ACM SIGACT News Distributed Computing Column 32

The Year in Review

Idit Keidar
Dept. of Electrical Engineering, Technion
Haifa, 32000, Israel
idish@ee.technion.ac.il

This column overviews the main events related to distributed computing in 2008. I begin with a citation of this year's Dijkstra Prize in Distributed Computing, which was awarded (in PODC'08) to Baruch Awerbuch and David Peleg for their paper "Sparse Partitions" published in FOCS in 1990. I include a reprint of the award statement, provided by Gadi Taubenfeld, the head of this year's award committee.

I then proceed with reviews of PODC– the ACM Symposium on Principles of Distributed Computing– and its European counterpart, DISC– the International Symposium on DIStributed Computing. I decided to also include a review of SPAA– the ACM Symposium on Parallelism in Algorithms and Architectures– since the boundaries between distributed and parallel computing are quite blurred these days: SPAA increasingly deals with classical distributed computing topics such as network (and graph) algorithms, while PODC and DISC continue to deal extensively with concurrent and shared memory-based computing, which arises in parallel architectures.

To review these three events, I invited students who have won Best Paper and Best Student Paper Awards in the respective conferences (PODC, DISC, and SPAA). I also asked them to include descriptions of their winning papers in their reviews.

The review of PODC is by Armando Castañeda from Universidad Nacional Autónoma de México, who won the Best Student Paper Award for his paper "New Combinatorial Topology Upper and Lower Bounds for Renaming", co-authored with his advisor Sergio Rajsbaum. This remarkable paper refutes a long-standing lower bound result on wait-free renaming, which was proven in a number of papers, including a Gödel Award winner. More specifically, the paper shows that the previously proven lower bound does not hold for all values of n, where n is the number of processes, while for other values of n it does hold. (See more below). PODC was co-located with CONCUR this year, and featured a special symposium to celebrate the contributions of Nancy Lynch in light of her sixtieth birthday. See more on the celebration (and a picture) in Armando's review of PODC below.

DISC is reviewed by the winners of two awards: Robert Danek and Wojciech Golab from the University of Toronto, who won the Best Paper Award for their paper "Closing the Complexity Gap Between FCFS Mutual Exclusion and Mutual Exclusion", and Wojciech Wawrzyniak from Adam Mickiewicz University in Poland, who won the Best Student Paper Award for the paper "Fast distributed approximations in planar graphs" he co-authored with M. Hańćkowiak and A. Czygrinow. The winning papers are discussed in the review below.

The review of SPAA is by Zvika Guz from the Technion, winner of SPAA's Best Paper Award, for the paper "Utilizing Shared Data in Chip Multiprocessors with the Nahalal Architecture", which he co-authored with Yours Truly, Avinoam Kolodny, and Uri Weiser, and discusses in his review of SPAA below.

Of course, these reviews do not cover all the interesting events where distributed computing is studied; distributed computing papers appear in numerous additional conferences. While it is clearly impossible to cover all the relevant venues in this column, I do try to provide a taste of this wide variety. For example, a recent column[1], has surveyed distributed computing research in systems conferences (SOSP and OSDI). In the current column, I chose to highlight another community that dabbles in distributed computing– the dependability community. To this end, I include a review by Gabi (Gabriel) Kliot of distributed computing papers in DSN 2008– the International Conference on Dependable Systems and Networks.

In all four reviews, you will find fun information about the venues, as well as technical content. Many thanks to Gadi, Armando, Robert, Wojciech, Wojciech, Zvika and Gabi for their colorful contributions!

Call for contributions: I welcome suggestions for material to include in this column, including news, reviews, open problems, tutorials and surveys, either exposing the community to new and interesting topics, or providing new insight on well-studied topics by organizing them in new ways.

The 2008 Edsger W. Dijkstra Prize in Distributed Computing

> The 2008 Edsger W. Dijkstra Prize in Distributed Computing was awarded, in PODC'08, to the paper "Sparse Partitions" by Baruch Awerbuch and David Peleg.

The Edsger W. Dijkstra Prize in Distributed Computing is awarded for an outstanding paper on the principles of distributed computing, whose significance and impact on the theory and/or practice of distributed computing has been evident for at least a decade.

The Dijkstra Award Committee has selected Baruch Awerbuch and David Peleg as the recipients of this year Edsger W. Dijkstra Prize in Distributed Computing. The prize is given to them for their outstanding paper: "Sparse Partitions" published in the proceedings of the 31st Annual Symposium on Foundations of Computer Science, pages 503–513, 1990.

The Sparse Partitions paper by Awerbuch and Peleg signified the coming-to-age of the area of distributed network algorithms. In this work, a line of research that started with Awerbuch's synchronizer and Peleg's spanner has culminated in this ground breaking paper that has had a profound impact on algorithmic research in distributed computing and in graph algorithms in general.

The paper presents concrete definitions of the intuitive concepts of locality and load, and gives surprisingly effective constructions to trade them off. The fundamental technical contribution in the paper is the

[1] Column 30, SIGACT News 39(2).

algorithm of coarsening, which takes, as input, a decomposition of the graph to possibly overlapping components, and generates a new decomposition whose locality is slightly worse, but whose load is far better. The desired balance between locality and load is controlled by a parameter provided by the user. While many other underlying ideas were present in prior work of Awerbuch and Peleg (separately), in the Sparse Partitions paper these ideas have come together, with a unified view, resulting in a new powerful toolkit that is indispensable for all workers in the field.

The magnitude of the progress achieved by the new techniques was immediately recognized, and its implications spawn much research to this day. In the Sparse Partitions paper itself, the authors improve on the best known results for two central problems of network algorithms, and many other applications of the results followed, quite a few of them in applications that were visionary at their time. To mention just a few, these include computation of compact routing tables and location services of mobile users (in the original paper), dramatically more efficient synchronizers, effective peer-to-peer network design, and scheduling in grid-like computing models. Besides these applications of the results, the paper can be viewed as one of the important triggers to much of the fundamental research that was dedicated to exploring other variants of the basic concepts, including the notions of bounded-growth graphs, tree metrics, general and geometric spanners.

