Distributed Computing

Proposal for a MAP-I optional curricular unit on Theory and Foundations
(2012-2013 edition)

José Pereira Paulo Almeida Pedro Souto Rui Oliveira

University of Minho and University of Porto

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). This course has been submitted for accreditaion by the CMU doctoral programme in August 2008 and awaits for the process' outcome.

This course has been taught in all three previous editions of the MAP-i programme. The current proposal for the 2012-2013 has the same structure and content of last year's edition, emphasizing critical systems with deeper study of agreement problems, state-of-the-art formal modeling of timed asynchronous networks, and real-time systems. It has also the same proponent team that encompasses researchers from two universities, Minho and Porto. Lecture material of the previous edition can be found in the course's web page.[1]

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

[1] `http://gsd.di.uminho.pt/teaching/DC/2011/`

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 Rachid Guerraoui.
- *"Advanced Operating Systems and Distributed Systems"* at CMU, by David Andersen.

In the CMU Computer Science Department course offers in Spring 2011, 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 systems;

- differentiate between synchronous, asynchronous and hybrid 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 1-2)
  - Formal model (lockstep rounds) and proof methods
  - Basic algorithms: Leader Election
  - Agreement with process and link failures
- Asynchronous networks: (Weeks 3-4)
  - Formal models (I/O automata) and proof methods
  - Basic algorithms: (revisited)
  - Logical time and State-machine simulation
- Agreement in asynchronous networks: (Weeks 5-6)
  - Impossibility of fault-tolerant consensus
  - Failure Detectors and Indulgence
  - Unreliable communication channels
  - Agreement problems: Distributed commit, Atomic broadcast
- Timed/Hybrid Asynchronous networks: (Week 7)
  - Formal model (timed I/O automata) and proof methods
  - Clock synchronization and Failure Detectors implementation
  - Timeliness and Real-time guarantees

5 Format

The course is organized around formal lectures, 3 hours per week, during half 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 previous editions 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. Aspnes, 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 2011/2012 edition of the Curricular Unit can be found at:
http://gsd.di.uminho.pt/teaching/DC/2011/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. Florianópolis, 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 proponent team consists of members of the Informatics Department of University of Minho, of the Mechanical Engineering Department and Informatics Engineering Department of the University of Porto.

The team has considerable experience of teaching and research in distributed systems, with strong emphasis on dependability and critical systems.

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.

Safety-critical networks This work has focused both at the level of the communications network and at the level of services. With respect to communications network, we have developed communication protocols to ensure reliable and real-time communication. Current research on this topic focuses on wireless communication. With respect to services, we have been working on algorithms for core services, such as group membership and reliable broadcast, that facilitate the development of safety critical applications.

Selected Related Publications

[1] A. Sousa, J. Pereira, F. Moura, R. Oliveira. Optimistic total order in wide area networks. In Proc. of the 21st IEEE International Symposium on Reliable Distributed Systems. 2002. IEEE Computer Society.

[2] 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.

[3] 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.

[4] J. Pereira, R. Oliveira. The Mutable Consensus Protocol. In Proc. of 23rd IEEE International Symposium on Reliable Distributed Systems. 2004. IEEE Computer Society.

[5] J. Almeida, P. Almeida, C. Baquero. Bounded version vectors. In Proc. of 18th Annual Conference on Distributed Computing, LNCS 3274, pages 102-116. 2004. Springer Verlag.

[6] V. Rosset, P. F. Souto, F. Vasques. A group membership protocol for communications systems with both static and dynamic scheduling. n Proc. of 6th IEEE International Workshop on Factory Communication Systems. 2006. IEEE Computer Society.

[7] V. Rosset, P. F. Souto, F. Vasques. Formal Verification of a Group Membership Protocol Using Model Checking. In Proc. of 9th International Symposium onthe Distributed Objects, Middleware, and Applications, LNCS 4803, pages 471-488. 2007. Springer Verlag.

Related Projects

GORDA: Open Replication of Databases
Funded by FP6 IST 004758, EUR 1.250.000, 2004-2008
More information at *http://gorda.di.uminho.pt*.

Safe-DuST: Services for Safety Critical Applications for Dual Scheduled TDMA Networks
Funded by FCT POSI/EIA/74313/2006, EUR 60.000, 2008-2010

ReD: Resilient Database Clusters
Funded by FCT PDTC/EIA-EIA/109044/2008, EUR 123.632, 2010-

CASTOR: Causality Tracking for Optimistic Replication in Dynamic Distributed Systems
Funded by FCT PTDC/EIA-EIA/104022/2008, EUR 66144, 2010-

DHT-Mesh: DHT-based services for increasing the Scalability of Highly available Wireless Mesh Networks
Funded by FCT PTDC/EEA-TEL/104185/2008, EUR 123.720,00, 2011-

CumuloNimbo: A Highly Scalable Transactional Multi-Tier Platform as a Service
Funded by FP7 ICT 257993, EUR 3.000.000, 2010-

B Proponents and Instructors

The current proposal is coordinated by Paulo Sérgio Almeida and supported by faculty from University of Minho and University of Porto. The unit's coordinator is responsible for scheduling sessions, establishing and conducting grading procedures, and for all administrative contact with students.

José Orlando Pereira is an Assistant Professor at the Department of Informatics of University of Minho, and a researcher at HASLab / INESC TEC (area of *Large Scale Distributed Systems*).

His research interests are in dependable distributed systems and are split between large scale reliable group communication and database replication. He was the Technical Manager of the IST GORDA project and the PI of the P-SON project on large scale reliable group communication. Currently, he coordinates the ReD project on resilient database clusters. He is currently supervising several Ph.D. projects on dependable distributed systems.

José Orlando Pereira lectures the Foundations of Distributed Systems and Transactional Systems components of the Distributed Systems UCE at the 2nd Cycle level to the Masters on Informatics at the U. Minho.

Paulo Sérgio Almeida is a Lecturer at the Department of Informatics of University of Minho, and a researcher member of HASLab / INESC TEC.

His scientific research activities are centered in distributed systems. The two main topics of research have been causality tracking mechanisms and distributed data aggregation algorithms. The main results achieved have been the "Flow Updating" distributed aggregation technique, the "Interval Tree Clocks" and "Bounded Version Vectors" logical clocks, and "Scalable Bloom Filters".

Pedro Ferreira do Souto is Assistant Professor at the Informatic Engineering Department of University of Porto, and a researcher at Instituto de Sistemas e Robótica-Porto.

His research interests are in the field of distributed systems, in particular in fault-tolerance and dependability and the application of formal methods for their assessment. He currently participates in the DHT-Mesh project, focusing on scalability issues. He was the principal investigator of the Safe-DUST project, funded by the FCT, that investigated novel fault-tolerant algorithms for dual-scheduled TDMA networks and the application of model checking to their evaluation.

Pedro Ferreira do Souto teaches a Distributed Systems course at the 2nd Cycle level at the Faculdade de Engenharia da Universidade do Porto.

Rui Oliveira is Associate Professor at the Department of Informatics of University of Minho, and a researcher at HASLab / INESC TEC 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 currently involved in the CumuloNimbo project. He was 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 Oliveira lectures Dependable Distributed Systems at 1st and 2nd Cycles levels.