Parallel and Distributed Algorithms

July 18–23, 1999

organized by

Bruce Maggs, Ernst W. Mayr, Friedhelm Meyer auf der Heide

The 6th Dagstuhl Seminar on *Parallel and Distributed Algorithms* was organized by Bruce M. Maggs (Carnegie Mellon University), Ernst W. Mayr (TU München), and Friedhelm Meyer auf der Heide (Paderborn University). It brought together 26 participants from Canada, Germany, Greece, Italy, Italy, Poland, UK, and the USA.

The presented talks covered a wide range of topics, like routing, load balancing, accessing global variables, graph partitioning, and many more. Furthermore, during an additional evening session, several open problems have been discussed.

The atmosphere provided by the Dagstuhl Center was a very pleasant one. The participants used the meeting rooms as well as the garden for lively and stimulating discussions. Also recreational aspects have not been neglected; due to the splendid weather the participants felt tempted to relax by means of hiking or biking. We would like to thank all people who contributed to the success of the seminar.

Participants

Petra Berenbrink, Paderborn, Germany, <pebe@uni-paderborn.de>
Thomas Erlebach, München, Germany, <erlebach@in.tum.de>
Torsten Fahle, Paderborn, Germany, <tef@uni-paderborn.de>
Faith Fich, Toronto, Canada, <fich@lri.fr>
Tom Friedetzky, München, Germany, <friedetz@in.tum.de>
Leslie Ann Goldberg, Warwick, U.K., <leslie@dcs.warwick.ac.uk>
Torben Hagerup, Frankfurt, Germany, <hagerup@informatik.uni-frankfurt.de>
Friedhelm Meyer auf der Heide, Paderborn, Germany, <fmadh@uni-paderborn.de>
Mirosław Kutyłowsky, Wrocław, Poland, <mirekk@tcs.uni.wroc.pl>
Stefano Leonardi, Roma, Italy, <leon@dis.uniroma1.it>
Bruce Maggs, Pittsburgh, U.S., <bmm@cs.cmu.edu>
Alberto Marchetti-Spaccamela, Roma, Italy, <alberto@dis.uniroma1.it>
Ernst W. Mayr, München, Germany, <mayr@in.tum.de>
Rafail Ostrovsky, New York, U.S., <rafail@research.telcordia.com>
Robert Preis, Paderborn, Germany, <robsy@uni-paderborn.de>
Yuval Rabani, Haifa, Israel, <rabani@csa.cs.technion.ac.il>
Rajmohan Rajaraman, Boston, U.S., <rraj@ccs.neu.edu>
Peter Rossmanith, München, Germany, <rossmani@in.tum.de>
Christian Scheideler, Paderborn, Germany, <chrsch@uni-paderborn.de>
Thomas Schickinger, München, Germany, <schickin@in.tum.de>
Jop Frederik Sibeyn, Saarbrücken, Germany, <jopsi@mpi-sb.mpg.de>
Paul G. Spirakis, Patras, Greece, <spirakis@cti.gr>
Angelika Steger, München, Germany, <steger@in.tum.de>
Rolf Wanka, Paderborn, Germany, <wanka@uni-paderborn.de>
Josef Weidendorfer, München, Germany, <weidendo@in.tum.de>
Matthias Westermann, Paderborn, Germany, <marsu@uni-paderborn.de>

Program

Monday

9:00 Leslie Ann Goldberg
Contention Resolution in Multiple-Access Channels

10:00 Robert Preis
Quality Matching and Local Improvement for Multilevel Graph-Partitioning

11:15 Jop F. Sibeyn
External Connected Components and Beyond

12:15 *Lunch*

16:00 Faith Fich
The Complexity of End-to-End Communication in Memoryless Networks

17:00 Torsten Fahle
Parallelization Strategies for the Vehicle Routing Problem with Time Windows

Tuesday

9:00 Rafail Ostrovsky
Universal $O(\text{congestion} + \text{dilation} + \log^{1+\epsilon} N)$ Local Control Packet Switching Algorithm

10:00 Thomas Erlebach
Time-Constrained Scheduling of Weighted Packets

11:15 Paul G. Spirakis
Efficient Redundant Assignments under Fault-Tolerance Constraints

12:15 *Lunch*

16:00 Stefano Leonardi
Parallel Machine Scheduling without Migration

17:00 Christian Scheideler
The PRESTO Project — Advances in designing a distributed real-time data server