It is interesting to view the Sparse Partitions paper in the historical context. The area of Network Algorithms has its roots in classical graph algorithms. Distributed algorithms have proved to be an algorithmically rich field with the Minimum Spanning Tree paper of Gallager, Humblet and Spira. Motivated by the asynchronous nature of distributed systems, Awerbuch invented the concept of a synchronizer. Peleg, coming from the graph theoretic direction, generalized the notion of spanning tree and invented the concept of spanners. In the Sparse Partitions paper, the additional ingredient of load was added to the combination, yielding a powerful conceptual and algorithmic tool. The results superseded the best known results for classical graph algorithms, thus showing the maturity of the field, which closed a circle by becoming a leading source for graph algorithms of any kind.

Award Committee 2008:

Yehuda Afek	Tel-Aviv Univ.
Faith Ellen	University of Toronto
Shay Kutten	Technion
Boaz Patt-Shamir	Tel-Aviv Univ.
Sergio Rajsbaum	UNAM
Gadi Taubenfeld, Chair	IDC

A Review of PODC 2008

Armando Castañeda
Instituto de Matemáticas
Universidad Nacional Autónoma de México
acastanedar@uxmcc2.iimas.unam.mx

The *ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing* (PODC) is the most important annual conference in North America dedicated to Distributed Computing. It focuses on research in the theory, design, specification and implementation of distributed systems. PODC covers every aspect of distributed computing, theoretical or experimental, and its goal is to understand the principles underlying distributed computing. The 27th version of the ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing was held on 18–21 August, 2008, in the beautiful city of Toronto, Canada, one of the most exciting cities in the world due to its cultural, entertainment and sports activities. This was a homecoming for PODC since the first one was held on Toronto in 1982, and fifteen PODCs have been held in Canada, including this one.

PODC 2008 was collocated with the 19th International Conference on Concurrency Theory (CONCUR) and various workshops:

1. Workshop on Approximate Behavioral Equivalences.
2. 5th SIGACT-SIGOPS Workshop on Foundations of Mobile Computing.
3. 15th International Workshop on Expressiveness in Concurrency.
4. Workshop on Formal Methods for Wireless Systems.
5. 10th International Workshop on Verification of Infinite-State Systems.
6. 6th International Workshop on Security Issues in Concurrency.

This PODC was a big one; 132 regular papers and 55 brief announcements were submitted and 40 regular papers and 44 brief announcements were selected by the program committee. The regular presentations were split into nine sessions according to their field of research. And the brief announcements were split into eight sessions, and some of them were presented in parallel in two different session rooms. All talks took place at the Bahen Center at the University of Toronto. Also, PODC 2008 featured three invited talks:

1. *Accountability for Distributed Systems*, Peter Druschel (Cornell University).

2. *Beyond Nash Equilibrium: Solution Concepts of the 21st Century*, Joseph Halpern (University of Massachusetts - Amherst).

3. *The Internet is Flat: A brief history of networking over the next ten years*, Don Towsley (Max Planck Institute for Software Systems).

I was happy to present the paper "New Combinatorial Topology Upper and Lower Bounds for Renaming", co-authored with my advisor Sergio Rajsbaum, which won the Best Student Paper Award. In the renaming task, $n + 1$ processes start with unique input names from a large space and must choose unique output names taken from a smaller name space, namely $0, 1, \ldots, K$. To rule out trivial solutions, a protocol must be anonymous: the value chosen by a process can depend on its input name and on the execution, but not on the specific process id. Attiya et al. showed in 1990 that renaming has a wait-free solution when $K \geq 2n$. Several proofs of a lower bound stating that no such protocol exists when $K < 2n$ have been published. In this paper we prove that, for certain exceptional values of n, this lower bound is incorrect, exhibiting a wait-free renaming protocol for $K = 2n-1$. For the other values of n, we present the first completely combinatorial lower bound proof stating that no such protocol exists when $K < 2n$. More precisely, our main theorem states that there exists a wait-free renaming protocol for $K < 2n$ if and only if the set of integers $\{\binom{n+1}{i+1} : 0 \leq i \leq \lfloor \frac{n-1}{2} \rfloor\}$ are relatively prime. Thus, such protocol exists for six processes, and not for less. Examples of exceptional numbers, are $n = 5, 9, 13$. The proof of the theorem uses combinatorial topology techniques, both for the lower bound and to derive the renaming protocol.

PODC introduced a new award this year– The Best Presentation Award. The award was split between Prasad Jayanti, for his presentation "Every problem has a weakest failure detector", and Thomas Moscibroda, for his presentation of the paper "Distributed Order Scheduling and its Application to Multi-Core DRAM Controllers".

Figure 1: Faith Ellen and Prasad Jayanti talking on coffee break.

The Edsger W. Dijkstra Prize is given for outstanding papers on the principles of distributed computing, whose impact on the theory or practice has been evident for at least ten years. This year the Dijkstra Prize was awarded to Baruch Awerbuch and David Peleg for their seminal 1990 paper *Sparse Partitions*, and it was presented in PODC 2008.

Figure 2: On the way to Monday's lunch.

A very special aspect of PODC 2008, is that it featured the mini symposium *Nancy Lynch Celebration: Sixty and Beyond*, in celebration of Nancy Lynch sixtieth birthday. This symposium included various invited talks illustrating the impact and importance of Nancy's work through all this time. The invited talks were the following:

1. *Evolution of Distributed Computing Theory: From Concurrency to Networks and Beyond*, Michael Fischer (Yale University).

2. *A World of (Im)possibilities*, Hagit Attiya (Technion) and Jennifer Welch (Texas A&M University).

3. *The Power of Simulation Relations*, Roberto Segala (Universitá di Verona).

4. *On Robustness, Fault-Tolerance and Wireless Networks*, Seth Gilbert (École Polytechnique Fédérale de Lausanne).

5. *The Future of Distributed Computing: Renaissance or Reformation?*, Maurice Herlihy (Brown University).

In the social part of the event, there was a banquet on Wednesday (see Figures 3 and 4) at the Paradise Restaurant at the Centre Island. The landscape from the island is excellent; the city looks great at the night. The banquet was excellent, good food, good friends and beers. We also ate a cake in celebrating Nancy's birthday (see Figure 5). Congratulations Nancy!!

The organizers were as follows:

Figure 3: The banquet.

General Chair: Rida Bazzi.

