos Baquero José Pereira
CCTC-DI – U. Minho

Abstract

This document describes a Ph.D. level course, corresponding to a Curricular Unit covering the Theory of Distributed Computing, currently running in the joint MAP-i doctoral programme in Informatics, organized by three Portuguese universities (Minho, Aveiro and Porto).

In the 2007-2008 edition of the MAP-i programme this course is being taught to 12 full time students. Lecture material for this edition can be found in the `slides` folder at `http://gsd.di.uminho.pt/teaching/DC/2007/`. Also in the 2007-2008 edition 6 MAP-i students initiated their PhD studies in the area of Large Scale Distributed Systems.

1 Context

1.1 Overview

Distributed computing refers to algorithms running on a set of machines connected by a network. Its importance has increased as computation migrated from monolithic mainframes to decentralized structures connected by the internet. Examples of distributed systems appear in many areas such as telecommunication, web applications, distributed data processing and massively multiplayer games.

While a distributed system can be built with redundancy (e.g. with replicated components) so as to provide availability in the presence of faults, this can be difficult to achieve if the software is programmed in an informal way, without strong theoretical foundations.

Concurrency, which occurs naturally in a distributed system, is already a difficult subject. On top of that, the problems that arise in asynchronous distributed systems subject to processor or link failures are difficult to comprehend. Even knowing what is possible to achieve may be non-intuitive. This means that people may waste years trying to solve an impossible problem; or they may build a software toolkit or a middleware platform (which will be used by many others) that will malfunction, or behave unpredictably in a non-repeatable and incomprehensible way.

1.2 Aims

This course aims at providing the theoretical foundations of distributed systems. It is targeted to graduate students and researchers wishing to advance the state-of-the-art in distributed systems. The course is technology agnostic and the abstractions presented are independent of any given technology. In fact, no technologies will be presented at all.

Although theoretical in nature, the course will also benefit students doing a more practical research. For example, database replication can be based on group communication protocols for which it is important to understand the agreement problem and algorithms.

The course focuses on formal models (e.g. I/O automata), abstractions (e.g. logical time), problems (e.g. agreement) and algorithms to solve them. It also focuses on impossibility results (e.g. the impossibility of fault-tolerant consensus in asynchronous networks).

1.3 Related Courses

From other graduate-level courses which are similar to this one, we highlight the following:

- *"Distributed Algorithms"* at the MIT, by Nancy Lynch.
- *"Theory of Distributed Computing"* at the EPFL, by Rashid Guerraoui.
- *"Advanced Operating Systems and Distributed Systems"* at CMU, by David Andersen.

In the CMU Computer Science Department course offers in Fall 2007, we refer the course 15-712 on *"Advanced Operating Systems and Distributed Systems"* by David Andersen. Altough our focus is different, with less systems component.

2 Objectives

The goal of this course is to provide an advanced theoretical background on distributed computing, addressing fundamental problems, models, algorithms and results. This provides a solid foundation for research on distributed computing in the context of a graduate program.

3 Learning Outcomes

Upon successful completion of this course, students should be able to:

- build formal models of distributed system;
- differentiate between synchronous and asynchronous models;
- understand the assumptions and limitations underlying models of distributed systems;
- describe the more relevant problems in distributed systems;

- reason about distributed algorithms;
- design new distributed algorithms;
- invoke impossibility results to avoid wasting time trying to solve an unsolvable problem;
- prove impossibility results.

4 Topics

- Synchronous networks: (Weeks 1-3)
  - Formal model (lockstep rounds) and proof methods
  - Basic algorithms: Leader Election, Spanning Trees
- Agreement in synchronous networks: (Weeks 4-5)
  - The abstract consensus problem
  - Agreement with process and link failures
  - Byzantine agreement
- Asynchronous networks: (Weeks 6-7)
  - Formal models (I/O automata,TLA) and proof methods
  - Basic algorithms: (revisited)
- Logical time in asynchronous networks: (Weeks 8-9)
  - Causality, real time and logical time
  - Vector clocks and version vectors
  - Stable properties: (Week 10)
    - Distributed termination
    - Global snapshots
    - Deadlock detection
- Agreement in asynchronous networks: (Week 11-14)
  - Impossibility of fault-tolerant consensus
  - Failure detectors and indulgence
  - Unreliable channels
  - Agreement problems:
    - Distributed commit
    - Atomic broadcast

5 Format

The course is organized around formal lectures, 3 hours per week, during one semester. The course is credited with 5 ECTS in the European Credit Transfer and Accumulation System. Some lecture time (around 1/4) is used for recitation, where a given student will have to present and defend a previously assigned research paper, leading to a discussion involving the other students.