Wednesday

9:00 Josef Weidendorfer
Transparent Load Balancing in an Explicit FE Solver

10:00 Thomas Schickinger
Load Balancing Using Bisectors — A Tight Average-Case Analysis

11:15 Matthias Westermann
Data Management in Networks

12:15 *Lunch*

13:30 *Excursion*

Thursday

9:00 Bruce Maggs
Protocols for Asymmetric Communication Channels

10:00 Rajmohan Rajaraman
Accessing Nearby Copies of Replicated Objects in a Distributed Environment

11:15 Yuval Rabani
Fairness in Routing and Scheduling

12:15 *Lunch*

16:00 Torben Hagerup
Improved Shortest Paths on the Word RAM

17:00 Mirosław Kutyłowski
Multi-party Finite Computations

19:30 *Open Problem Session*

Friday

9:00 Petra Berenbrink
Allocating Weighted Balls

10:00 Tom Friedetzky
Randomized and Adversarial Load Balancing

11:15 Rolf Wanka
Local Divergence of Markov Chains

12:15 *Lunch*

Contents

Leslie Ann Goldberg

Contention Resolution in Multiple-Access Channels

(Joint work with Hesham Al-Ammal, Mark Jerrum, Sampath Kannan, Phil MacKenzie, and Mike Paterson)

A multiple-access channel is a broadcast channel (such as an Ethernet) that allows multiple users to communicate with each other by sending messages onto the channel. If two or more users simultaneously send messages, then the messages interfere with each other (collide) and the messages are not transmitted successfully. The channel is not centrally controlled. Instead, the users use a contention-resolution protocol to resolve collisions. Thus, after a collision, each user involved in the collision waits a random amount of time (which is determined by the protocol) before resending. Informally, a contention-resolution protocol is "stable" if the backlog of messages stays finite. This talk surveys recent and ongoing work in the area, focusing on the following questions. (1) Are there any stable backoff protocols for positive arrival rates (in the infinitely-many users model)? (2) What is the best throughput that can be achieved by a backoff protocol (in the infinitely-many users model)? (3) For which arrival rates is Binary Exponential Backoff stable (in the finitely-many users model)?

Robert Preis

Quality Matching & Local Improvement for Multilevel Graph Partitioning

(Joint work with Ralf Diekmann and Burkhard Monien)

Multilevel strategies have been shown to be very powerful approaches for efficient graph-partitioning. Their efficiency is dominated by two parts; the coarsening and the local improvement strategies. Several methods have been developed to solve these problems, but their efficiency has only been proved on an experimental basis. We present new and efficient methods for both problems, satisfying certain quality measurements. For the coarsening part we develop a new approximation algorithm for maximum weighted matching in general edge-weighted graphs. It calculates a matching with an edge weight of at least $\frac{1}{2}$ of the edge weight of a maximum weighted matching. Its time complexity is $O(|E|)$, with $|E|$ being the number of edges in the graph. Furthermore, we show that the Helpful-Set

strategy, which can be used for iterative local improvement of existing partitions, guarantees an upper bound of $\frac{k-1}{2}|V| + 1$ on the bisection width of graphs with maximum degree of $2k$. These quality methods used for the two parts of the multilevel approach lead to an efficient graph-partitioning concept.

Jop F. Sibeyn

External Connected Components and Beyond

(Joint work with Ulrich Meyer)

The underlying idea of an earlier parallel list ranking algorithm is abstracted and turned into a general algorithmic method, comparable to, but different from, divide-and-conquer. As an application we derive novel algorithms for external connected components, list-ranking, tree rooting, minimum spanning trees and all-pairs shortest-paths. The method allows a more structured and more efficient reduction of the number of nodes than earlier algorithms such as those by Hirschberg for connected components and Boruvka for minimum spanning trees. Furthermore, in our algorithms the edges need to be stored only once, while in all other algorithms we know edges are stored at both endpoints. For many classes of graphs our connected components algorithm has excellent practical performance. Results are presented. The external implementation of our all-pairs shortest-paths algorithm has the *same* amount of paging as the standard matrix multiplication algorithm, this improves the best previous result by a factor of two. As a side result we present an improved external algorithm for matrix multiplication. Multiplying two $n \times n$ matrices on a computer with main memory size M and block size B requires only $(2 \cdot n^3/\sqrt{M})/B$ I/Os.