Program Chair: Boaz Patt-Shamir.

Local Arrangements Chair: Eric Ruppert.

Publicity: Petr Kuznetsov and Thomas Moscriboda.

Treasurer: Srikanta Tirthapura.

Steering Committee Chair: Faith Ellen.

Steering Committee: Rida Bazzi, Dahlia Malkhi, Boaz Patt-Shamir, Michel Raynal, Srikanta Tirthapura and Roger Wattenhofer.

Program Committee: Azer Bestavros, Anat Bremler-Barr, Gregory Chockler, Antonio Fernández Anta, David Gamarnik, Leszek Gasieniec, Cyril Gavoille, Rachid Guerraoui, Emin Gün Sirer, Fabian Kuhn, Zvi Lotker, Nancy Lynch, Yishay Mansour, Mark Moir, Yoram Moses, Boaz Patt-Shamir, Sergio Rajsbaum, Adi Rosén, Dan Rubenstein, Eric Ruppert, Peter Sanders, Jennifer Welch and Lisa Zhang.

There are a lot of people that were not mentioned here but gave to us a great event. Thanks all of them. In addition, I want to thank Jonathan Lung for his photos.

Figure 4: A good time at the Centre Island.

Figure 5: Happy Birthday Nancy.

Review of DISC 2008

Robert Danek
University of Toronto
Canada
rdanek@cs.toronto.edu

Wojciech Golab
University of Toronto
Canada
wgolab@cs.toronto.edu

Wojciech Wawrzyniak
Adam Mickiewicz University
Poland
wwawrzy@amu.edu.pl

1 Introduction

The 22nd International Symposium on Distributed Computing, DISC, took place between September 22-24, 2008 in Arcachon, France. Arcachon is a beautiful coastal town in the south west of France, and is a popular holiday destination for Parisians. The conference venue was the Palais des Congrès, which is located right off the beach. The conference room overlooked the tranquil waters of Arcachon Bay, giving the presentations a picturesque backdrop and providing for a relaxed atmosphere during lunch. The beach's close proximity also allowed attendees the opportunity to go for a brief stroll after lunch before returning for the afternoon sessions.

There were a number of different topics covered at the conference this year. This included sessions on graph algorithms, shared memory synchronization, message passing algorithms, and failure detectors to name a few. There was also a brief announcements session where researchers presented shorter snapshots of their recent research.

DISC was co-located with several workshops: GRAAL, DYNAMO, and ANR/ARC FRACAS & MALISSE. These workshops focused on graphs and algorithms in communication networks (GRAAL), algorithmic aspects of dynamic networks (DYNAMO), and issues related to malice-tolerance, and self-organization in sensor and wireless networks (FRACAS & MALISSE).

There were two best paper awards given at DISC this year. The first one was for "best paper", and the second for "best student paper". The best paper award was given to Robert Danek and Wojciech Golab for "Closing the Complexity Gap Between FCFS Mutual Exclusion and Mutual Exclusion", which presented the first known FCFS Mutual Exclusion algorithm that uses only reads and writes and in which a process makes at most a logarithmic number of remote memory references (RMRs). (Measuring the number of RMRs a process makes is a common way of evaluating the time complexity of shared memory synchronization algorithms.) The algorithm Danek and Golab presented had the added bonus of being adaptive to the RMR time complexity measure, i.e., time complexity was a function of the number of processes executing the algorithm concurrently.

The best student paper award was given to W. Wawrzyniak for joint work with M. Hańćkowiak and A. Czygrinow titled "Fast distributed approximations in planar graphs". This paper gives some new approximation algorithms for typical graph theoretical problems, such as Maximal Independent Set (MIS), Maximal

Matching, and Minimal Dominating Set (MDS). Based on earlier papers by Naor and Linial, a new lower bound for approximating MIS and MDS in planar graphs is also presented. This result is connected with another paper in this conference, *Leveraging Linial's Locality Limit*, by C. Lenzen and R. Wattenhofer.

2 Excursion and Dinner Banquet

On Tuesday evening, the organizers of DISC scheduled an excursion to the Dune de Pyla. After a busy day of talks, participants were taken to the destination by two busses. On arrival, at the starting point of our trip, we saw the highest dune in Europe (107 m).

Figure 1: Dune de Pyla.

Before seeing one of the most beautiful sights in Europe, everybody had to overcome their own weakness climbing hundreds of stairs up the steep sandy slope. The effort was all the more big due to the beautiful warm weather, but consequently the emerging view was amazing. From the top of the dune we saw a fantastic view of the bay on the right, the ocean in front and a tract of forest on the left. In the age of the metropolis, the sight is so marvelous that we think it is worth coming here just to see it. As proof we give some photos in Figure 1. (More photos can be found at `http://dept-info.labri.fr/~gavoille/disc2008/`.)