In the 2007-2008 edition the following papers were discussed:

- A New Approach to Proving the Correctness of Multiprocess Programs. by Leslie Lamport, 1979.
- A Distributed Algorithm for Minimum-Weight Spanning Trees. by G. Gallager,P. A. Humblet, P. M. Spira, 1983.
- Using Time Instead of Timeout for Fault-Tolerant Distributed Systems. by Leslie Lamport, 1984.
- Reaching Agreement in the Presence of Faults. by M. Pease, R. Shostak, L. Lamport, 1980.
- Model Checking TLA+ Specifications. by Y. Yu, P, Manolios, L. Lamport, 1999.
- Reliable Communication over Unreliable Channels. Y. Afek, H. Attiya, A. Fekete, M. Fischer, N. Lynch, Y. Mansour, D. Wang, L. Zuck, 1994.
- Proving Safety Properties of an Aircraft Landing Protocol Using I/O Automata and the PVS Theorem Prover: A Case Study. S. Umeno, N. Lynch, 2006.
- Optimal Time Self Stabilization in Dynamic Systems. S. Dolev, 1993.
- Computation in Networks of Passively Mobile Finite-State Sensors. D. Angluin, J. Aspnesm, Z. Diamadi, M. Fischer, R. Peralta, 2004.
- Renaming in an Asynchronous Environment. H. Attiya, A. Bar-Novm D. Dolev, D. Peleg, R. Reischuk. 1990.

Lectures notes of the 2007/08 edition of the CU can be found at:
http://gsd.di.uminho.pt/teaching/DC/2007/slides/

6 Grading

The grading is based on two components:

- continuous grading along the semester, involving recitations and research paper analysis;
- individual monograph at the end of the course.

7 References

[1] N. Lynch. *Distributed Algorithms.* Morgan-Kaufmann, 1996.

[2] L. Lamport. Time, clocks and the ordering of events in a distributed system. *Communication of the ACM* 21, no.7, July 1978.

[3] F. Mattern. Virtual time and global states of distributed systems. In Z. Yang, and T. Marsland (eds.) *Global States and Time in Distributed Systems*, IEEE, 1994.

[4] L. Lamport. The part-time parliament. *Transactions on Computer Systems* 16, no.2, May 1989.

[5] T. Chandra and S. Toueg. Unreliable failure detectors for reliable distributed systems. *Journal of the ACM*, Volume 43, Issue 2, March 1996.

[6] T. Chandra, V. Hadzilacos, and S. Toueg. The weakest failure detector for solving consensus. In *Proceedings of the Annual ACM Symposium on Principles of Distributed Computing*, 1992.

[7] P. Dutta and R. Guerraoui. The inherent price of indulgence. *Distributed Computing* 18(1), Springer, 2005.

[8] J. B. Almeida, P. S. Almeida, C. Baquero. Bounded version vectors. In Rachid Guerraoui, editor, Proceedings of DISC 2004: 18th international symposium on distributed computing, number 3274 in LNCS, pages 102–116. 2004. Springer Verlag.

[9] J. Pereira, R. Oliveira. The mutable consensus protocol. Proc. 23rd international symposium on reliable distributed systems, pages 218-227. Florianpolis, Brazil, 2004. IEEE, IEEE Computer Society.

[10] V. Hadzilacos, S. Toueg, Fault-tolerant broadcasts and related problems, In *Distributed systems (2nd Ed.)*, ACM Press/Addison-Wesley Publishing Co., New York, NY, 1993.

A Research Background

The teaching team consists of members of the Distributed Systems Group (GSD) of the Informatics Department of Minho University. The team has considerable experience of teaching and research in distributed systems. The GSD's research activities have been focused on two main areas: dependability on large-scale networks and weakly consistent data replication.

Dependability on large-scale networks The work has been focused on fundamental and applied research on models, algorithms and tools enabling to build dependable services and applications on large-scale networks. The approaches being pursued depart from solid ground on fault-tolerant distributed coordination and group communication protocols and explore novel ideas and intuitions deemed to adapt well to large-scale networks. Current research, namely on optimistic and semantically reliable group protocols and on models of partial replication, is strongly supported by ongoing projects and represent the basis of future research.

Weakly consistent data replication In line with our past research in distributed and mobile filesystems, we have been conducting research in theoretical and systems support to weak consistency file replication and versioning. Recent work produced a versioning/replication tool that targets partitioned filesystems, and a theoretical model for dependency tracking in autonomous systems with a dynamically evolving number of replicas. Current and future research in this line targets new developments in the dynamic replication theory, the development of a novel approach to dependency tracking in fixed replication sets, and the use of dependency tracking for mutable data management in global peer-to-peer systems.

Related Publications

[1] J. Pereira, R. Oliveira. The mutable consensus protocol. Proc. 23rd international symposium on reliable distributed systems, pages 218-227. Florianopolis, Brazil, 2004. IEEE, IEEE Computer Society.

[2] J. B. Almeida, P. S. Almeida, C. Baquero. Bounded version vectors. In Rachid Guerraoui, editor, Proceedings of DISC 2004: 18th international symposium on distributed computing, number 3274 in LNCS, pages 102–116. 2004. Springer Verlag.

[3] A. Sousa, J. Pereira, F. Moura, R. Oliveira. Optimistic total order in wide area networks. Proc. 21st ieee symposium on reliable distributed systems, pages 190-199. 2002. IEEE CS.