Faith Fich

The Complexity of End-to-End Communication in Memoryless Networks

(Joint work with Micah Adler)

End-to-end communication is the problem of sending a sequence of messages from a sender to a receiver, when the network through which they communicate is unreliable. The model considered is an asynchronous network with dynamic link failures, where intermediate processors store no information. A number of algorithms for solving this problem are surveyed and a new lower bound is presented. The lower bound is in terms of the network topology and implies, for example, that fixed degree meshes with n nodes and the complete network on n nodes require headers of length $\Omega(n)$ and $\Omega(n \log n)$, respectively. In many cases, including these networks and all series parallel networks, the lower bounds match known upper bounds.

Torsten Fahle

Parallelization Strategies for the Vehicle Routing Problem with Time Windows

(Joint work with Jürgen Schulze)

In this talk we describe two different approaches to parallelize a tabu search heuristics for the vehicle routing problem with time window constraints (VRPTW). The objective of the VRPTW is to deliver a set of customers with known demands on minimum-cost vehicle routes. The routes originate and terminate at a central depot and customers can only be delivered within a certain time window. Both algorithms start from a poor quality solution found by an insert-heuristic and improve it using a tabu search approach. The neighborhood of this local search is defined by simple customer shifts and allows to consider infeasible interim-solutions. The main difficulty with the parallelization of iterative local search strategies is their inherent sequential way of finding good solutions. Our first parallel algorithm is a master-slave approach where each iteration is performed by all processors in parallel. We describe the concept and how to cope with synchronization problems. Our second parallel algorithm performs several search threads concurrently. Periodically a fast set covering heuristic is used to

construct new solutions from older ones. We present extensive computational results for standard benchmark problems to show the behavior of both algorithms on Multiple-Instruction Multiple-Data computer architectures. Furthermore, we show that significant speed-ups can be achieved by our second algorithm while maintaining solution quality.

Rafail Ostrovsky

Universal $O(\text{congestion} + \text{dilation} + \log^{1+\epsilon} N)$ Local Control Packet Switching Algorithm

(Joint work with Yuval Rabani)

In 1988, Leighton, Maggs and Rao proved a much celebrated result: that for any network, given any collection of packets with a specified route for each packet, there exists an "optimal" schedule for all these packets. That is, there exists a schedule of the motion of the packets such that at each step, every edge is crossed by at most one packet, and all the packets are delivered to their destinations in $O(C + D)$ steps, where C is the "congestion" (i.e., the maximum number of paths that share the same edge), and D is the "dilation" (i.e., the length of the longest path). The proof was non-constructive and relied on Lovász Local Lemma. In a followup paper, Leighton and Maggs gave a centralized algorithm for finding the schedule. The original paper left open the question whether there exists a constructive distributed "on-line" algorithm with the same optimal performance. Last year, Rabani and Tardos presented a randomized local-control algorithm which with high probability delivers all packets in time $O\left(C + D \cdot \left((\log^* N)^{O(\log^* N)}\right) + (\log N)^6\right)$. In this paper, we show a nearly optimal local control algorithm for this long-standing open problem. We show a randomized algorithm which for any network topology delivers all the packets to their destinations in time $O(C + D + \log^{1+\epsilon} N)$ with high probability, where N is the size of the problem, and $\epsilon > 0$ is arbitrary. Our result has implications to ATM (Asynchronous Transfer Mode) packet switching algorithms and other applications.

Thomas Erlebach

Time-Constrained Scheduling of Weighted Packets

Using an LP-relaxation and a deterministic rounding procedure based on a coloring subroutine, we devise approximation algorithms for time-constrained scheduling of weighted packets in tree networks and in mesh networks with dimension-order routing. Every packet is given by a directed path, a release time, a deadline, and a positive weight. We consider the bufferless case, where a packet, once it has been sent out from its source node, must travel one link of its path in every time step until it reaches its destination. The goal is to find a feasible schedule that maximizes the total weight of packets that reach their destinations before their deadlines. For every fixed $\varepsilon > 0$, we obtain polynomial-time algorithms with approximation ratio $3 + \varepsilon$ in trees and $6 + \varepsilon$ in meshes, matching the best known bounds for the unweighted versions of the problems. Furthermore, using the same technique, we obtain a $(\frac{5}{3} + \varepsilon)$-approximation algorithm for the maximum weight edge-disjoint paths problem in bidirected trees and a $(4 + \varepsilon)$-approximation algorithm for the same problem in bidirected mesh networks with dimension-order routing.