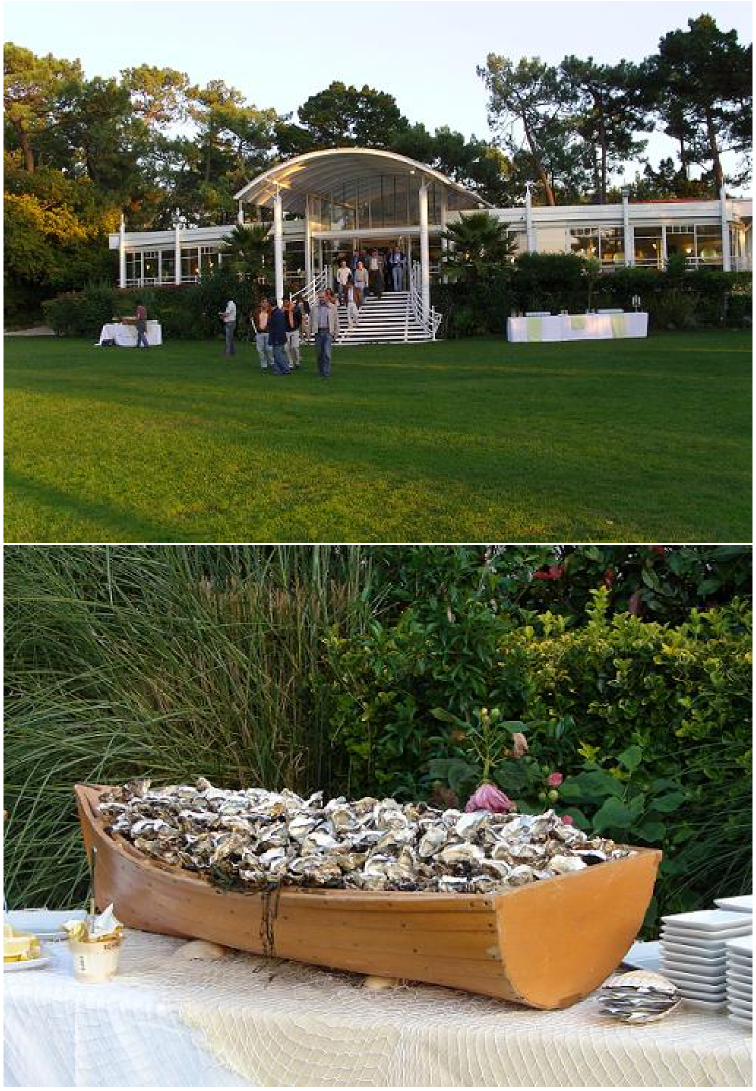


Figure 2: “Tir au Vol” and awesome oysters.

Shortly after the trip, the dinner banquet took place at “Tir au Vol”, located in Parc Pereire on the bay. One of the most amazing things at the conference banquet was the food. In addition to many delicious but quite rife dishes, a serving of awesome oysters was delivered in a small boat, which for some required a lot of courage to enjoy. Figure 2 gives snapshots of our fond memories from the banquet.

3 Conference program

This year's committees were chaired as follows: Gadi Taubenfeld – program committee and award committee; Rachid Guerraoui – steering committee; Cyril Gavoille – organization committee.

3.1 Keynote Talks

Each day of the conference opened with a cup of fresh coffee followed by an equally stimulating keynote talk.

Prof. David Harel of the Weizmann Institute in Israel entertained us on the first day with the talk titled "In Silico Biology, or On Comprehensive and Realistic Modeling". This cross-disciplinary work looks at realistic models of biological entities, such as tissues and organs, as reactive distributed systems that can be simulated interactively. The premise of these models is to accurately reproduce the behaviour of reactive components, such as cells, and, predict emergent properties of the system, such as the shape of an organ that develops from these cells under the control of complex biological interactions. Prof. Harel demonstrated exciting recent results, but also stressed that the research area is still in its infancy, challenging researchers to continue striving for the grand challenge – a complete model of a multi-cellular organism.

The second keynote talk, titled "Distributed Cooperation and Adversity: Complexity Trade-Offs", was given by Prof. Alexander Shvartsman from the University of Connecticut. This talk gave a survey of complexity results for the fundamental problem of executing a collection of tasks in an unreliable decentralized computing environment. Since the problem is generally solvable, it is a rich source of interesting complexity tradeoffs as it remains challenging to successfully combine efficiency with fault-tolerance under various forms of adversity. The talk surveyed fundamental results as well as bounds specific to message-passing and shared memory systems under various failure models and synchrony assumptions.

Dr. Phillip Gibbons from Intel delivered the final presentation, titled "Fun with Networks: Social, Sensor, and Shape Shifting". This talk discussed interesting algorithmic problems arising in novel network settings, and their solutions. In social networks, a Sybil attack involves a malicious user out-voting honest ones by assuming multiple fake identities, and can be dealt with using randomized routes over a topology that represents trust relationships among users. In-network energy-efficient aggregation of data in wireless sensor networks was the second problem considered. Synopsis Diffusion marries the efficiency of routing over spanning trees with the robustness of multi-path routing by using synopses to prevent double-counting of redundant data. The talk ended with an exciting look at using "catoms" – tiny robotic modules under software control – to dynamically form arbitrary physical shapes.

3.2 Full Paper Track

Contributions accepted to the full paper track were presented in seven sessions:

Session 1: Consensus - Chair: Philippas Tsigas
Several efficient consensus protocols were presented for message-passing systems under various failure and synchrony models. "How to Solve Consensus in the Smallest Window of Synchrony" by Alistarh et al. considers eventually synchronous systems with crash failures. "Constant-space Localized Byzantine Consensus" by Dolev and Hoch and "Bosco: One-Step Byzantine Asynchronous Consensus" by Song and van Renesse both deal with Byzantine failures, and consider special cases, namely a local adversary in a synchronous system, and contention-free runs in an asynchronous system, respectively. "Continuous Consensus with Failures and Recoveries" by Mizrahi and Moses considers synchronous systems with crash and

omission failures and looks at continuous consensus, where processes repeatedly reach common knowledge. Finally, “No Double Discount: Condition-based Simultaneity Yields Limited Gain” by Moses and Raynal looks at the complexity of consensus in synchronous systems with crashes under a combination of condition-based and simultaneity-based specifications.

Session 2: Network and graph algorithms - Chair: Dariusz Kowalski
The second session of DISC contained information about network and graph algorithms. We can see some new trends in the expanding area of techniques for analyzing wireless networks. Wireless network model graphs are among the main interesting issues here. Particularly fitting in this session was the second work involving the UDG model, titled “Computing Lightweight Spanners Locally”. Kanj et al. present a great local distributed algorithm that computes a bounded-degree plane lightweight spanner for a given UDG graph. This is a very important problem because it gives an efficient algorithm for constructing a topology for broadcast in wireless networks.

In the paper “A limit to the Power of Multiple Nucleation in Self-assembly”, Sterling shows that it is impossible to use multiple nucleation to accelerate tiling a surface in constant time for tile assembly models that are reversible, irreversible, error-permitting or error-free.

Dimitrov et al. in “Resilient Random Regular Graphs” show an algorithm that w.h.p. builds a resilient (connected) regular graph. They describe algorithms also to add and delete nodes from such graphs, while retaining resilience properties.