[4] P. S. Almeida, C. Baquero, V. Fonte. Version stamps – decentralized version vectors. Proceedings of the 22nd international conference on distributed computing systems (ICDCS), pages 544–551. 2002. IEEE Computer Society.

[5] P. Almeida, C. Baquero, Vi. Fonte: Panasync: dependency tracking among file copies. In *ACM SIGOPS European Workshop*, 2000.

[6] J. Pereira, L. Rodrigues and R. Oliveira. Semantically reliable broadcast: Sustaining high throughput in reliable distributed systems. In *Concurrency in Dependable Computing*, Paul Ezhilchelvan and Alexander Romanovsky (eds.), Chapter 10, Kluwer Academic Publishers, 2002.

[7] J. Pereira, L. Rodrigues and R. Oliveira. Semantically reliable multicast: Definition, implementation and performance Evaluation. *IEEE Transactions on Computers, Special Issue on Reliable Distributed Systems*, 2003.

Related Projects

SHIFT: Group Communication with Differentiated Messages
Funded by FCT POSI/CHS/33792/1999, EUR 20.000, 2000-2003
More information at *http://shift.di.fc.ul.pt.*

ESCADA: Fault-Tolerant Scalable Distributed Databases
Funded by FCT POSI/CHS/33792/1999, EUR 55.900, 2000-2003
More information at *http://escada.lsd.di.uminho.pt.*

StrongRep: Strongly Consistent Replicated Databases in Geographically Large-Scale Systems
Funded by FCT POSI/CHS/41285/2001, EUR 44.000, 2002-2005
More information at *http://strongrep.lsd.di.uminho.pt.*

GORDA: Open Replication of Databases
Funded by FP6 IST 004758, EUR 1.250.000, 2004-
More information at *http://gorda.di.uminho.pt.*

FEW: Files EveryWhere
Funded by FCT POSI/EIA/59064/2004, EUR 83.000, 2005-
More information at *http://asc.di.fct.unl.pt/few/.*

B Instructors

Francisco Soares de Moura is associate professor at the Department of Informatics of Minho University, and a researcher at CCTC in the area of Distributed Systems.

His research interests include operating systems and distributed systems, namely the areas of file and database replication. He is a member of the research team of IST GORDA project on open replication of databases and previously led the FCT funded Mobisnap project on databases and mobility.

Francisco Moura has lectured Distributed Systems at 1st and 2nd Cycles levels.

For more information, visit *http://gsd.di.uminho.pt/members/fsm.*

Rui Carlos Oliveira is associate professor at the Department of Informatics of Minho University, and a researcher at CCTC in the area of Distributed Systems.

His research interests are in dependable distributed systems and cover consistent database replication, distributed agreement problems and gossip-based communication. He is the coordinator of the IST GORDA project on open

replication of databases and previously led two related, FCT funded, research projects ESCADA and StrongRep. He currently supervises two Ph.D. students with projects on the topic of database replication.

Rui Carlos Oliveira lectures Dependable Distributed Systems at 1st and 2nd Cycles levels.

For more information, visit *http://gsd.di.uminho.pt/members/rco*.

Paulo Sérgio Almeida is a lecturer at the Department of Informatics of Minho University, and a researcher member of CCTC.

His scientific research activities are centered in distributed systems. The two main topics of research have been time/version stamping mechanisms and distributed data aggregation algorithms. The mains results of late have been Dynamic Version Stamps, Bounded Version Vectors and Scalable Bloom Filters.

For more information, visit *http://gsd.di.uminho.pt/members/psa*.

Carlos Baquero is a lecturer at the Department of Informatics of Minho University, and a researcher member of CCTC in the area of Distributed Systems.

His research interests are focused on distributed systems, in particular in causality tracking, peer-to-peer systems and mobile computing. Recent research is focused on highly dynamic distributed systems, both in internet P2P settings and in mobile and sensor networks. He his the local coordinator of FCT founded project FEW, where optimistic replication of file-systems is addressed. He currently supervises two Ph.D. and two M.Sc. projects in topics related to data dependency tracking, P2P indexing and distributed data aggregation.

Carlos Baquero lectures Distributed Systems and Data Management in Mobile Computing at the 2nd Cycle level.

For more information, visit *http://gsd.di.uminho.pt/members/cbm*.

José Orlando Pereira is a lecturer at the Department of Informatics of Minho University, and a researcher at CCTC (area of *Distributed Systems*).

His research interests are in dependable distributed systems and are split between gossip-based communication and database replication. Regarding gossip-based communication, he leads the P-SON research project, which focuses on gossip-based communication protocols. Regarding database replication, he is the Technical Manager of the IST GORDA project. He is currently supervising a Ph.D. project on *Database Replication on Shared Storage Clusters*.

José Orlando Pereira lectures Distributed Systems at the 2nd Cycle level to the M.Sc. program on Biomedical Engineering, with emphasis multi-tiered architectures, and to the M.Sc. program on Mobile Computing, with emphasis on message passing algorithms and systems.

For more information, visit *http://gsd.di.uminho.pt/members/jop*.

Rui Oliveira coordinates the Distributed Computing curricular unit.