Paul G. Spirakis

Efficient Redundant Assignments under Fault-Tolerance Constraints

(Joint work with D. Fotakis)

In this work we examine the problem of reliably sending messages through the use of unreliable channels (each having a probability of failure independently). We examine static assignements with the use of copies of messages to more than one channels. The objectives are (a) to find a static assignment that guarantees safe delivery of all the messages with probability above a certain threshold and (b) to minimise the maximum load created.
The problem is NP hard even in the case of channels of equal capacity. We first consider equal capacity channels. We show that the most robust schedules are of the form of partition schedules, ie where the channels are partitioned into reliable disjoint groups and the messages are split in an equibalanced way among

these reliable "effective channels". This result extends work by Lomonosov on construction of reliable multigraphs.
Through this result we achieve constant ratio approximations for the second objective (while satisfying the first) for the problem. We also provide efficient approximations for the case of channels of different capacities.
Some of the results of the talk appeared in APPROX-RANDOM 1999.

Stefano Leonardi

Parallel Machine Scheduling without Migration

(Joint work with B. Awerbuch, Y. Azar, O. Regev, L. Becchetti, and S. Muthukrishnan)

In this work we study parallel machine scheduling problems when preempted jobs cannot be migrated to a different machine. Jobs are released over time and must be processed for their processing time up to completion. Once a job is started on a machine, its processing can be interrupted and resumed later on the same machine. Awerbuch, Azar, Leonardi and Regev (Stoc '99) proposes a new on-line algorithm for parallel machine scheduling that allows preemption but it does not allow migration. The algorithm is proved to have optimal logarithmic competitive ratio for minimizing the average response time of a set of jobs, where the response time of a job is the time interval between the release and the completion of the job. In a later paper, Becchetti, Leonardi and Muthukrishnan (unpublished paper, '99) prove that the algorithm proposed in AALR also achieves $O(1)$ competitive ratio for optimizing the average stretch, where the stretch of a job is the ratio between the response time and the processing time of the job. Both average response time and average stretch objective functions measure the quality of service provided by the system. Average stretch is also a good measure of the load of the system. These results basically match the bound obtained for the same objective functions from the popular heuristic Shortest Remaining Processing Time that allows preemption *and* migration.

Christian Scheideler

The PRESTO Project — Advances in designing a distributed real-time data server

(Joint work with Petra Berenbrink and Andre Brinkmann)

PRESTO is the abbreviation for Paderborn REal-time STOrage network. The PRESTO project is done in a cooperation between working groups in the computer science department and electrical engineering department of the Paderborn university, funded by grants from the German government. In this project, we aim at constructing a scalable and highly fault-tolerant storage network that manages a set of parallel disks in a resource efficient way and that is able to support a fast real-time delivery of data. I present in this talk the principal issues and innovations involved in the design of the PRESTO storage network, concentrating on data placement, load balancing and routing strategies.

Josef Weidendorfer

Transparent Load Balancing in an Explicit FE Solver

The talk presents work done in the context of the EPaCTS project, a cooperation between TUM, BMW AG, SGI and ESI. The latter one sells a FE solver software specifically adopted to crashtest simulation simulation for the automobile industry. The project concentrates on aptimization of the scalability of the parallel version of this software. As the solver works in an explicit manner, meaning that normal FE calculation is only done among direct neighbours of the FE mesh, contact search and treatment has to be computed separately. This leads to changing and therefore unbalanced workload even without techniques like adaptive meshing. We propose a application transparent framework for load balancing. Here a underlying data abstraction independant from actual communication hardware is used. By framing accesses to shared data with Acquire/Release operations, the system is possible to automatically synchronize accesses in a specified manner. On top of this, an LB module (allowing arbitrary algorithms to be plugged in), a monitoring component actually triggering LB, and the original application is based.