Session 3: Fault tolerance and distributed data - Chair: Shay Kutten
This session covered distributed data-sharing problems and mutual exclusion in message passing systems. The first paper, “A Self-stabilizing Algorithm with Tight Bounds for Mutual Exclusion on a Ring” by Chernoy et al., presents a three-state mutual exclusion algorithm for ring networks and a tight bound on its worst-case complexity. “Optimizing Threshold Protocols in Adversarial Structures” by Herlihy et al. looks at transforming threshold-based fault-tolerant replication protocols to ones that tolerate dependent failures. “Matrix Signatures: From MACs to Digital Signatures in Distributed Systems” by Aiyer et al. shows how to achieve the properties of digital signatures using MACs. “On the Robustness of (Semi)Fast Quorum-Based Implementations of Atomic Shared Memory” by Georgiou et al. explores the trade-off between fault-tolerance and latency for implementing atomic read/write objects in message-passing systems. “Optimistic Erasure Coded Distributed Storage” by Dutta et al. is about using erasure codes to implement atomic registers in the asynchronous crash-recovery message passing model.

Session 4: Shared memory synchronization - Chair: Michel Raynal
The shared memory synchronization session included a number of interesting papers. Transactional memory, an extremely popular topic in recent years, was represented with a paper by Geurraoui et al. This paper tackled the foundations of transactional memory by introducing the concept of permissiveness. A transactional memory implementation is said to be permissive if it never aborts a transaction unless necessary for correctness.

Another highlight of the session was “The Mailbox Problem”, in which Aguilera et al. introduced the attendees to a synchronization problem between a housewife and a postman. The postman signals the housewife when he has put mail in the mailbox, and the housewife must retrieve all letters from the mailbox. Moreover, the housewife can only retrieve mail when the box is non-empty. The motivation for the problem arises from the interaction between devices and the CPU in a computer, such as the interaction that occurs when an I/O device interrupts the CPU.

Session 5: Radio Networks and message passing algorithms - Chair: Seth Gilbert
As mentioned earlier, models for wireless networks are widely researched nowadays. This is evidenced by the topics of this session where three of five papers involved broadcasting in UDG or similar models.

In the first talk Elsaesser et al. presented the paper "On Radio Broadcasting in Random Geometric Graphs", in which they consider radio broadcasting algorithms in random geometric graphs (with different transmission radii and assuming full synchronization). They consider the graph model $G_{\leq c}$ and give a randomized algorithm for broadcasting messages in $O(\log \log n)$ steps (w.h.p). In "Efficient Broadcasting in Known Geometric Radio Networks with Non-uniform Ranges" by Gasieniec et al., lower and upper bounds for the broadcasting problem are presented but in the conflict-embodied model. "Broadcasting in UDG Radio Networks with Missing and Inaccurate Information", by Guido et al., was the third work from the UDG or similar model series of papers.

This session also included a presentation by Chalopin et al. of the paper "Local Terminations and Distributed Computability in Anonymous Networks", which examined the non-trivial problem of local terminations. In this paper, the authors give the first partition of a distributed task that can be computed with weak local termination.

The last paper of this session, by Tsuchiya and Schiper, was "Using Bounded Model Checking to Verify Consensus Algorithms", which proposed an approach for automatic verification of asynchronous consensus.

Session 6: Network algorithms and sensor networks - Chair: Sébastien Tixeuil
The papers in the sixth session reflected a mixture of network routing and topics in the increasingly popular area of ad-hoc mobile networks. In "Dynamic Routing and Location Services in Metrics of Low Doubling Dimension", Konjevod et al. present a dynamic compact routing algorithm that has applications to ad-hoc mobile networks and distributed hash tables. In "Online, Dynamic, and Distributed Embeddings of Approximate Ultrametrics", Dinitz gives the first metric embeddings with low distortion that require a small number of changes in their structure when then network changes. "On the Emulation of Finite-buffered OQ Switches using Combined Input-Output Queuing" by Elhaddad and Melhem examines the relationship among scheduling policies, and also characterizes the tradeoff between CIOQ speedup and input buffer occupancy. The session concluded with "Theoretical Bound and Practical Analysis of Minimum Connected Dominating Set in Ad Hoc and Sensor Networks" by Vahdatpour et al., which gives a tight bound on the ratio of the size of maximum independent set and size of the mininum connected dominating set.

Session 7: Mobile agents and failure detectors - Chair: Amos Korman
Last but not least was the session on mobile agents and failure detectors. This session included the presentation of "The Weakest Failure Detector for Message Passing Set-Agreement". The failure detector introduced was the Loneliness detector, which outputs *true* or *false* depending on a number of conditions. This was a very elegant result, and the presentation was very straightforward. A related result for the shared memory model, "Anti-Ω: The Weakest Failure Detector for Set Agreement", was presented by Piotr Zieliński at PODC 2008 a month earlier, although the correctness details of that result were much more involved.

Due to space constraints we don't have the opportunity to mention every paper presented at DISC this year. However, all the papers presented at DISC this year were of excellent quality and advanced the current state of knowledge in distributed computing. We look forward to next year's DISC in Spain.

Acknowledgement. Thanks to Cyril Gavoille, Phuong Hoai Ha, David Ilcinkas, Hyonho Lee, Yee Jiun Song, Gadi Taubenfeld, and Shmuel Zaks for sharing their excellent photos.

Review of SPAA'08

Zvika Guz
Dept. of Electrical Engineering, Technion
Haifa, 32000, Israel
zguz@tx.technion.ac.il

The *20th ACM Symposium on Parallelism in Algorithms and Architectures*, SPAA 2008, was held in Munich, Germany, between June 14 and June 16, 2008. This year the conference has celebrated its 20th anniversary with a series of invited talks, a special track on hardware and software techniques for multicore machines and a poster session. The program committee, chaired by Nir Shavit, has accepted 36 regular papers out of 128 submissions, 13 brief announcements and 5 posters, creating a varied program that brought together the theory and practice of parallel computing. SPAA papers have covered a large variety of topics - graph algorithms, broadcasting in networks, parallel and distributed scheduling, transactional memory, multicore systems, and others.

The conference took place at the Main Campus and the Garching Campus of the Technical University of Munich (TUM). Together with SPAA, Munich has also celebrated its anniversary (850th), providing the conference participants with two days of artists market, folk music, folk dance groups, and lots and lots of beer. The banquet dinner was held at a restaurant observing the beautiful Bavarian Opera House, and has featured a very fun presentation by Phillip Gibbons that reviewed the 20 years of SPAA via pictures, stories, trivia questions and funny anecdotes. The other contenders for the peak-of-SPAA'08 title were the two indoor slides in the Faculty of Mathematics at Garching Campus which have been tested by almost all participants of the conference on the last day. (See Figure 1).