Thomas Schickinger

Load Balancing Using Bisectors — A Tight Average-Case Analysis

(Joint work with Stefan Bischof and Angelika Steger)

In parallel computation we often need an algorithm for dividing one computationally expensive job into a fixed number, say N, of subjobs, which can be processed individually (with reasonable overhead due to additional communication). In practice it is often easier to repeatedly bisect jobs, i.e., split one job into exactly two subjobs, than to generate N subjobs at once. In order to balance the load among the N machines, we want to minimize the size of the largest subjob (according to some measure, like cpu-time or memory usage). In this talk we study a recently presented load balancing algorithm, called Heaviest First Algorithm (Algorithm HF), that is applicable to all classes of problems for which bisection can be computed efficiently. This algorithm implements a very natural strategy: During $N - 1$ iterations we always bisect the largest subproblem generated so far. We present results concerning the worst-case performance of Algorithm HF, which were obtained by Bischof, Ebner and Erlebach. They showed that the maximum load differs from the optimum only by a constant factor if a lower bound on the quality of the bisections is assumed. Furthermore, we intensively study the average-case (joint work with S. Bischof and A. Steger), assuming a natural and rather pessimistic distribution for the quality of the bisections. In this model the maximum load generated by Algorithm HF is proved to be only twice as large as the optimum with high probability. Additionally, our analysis suggests a simpler version of Algorithm HF which can easily be parallelized.

Matthias Westermann

Data Management in Networks

(Joint work with Christof Krick, Bruce Maggs, Friedhelm Meyer auf der Heide, Harald Räcke, and Berthold Vöcking)

This talk deals with data management for large parallel and distributed systems such as massively parallel processor systems (MPPs) and networks of workstations (NOWs) that consist of a set of nodes each having its own local memory module. These nodes are usually connected by a relatively sparse network such

that communication is often the major bottleneck. The only way to bypass this bottleneck is to reduce the communication overhead by exploiting locality. A dynamic data management service allows to access shared data objects from the individual nodes in the network. These objects are, e.g., global variables in a parallel program, pages or cache lines in a virtual shared memory system, shared files in a distributed file system, or pages in the World Wide Web. In this talk, we analyze theoretically and evaluate experimentally a data management strategy called the "access tree strategy".

The theoretical analysis of the access tree strategy considers data management in a competitive model. It is shown that the access tree strategy minimizes the congestion up to small factors by anticipating the locality included in an application. Thus, the access tree strategy prevents that some of the links become a communication bottleneck. In addition, the access tree strategy can deal with the problem that each node has only limited memory resources. Several classes of networks are considered. For example, it is shown that the access tree strategy achieves optimal competitive factor $O(\log n)$ for every application running on 2-dimensional meshes with n nodes. Furthermore, fat-trees, hypercubic networks, complete networks, and Internet-like clustered networks are investigated.

In the experimental evaluation of the access tree strategy we test several variations of this strategy on two different applications of parallel computing, which are matrix multiplication and Barnes-Hut N-body simulation. We compare the congestion and the execution time of the access tree strategy and their variations with a standard CC-NUMA strategy that uses a fixed home for each data object. Additionally, we do comparisons with hand-optimized message passing strategies producing minimal communication overhead. At first, we will see that the execution time of the applications heavily depends on the congestion produced by the different data management strategies. At second, we will see that the access tree strategy clearly outperforms the fixed home strategy and comes relatively close to the performance of the hand-optimized strategies. In particular, the larger the network is the more superior the access tree strategy is against the fixed home strategy.

Bruce Maggs

Protocols for Asymmetric Communication Channels

(Joint work with Micah Adler)

This talk will begin by briefly describing a wireless asymmetric network that the speaker has deployed at Carnegie Mellon to provide internet access to students living in neighborhoods surrounding the campus. It will then focus on the following problem inspired by the asymmetric network project. Suppose that a client and a server are connected by an asymmetric communication channel where the transmission bandwidth from the server to the client greatly exceeds the bandwidth from the client to the server. Suppose further that the client wishes to communicate an n-bit data item to the server, where the data is drawn from a distribution D that is known to the server but not to the client. Can the bandwidth from the server to the client be used to reduce the number of bits transmitted by the client? We show that the answer is yes. In particular, we present protocols in which the expected number of bits transmitted by the server and client are $O(n)$ and $O(H(D))$, respectively, where $H(D)$ is the entropy of D. In the simplest of these protocols, the expected number of rounds of communication is $O(H(D))$. We also give a protocol for which the expected number of rounds is only $O(1)$, but which requires more computational effort on the part of the server. These protocols are complemented by lower bounds and impossibility results. We show that all of the protocols are existentially optimal in terms of the number of bits sent by the server, i.e., there are distributions for which the expected total number of bits exchanged must be at least $n - 1$. In addition, we show that there is no protocol that is optimal for every distribution. We demonstrate this by proving that the problem of computing, for an arbitrary distribution D, a string of bits that the server should send to the client in order to minimize the expected total number of bits is undecidable.