While many of this year's papers have dealt with "classical" SPAA topics, the special track was unique in the sense that it featured papers dealing with the shift to multicore machines and the emergence of the multicore era. Fourteen papers were selected for the special track, entitled *"Hardware and Software Techniques for Multicore Machines"*, and chaired by Jim Larus. The track has spanned four sessions - a session on Multicore Systems, a session on Transactional Memory, a session on Multicore Algorithms, and a session on STM Design and Locks. Below I will review several representative papers from the special track.

The first session on Multicore Machines, which was also the very first session of the conference, has been preceded by a keynote talk by Prof. Kunle Olukotun from Stanford University who is one of the key people behind the Sun Niagra architecture. Olukotun set the stage for all subsequent presentations of the track as he touched both the hardware and software aspects of the shift to multicores. The talk opened with an overview of the Niagra-2 architecture and continued with a sketch of how future multicore machines may look like. Next, Olukotun presented his viewpoint on the key challenges of the multicore era, namely, developing software that will step up and exploit the plethora of parallelism that these machines will make available.

Figure 1: The indoor slides at the TUM Faculty of Mathematics.

I was happy to give the first regular talk of the conference, presenting Nahalal - a new cache architecture for Chip-Multiprocessors. The paper, which is a joint work with Idit Keidar, Avinoam Kolodny, and Uri C. Weiser has won the Best Paper Award of SPAA this year. The work is motivated by two key observations: (1) in many multi-threaded applications, a small set of shared cache lines accounts for a significant portion of the memory accesses; and (2) in future CMPs, cache access times will no longer be a constant but will rather depend on the concrete distance in the on-die layout between the data being accessed and the core accessing it. Our work leverages these two observations to devise Nahalal – a cache whose novel floorplan topology partitions cached data according to its usage (shared versus private data), and thus enables fast access to shared data for all processors while preserving the vicinity of private data to each processor. While our paper presents a specific optimization of CMP cache architectures, namely optimizing shared data access, it fits within the broader picture of re-thinking traditional designs in the context of CMP, and demonstrates why such systems cannot directly inherit the traditional principles and know-how's of neither uniprocessor nor multi-processor machines.

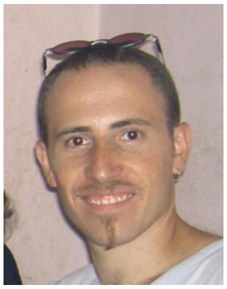
The next paper by Edya Ladan-Mozes and Charles Leiserson, has followed the same concept and presented a new cache consistency architecture designated for CMP – the Hierarchical Cache Consistency (HCC). Caches in HCC are shared hierarchically and protocol messages are embedded in the message-routing network that interconnects the caches. Every message in HCC makes monotonic progress without timeouts, retries, negative acknowledgments or retreating. Hence, HCC not only avoids few of the main drawbacks of other cache consistency models, but it also minimizes interconnect bandwidth which is a crucial factor in CMPs. The HCC model is fully distributed and highly scalable, thus particularly suitable for future many-core processors. The authors prove that the protocol is deadlock-free and that it provides sequential consistency, and also present an implementation of the MSI coherency protocol.

The two sessions on Transactional Memory (TM), and the large number of TM papers in this year's SPAA reflect the depth of research done in this field and the importance of TM as a tool to harness the parallelism of future multi-core machines. In this context, two papers have dealt with practical issues that have to be answered if Transactional Memory is to advance from a proof of concept into full-fledged commercial tool. First, Adam Welc, Bratin Saha and Ali-Reza Adl-Tabatabai have presented a work on irrevocable transactions. This work tackles the problem of supporting irrevocable action whose side effects cannot be rolled back (such as I/O operation or system call) within a transaction. While most TM assume that all operations executed within an atomic block are revocable (that is, their side-effects can be automatically rolled back), efficient support for irrevocable operations is a must in order to apply TM to any practical real-world program. The paper describes a TM system that supports irrevocable operations within transactions by allowing transactions to transition to an *irrevocable state*. When a transaction switches to *irrevocable state*, it is guaranteed that no subsequent action by any other transaction will ever cause this atomic block to be revoked. The paper describes a mechanism for efficient implementation of irrevocable transactions in which other transactions can still execute and commit concurrently, and also shows how irrevocability can be leveraged for contention management.

A second paper by Richard Yoo, Yang Ni, Adam Welc, Bratin Saha, Ali-Reza Adl-Tabatabai and Hsien-Hsin Lee has studied STM performance over large-scale real-world programs. This work shows that while most STM have been tested over small-scale workloads, there are major performance bottlenecks that are unique to the case of running large scale workloads on STM, and that realistic workloads must be used in order to take STM to the next level. Specifically, the authors identify four major performance bottlenecks: (1) false positive detection of conflicts, resulting from the coarse granularity used by the conflict detection mechanism; (2) over-instrumentation of read/write barriers generated by the compiler due to its lack of application-level knowledge; (3) privatization-safety cost, which refers to overheads resulting from the need to guarantee correctness of program idioms that are costly to enforce but in some cases are never used; and (4) poor amortization of transaction's startup/teardown overhead when transactions have a very short length.

Review of DSN'08

Gabriel Kliot
Department of Computer Science, Technion
Haifa, 32000, Israel
gabik@cs.technion.ac.il



The 38th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2008, was held in Anchorage, Alaska, USA between June 24 and June 27 2008.

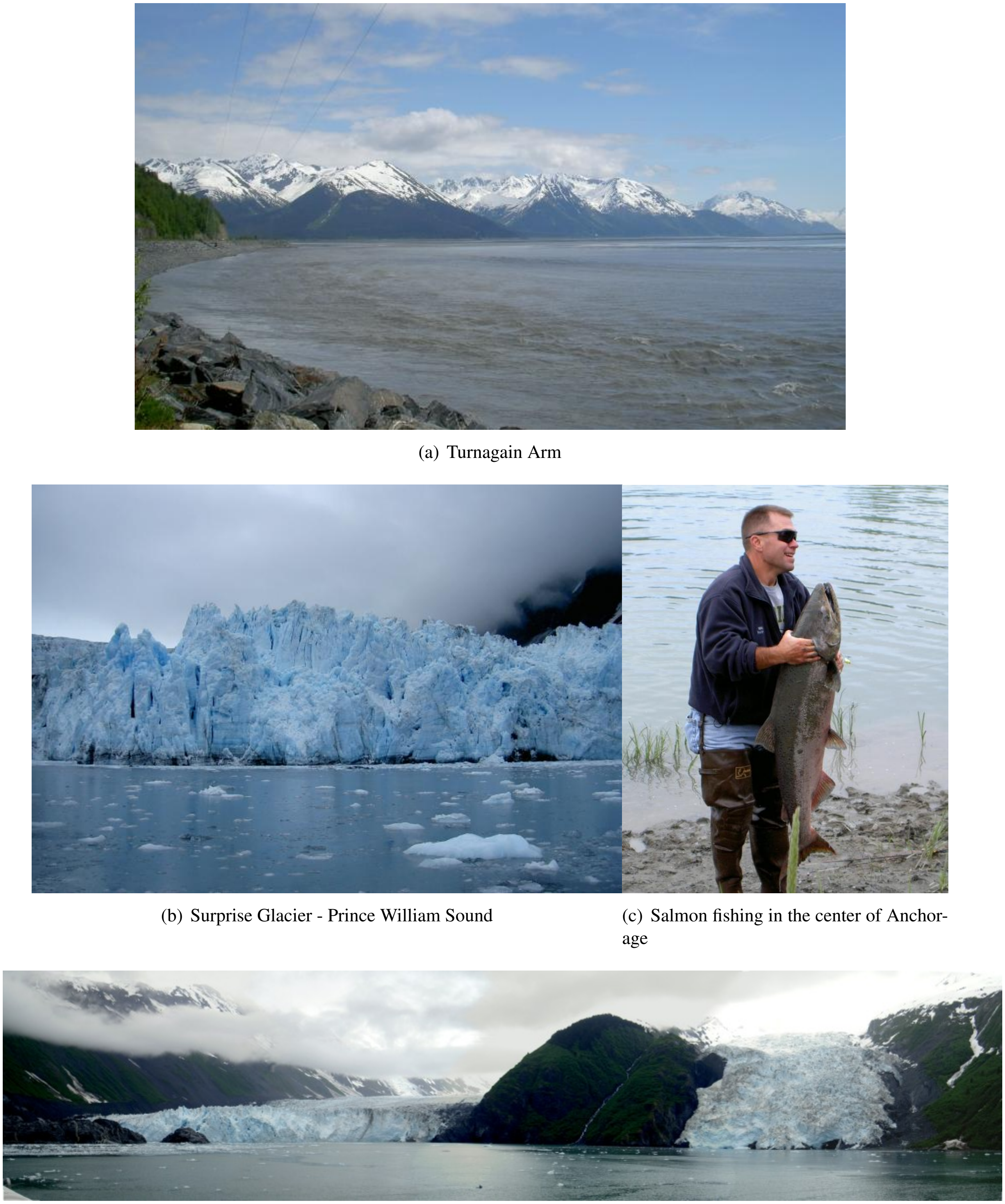
The conference included two tracks: the Dependable Computing and Communication Symposium (DCCS) and the Performance and Dependability Symposium (PDS), as well as four workshops, tutorials, a student forum, an industry forum on critical topics in dependability, fast abstract sessions, and 3 Birds of a Feather (BoF) sessions. The DCCS track, chaired by Neeraj Suri, featured 35 paper with acceptance rate of 23.5% and the PDS track, chaired by Kimberly Keeton, featured 23 paper with acceptance rate of 26.5%. Around one third of the paper dealt either directly or indirectly with different areas of distributed computing.

DSN'08 featured 2 keynotes. Alex Hills from CMU talked about the Alaskan aviation and its dependability (and sometimes non-dependability) across last decades, in his lecture titled "Dependability on the Last Frontier". After his lecture some of us felt we were lucky that the excursion will be by boat and not by airplane. Thomas Seder from General Motors talked about the challenges of Automotive Enhanced Vision Systems, concentrating on a set of issues that emerge from the nature of user-system interaction.

The annual William C. Carter Award was rewarded to a graduate student to recognize a significant contribution to the field of dependable computing through his graduate dissertation. The winner this year was Karthik Pattabiraman from of University of Illinois at Urbana-Champaign for the paper entitled "SymPLFIED: Symbolic Program Level Fault Injection and Error Detection Framework", by K. Pattabiraman, N. Nakka, Z. Kalbarczyk, and R. Iyer.

The organizing committee, chaired by Philip Koopman, has done an outstanding job on organizing the conference and the accompanying venue events.

One of the conference highlights was the excursion to the "26 Glacier Cruise" on Prince William Sound. We were taken by bus along the Turnagain Arm (see Figure 1(a)) with its famous, third largest tide in the world, crossed by a one-way tunnel to Whittier and boarded the Klondike Express luxurious catamaran. The catamaran took us along the numerous glaciers. The cruise highlight was the Surprise Glacier (Figures 1(b) and 1(d)), where the ship stopped for almost an hour to let everyone admire the magnificent view and take enough pictures. Back in Anchorage, those who wanted to get some original Alaskan experience could try their luck with fishing. Just 10 minutes walking outside the hotel enormous-sized salmons could be caught (see Figure 1(c) to apprehend the size)! And those lucky, who had some more time to explore Alaska, could experience the true Alaska by hiking some of its famous parks.



(a) Turnagain Arm

(b) Surprise Glacier - Prince William Sound

(c) Salmon fishing in the center of Anchorage

(d) Surprise Glacier - Prince William Sound

Figure 1: Prince William Sound cruise

DSN'08 featured a wide selection of papers on distributed and networking related subjects, including both theoretical and practical experience hands-on papers. I will now briefly review five representative distributed computing related papers.