Rajmohan Rajaraman

Accessing Nearby Copies of Replicated Objects in a Distributed Environment

(Joint work with Greg Plaxton and Andréa Richa)

Consider a set of shared objects in a distributed network, where several copies of each object may exist at any given time. To ensure both fast access to the objects as well as efficient utilization of network resources, it is desirable that each access request be satisfied by a copy "close" to the requesting node. Unfortunately, it is not clear how to efficiently achieve this goal in a dynamic, distributed environment in which large numbers of objects are continuously being created, replicated, and destroyed. In this talk, we describe a simple randomized algorithm for accessing shared objects that tends to satisfy each access request with a nearby copy. The algorithm is based on a novel mechanism to maintain and distribute information about object locations, and requires only a small amount of additional memory at each node. We analyze our access scheme for a class of cost functions that captures the hierarchical nature of wide-area networks. We show that under the particular cost model considered: (i) the expected cost of an individual access is asymptotically optimal, and (ii) if objects are sufficiently large, the memory used for objects dominates the additional memory used by our algorithm with high probability. We also address dynamic changes in both the network as well as the set of object copies.

Yuval Rabani

Fairness in Routing and Scheduling

(Joint work with Jon Kleinberg and Eva Tardos)

We consider the issue of network routing subject to explicit fairness conditions, namely max-min fairness. The optimization of fairness criteria interacts in a complex fashion with the optimization of network utilization and throughput. In this work we consider the problem of selecting paths for routing so as to provide a bandwidth allocation that is as fair as possible (in the max-min sense). We obtain the first approximation algorithms for this basic optimization problem, for single-source unsplittable routings in an arbitrary directed graph. Special cases of our model include several fundamental load balancing problems, endowing them with

a natural fairness criterion to which our approach can be applied. Our results form an interesting counterpart to earlier work of Megiddo, who considered max-min fairness for single-source fractional flow. The optimization problems in our setting become NP-complete, and require the development of new techniques for relating fractional relaxations of routing to the equilibrium constraints imposed by the fairness criterion.

Torben Hagerup

Improved Shortest Paths on the Word RAM

We prove two new results concerning the deterministic solution of single-source shortest-paths (SSSP) problems in almost linear space for networks with nonnegative edge lengths on a unit-cost random-access machine with a word length of w bits and a usual instruction set that does not include multiplication. (1) In directed networks with n vertices and m edges, the SSSP problem can be solved in $O(n + m \log w)$ time; (2) In undirected networks with n vertices and m edges, the SSSP problem can be solved in $O(n + m\alpha(m, n))$ time, where α is an "inverse Ackermann" function. Result (1) is new in that the space used is almost linear, and result (2) is new in that no multiplication is needed. The algorithms realizing (1) and (2) are more practical and, in the case of (2), significantly simpler than the theoretically fastest known algorithms.

Mirosław Kutyłowski

Multi-party Finite Computations

(Joint work with Tomasz Jurdziński and Krzysztof Loryś)

We consider computations executed by groups of finitely many finite devices working on common read-only input data and communicationg through messages. We consider the number of messages (of a finite length) as a complexity measure of a computation. This relates to multi-party communiaction complexity of problems

except we assume that the whole input data is available to each processing unit and that memory of each unit is severely restricted. We show a number of hierarchy results for this complexity measure: for each constant k there is a language, which may be recognized with $k+1$ messages and cannot be recognized with $k-1$ messages. We give an example of a language that requires $\Theta(\log\log n)$ messages and claim that $\Omega(\log\log(n))$ messages are necessary, if a language requires more than a constant number of messages. We present a language that requires $\Theta(n)$ messages. For a large family of functions f, $f(n) = \omega(\log\log n)$, $f(n) = o(n)$, we prove that there is a language which requires $\Theta(f(n))$ messages. Finally, we present functions that require $\omega(n)$ messages.