Nicolas Schiper and Sam Toueg conducted an experimental evaluation of three known fault-tolerant leader election algorithms ([1, 2]). All algorithms use a stochastic failure detector algorithm of Chen *et al.* [4] and a link quality estimator in order to provide some degree of QoS control and to adapt to dynamic network conditions. Two of the proposed algorithms posses a desirable property of leader stability – current leader is not demoted and replaced if it is still operational. The algorithms were compared along four metrics: (1) leader recovery time – the time that elapsed from the time when the current leader has crashed until a new leader is elected again; (2) average mistake rate – the rate at which the service makes a mistake by demoting a functional leader (this could happen both due to failure detector's inaccuracy and election instability); (3) leader availability – the amount of time that the group has a commonly agreed and alive leader out of the total time. (4) CPU and bandwidth overhead. The comparison was performed in a LAN environment with simulated link crashes, message drops and workstation temporal crashes under a wide variety of settings. Two of the algorithms were found to behave well in extremely unfavorable conditions, i.e., in networks with very high processor failures and very poor communication links. This robustness is due to the combination of leader election algorithms that were proven to work under weak systems assumptions [1, 2], with an underlying failure detector algorithm that provides some QoS control [4].

The paper by Yair Amir *et al.* explored the impact of Byzantine attacks on the performance of Byzantine state-machine replication (SMR) protocols. As noted by the authors, those protocols usually target in providing correctness (safety and liveness) in a Byzantine environment, however their performance could still be greatly degraded by possible attacks. Specifically, they show how the famous BFT protocol [3] could be significantly slowed down by a malicious leader, by manipulating protocol specific timeouts and delays, without triggering suspicions and reelection by correct processes. Thus, while correct in the traditional sense (both safety and liveness are met), systems vulnerable to such performance degradation are of limited practical use in adversarial environments. The authors propose a new Byzantine fault-tolerant SMR protocol, resilient to performance degradation under attacks. The protocol meets a new performance-oriented correctness criterion, BOUNDED-DELAY, which makes a stronger guarantee than traditional liveness criteria. This is achieved by putting certain constrains on the leader and by aggressive monitoring leader's actions by the non-leader nodes and comparing its actions against a threshold level of acceptable performance.

Roxana Geambasu presented a paper resulting from her summer internship with Microsoft Research on "Fault-tolerant System Specification". In her project, Roxana targeted a goal of formally specifying several fault-tolerant file systems, thus enabling easier analysis, design, and comparison. The authors wrote formal TLA+ specifications for three systems (Chain Replication, Niobe (Microsoft) and GFS (the Google File System)) and used those specifications for three purposes: (1) to crystallize design differences and similarities; (2) to understand and mechanically verify consistency properties; and (3) to experiment with alternative designs. The authors reported specifications to be relatively easy to produce, useful for a deep understanding of system functioning, and valuable for system comparison. Such findings could potentially motivate other researchers to write formal specifications for their critical distributed systems as well.

Tudor Marian presented his joint work with Mahesh Balakrishnan, Ken Birman, and Robbert van Renesse on building soft state replication mechanism (called Tempest) in the service tier. Partially caching and managing soft state in the service tier (in addition to durably store the state in the databases or filesystem in the third tier) is a well-know technique to improve the performance of Service-Oriented Architectures, via increasing responsiveness and scalability. To further increase the availability of soft state stored in the service tier the authors consider replication. Tempest provides developers with TempestCollections: custom data structures that look similar to conventional Java Collections and are being transparently replicated across multiple machines, providing fail-over and load-balancing for soft state with zero extra effort by the developer. This is an opposite to other well known techniques to manage soft state, that require the developer to specially adapt the code to be replication-aware. Tempest uses a fast and unreliable IP multicast operation to spread/broadcast state changes to multiple service instances and then uses gossip-based reconciliation to maintain replica consistency in the face of faults and overloads. TempestCollections provide eventual consistency - "if the system is quiescent (i.e., if no more transactions are allowed to enter the system), the system eventually stabilizes in a globally consistent state." This is similar to the eventual consistency semantics provided by Amazon's Dynamo [5]. As previously reported by other works (e.g., [5]), such weaker consistency allows for an increased availability and responsiveness. Tempest was implemented as a new transport protocol in the Apache AxisSoap web services stack and compared against other in-memory databases, outperforming them in realistic settings.

Gabriel Kliot presented his joint work with Roy Friedman and Chen Avin on building Probabilistic Quorum Systems in wireless ad hoc networks. Quorums are a basic construct in solving many fundamental distributed computing problems, such as consensus and distributed storage. Probabilistic quorums were previously suggested by Malkhi *et al.* [6] to further improve the resilience, efficiency, and scalability of quorum systems. The basic idea in this work is that the quorums access strategy does not necessarily have to be symmetric. In particular, in a Bi-quorum system the nodes of only one of the two quorums should be picked uniformly at random out of all nodes, while the second quorum can be picked in an arbitrary (non adversarial) way, while still guaranteeing the same intersection probability. Gained with this insight the authors explore a number of ways to implement various asymmetric access strategy mixes, finding that the usage of random walks results in the best performance, both in static and mobile ad hoc networks. Along the lines the authors also present a number of useful theoretical results about random walks in random geometric graphs (graphs that are usually used to model ad hoc networks).

References

[1] M. K. Aguilera, C. Delporte-Gallet, H. Fauconnier, and S. Toueg. On implementing Omega with weak reliability and synchrony assumptions. In *Proc. of PODC*, pages 306–314, 2003.

[2] M. K. Aguilera, C. Delporte-Gallet, H. Fauconnier, and S. Toueg. On implementing Omega with weak reliability and synchrony assumptions. Technical Report HAL-00259018, CNRS - France, November 2007.

[3] M. Castro and B. Liskov. Practical Byzantine fault tolerance. In *Proc. of the 3rd Symposium on Operating Systems Design and Implementation (OSDI)*, pages 173–186, 1999.

[4] W. Chen, S. Toueg, and M. K. Aguilera. On the quality of service of failure detectors. *IEEE Transactions on Computers*, 51(5):561–580, May 2002.

[5] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: Amazon's highly available key-value store. In *Proc. of 21st ACM SIGOPS Symposium on Operating Systems Principles (SOSP)*, pages 205–220, 2007.

[6] D. Malkhi, M. Reiter, A. Wool, and R. Wright. Probabilistic Quorum Systems. *The Information and Computation Journal*, 170(2):184–206, November 2001.