Petra Berenbrink

Allocating Weighted Balls

(Joint work with Friedhelm Meyer auf der Heide and Klaus Schröder)

The classic balls-into-bins game considers the experiment of placing m balls independently and uniformly at random(i.u.r.) in n bins. For $m = n$, it is well known that the maximum load, i.e., the number of balls in the fullest bin is $\Theta(\log n/\log\log n)$, with high probability. It is also known that a maximum load of $\emptyset\left(\frac{m}{n}\right)$ can be obtained for all $m \geq n$ if each ball is allocated in one (suitably chosen) of two (i.u.r.) bins. Stemann presents a distributed algorithm to find the "suitable" bin for each ball. The algorithm uses r communication rounds to achieve maximum load of $\max\left\{\sqrt[r]{\log n}, \emptyset\left(\frac{m}{n}\right)\right\}$, with high probability.
We generalize Stemann's upper bound to weighted balls: Let W^A (W^M) denote the average (maximum) weight of the balls. Furthermore, let $\Delta = W^A/W^M$. Then the optimal maximum load is $\Omega(m/n \cdot W^A + W^M)$. We present a protocol that achieves a maximum load of $\gamma \cdot \left(\frac{m}{n} \cdot W^A + W^M\right)$ using $\emptyset\left(\frac{\log\log n}{\log(\gamma\cdot((m/n)\cdot\Delta+1))}\right)$ communication rounds. For uniform weights this matches the results of Stemann. In particular, we achieve a load of $\emptyset\left(\frac{m}{n} \cdot W^A + W^M\right)$ using $\log\log n$ communication rounds, which is optimal up to a constant factor.
Our approach can also be used to achieve similar results for weighted balls in the case of static or dynamic versions of the sequential balls-into-bins games where every ball is placed into the less loaded bin of several randomly chosen ones. Furthermore, the approach can be used to show upper bounds for a parallel and dynamic balls-into-bins game where for each weighted ball a copy is inserted into two randomly chosen bins. The bins store the copies in a FIFO queue. Every bin is allowed to delete balls with a fixed total weight per round out of its queue

and also deletes the other copies of these balls.
All the balls-into-bins games model load balancing problems: The balls are jobs, the bins are resources, the task is to allocate the jobs to the resources in such a way that the maximum load is minimized. Our extension to weighted balls allows us to extend previous bounds to models where resource requirements may vary. For example, if the jobs are computing tasks, their running time may vary. Applications of such load balancing problems occur, e.g. for client-server networks and for multimedia-servers using disk arrays.

Tom Friedetzky

Randomized and Adversarial Load Balancing

(Joint work with Petra Berenbrink and Angelika Steger)

In this talk we consider dynamic load balancing algorithms for randomized and adversarial load generation models. Consider a system of n processors. In our randomized generation models every processor may generate a task with a certain probability at each time step, leading to an expected system load of $\mathcal{O}(n)$. We present a load balancing algorithm that assures that with high probability no processor has a load exceeding $\mathcal{O}(\log \log n)$ at an arbitrary point of time. This improves upon the previously known bound of $\mathcal{O}\left((\log \log n)^2\right)$. In the case of the adversarial load generation model every processor can change its load by some constant at each time step. Thus, the system load may become arbitrarily large. We present a balancing algorithm and show that if at some point of time τ no processor has a load exceeding some constant times the average, with high probability this holds for the next polynomial number of steps.

Rolf Wanka

Local Divergence of Markov Chains

(Joint work with Yuval Rabani and Alistair Sinclair)

We develop a general technique for the quantitative analysis of iterative distributed load balancing schemes. We illustrate the technique by studying two

simple, intuitively appealing models that are prevalent in the literature: the diffusive paradigm, and periodic balancing circuits (or the dimension exchange paradigm). It is well known that such load balancing schemes can be roughly modeled by Markov chains, but also that this approximation can be quite inaccurate. Our main contribution is an effective way of characterizing the deviation between the actual loads and the distribution generated by a related Markov chain, in terms of a natural quantity which we call the *local divergence*. We apply this technique to obtain bounds on the number of rounds required to achieve coarse balancing in general networks, cycles and meshes in these models. For balancing circuits, we also present bounds for the stronger requirement of perfect balancing, or counting.