Research Statement

Murat Demirbas

My research is in the broad area of distributed systems. In distributed systems, nodes execute concurrently with limited information about what the other nodes are doing. A fundamental problem in distributed systems is to coordinate the behavior of these independent nodes effectively. For safety reasons concurrency needs to be tamed to prevent unintentional nondeterministic executions, on the other hand, for performance reasons, concurrency needs to be boosted to achieve timeliness. My research focus is on designing efficient and resilient algorithms and services for coordination in distributed systems.

Distributed coordination faces many challenges. A major challenge is the scalability of coordination. While communication and synchronization is needed for coordination, they should be minimized to achieve scalability. Another major challenge is the unreliability of nodes and communication links between nodes. Several impossibility results haunt the distributed coordination problem. Moreover, large-scale distributed systems introduce complex unanticipated failure modes and cascading failures that complicate distributed coordination.

Motivated by these challenges, my research is focused on designing efficient and resilient algorithms and services for coordination in distributed systems. I have been investigating distributed coordination in the context of different application domains, including wireless sensor/actor networks, crowdsourced smartphone-based sensing and collaboration, and most recently in the context of large scale coordination in cloud computing. I have been successful in attracting steady funding for my research projects. I am the Principal Investigator on all of the research projects listed below, except for the National Institute of Health (NIH) and National Science Foundation(NSF)–CI projects, for which I am a Co-Principal Investigator.

1. Scalable coordination for wide-area distributed systems *(NSF CSR—Computer Systems Research 2015-2018)*
2. Synchrony-aware primitives for building highly auditable, highly scalable, highly available distributed systems *(NSF XPS—Exploiting Parallelism and Scalability 2015-2019)*
3. Smartsourced sensing and collaboration *(Google Research Award 2010, NIH/NIEHS award 2010-2012, NSF CI—Cyber Infrastructure 2012-2015)*
4. Efficient and resilient querying and tracking services for wireless sensor networks *(Office of Naval Research 2009-2012)*
5. Tool support for producing high-assurance and reliable software for wireless sensor/actor networks *(NSF CSR—Computer Systems Research 2009-2012)*
6. An in-network collaboration and coordination framework for wireless sensor/actor networks *(NSF CAREER Award 2008-2013)*

Next, I briefly summarize my research approach and contributions in these projects.

1 Scalable coordination for wide-area distributed systems (funded by NSF CSR 2015-2018)

Problem: For loosely dependent tasks that require little or no communication, simple abstractions such as barrier synchronization suffice for coordination, however, for large-scale cloud-computing and web-services applications that require tighter synchronization (such as online transaction processing systems, large-scale web-services, distributed file systems, social networking, and graph processing applications), a fine-grained complex coordination service is needed. Traditional distributed coordination techniques fail to scale for wide-area networks to support these emerging applications. Centralized coordination fails to scale with respect to the increased distances in the wide-area, whereas decentralized coordination fails to scale with respect to the number of nodes involved.

Approach: We propose to achieve scalable coordination over wide-area using a novel hybrid design, called Maestro. The Maestro framework will address the following research questions: (1) What are the limits of fully-centralized and fully-decentralized solutions to coordination, and what are the scalability benefits of a hybrid hierarchical approach? (2) How can locality-awareness be utilized to achieve high-performance across wide-area? (3) How can partition-awareness be utilized to achieve consistency across wide-area?

Expected contributions: The Maestro framework introduces a hierarchical lock broker architecture with a novel lock-leasing mechanism and smart/adaptive lock migration. This architecture allows flexibility of control and provides the best of both centralized and decentralized approaches. As the authority of their respective domains, the brokers learn and adapt to the access patterns at runtime to improve lock-locality and hence scalability, while they also have autonomy to allow independent accesses to be initiated and executed in a decentralized manner. Maestro will provide optimizations such as proactive leasing of locks to servers even before they are requested, lock migration (changing primary site assignment of locks), and shared/fractional lock-leasing to selectively allow decentralized coordination and relaxed consistency when appropriate. The paper we outlined the design of the Maestro lock broker won a best paper award in IEEE Big Data Conference in October 2015 [1].

Our Maestro framework will fill an important gap in the wide-area scalable coordination of tightly-coupled consistency-critical distributed applications. We will evaluate Maestro on two popular distributed application domains: wide-area ZooKeeper for distributed coordination, and wide-area distributed metadata management. We also started collaborations with the Microsoft DocumentDB team as part of this project.

2 Synchrony-aware primitives for building highly auditable, highly scalable, highly available distributed systems (funded by NSF XPS 2015-2019)

Problem: Auditability is a key concept for enabling highly scalable and highly available distributed systems (such as Facebook, Google, and Twitter); auditability enables identifying latent

concurrency bugs, faulty states and services, and performance bottlenecks. In turn, for the auditability of a system, time is a key concept. Unfortunately, there is a gap between the theory and the practice of distributed systems in terms of the use of time. The theory of distributed systems shunned the notion of time and considered asynchronous systems, whose event ordering is captured by logical clocks. The practical distributed systems employed NTP synchronized clocks to capture time but did so in ad hoc undisciplined ways.

Approach: Our work focuses on providing auditability by combining two key concepts: time and causality. In particular, we introduce hybrid logical clocks (HLC) [2,3] which offer the functionality of logical clocks while keeping them *close* to physical clocks. HLC combines the theoretical underpinnings of causality and the practicality of physical clocks by identifying how logical clocks can be improved and tuned based on the availability of NTP synchronization. The principle guiding HLC design is "uncertainty resilience". HLC is designed to be always wait-free/nonblocking and correct (albeit with reduced efficiency) even when time synchronization has degraded or is not available.

Expected contributions: HLC can provide efficient global consistent-state snapshots without needing to wait out clock synchronization uncertainties and without requiring prior coordination. Leveraging on the auditability primitives provided by HLC [4], we will build support for scalable and available systems. To enable highly available systems, we will investigate the design of a monitor component that detects and corrects distributed system state corruptions. The principle guiding the design of the monitor component will be "centralized oversight and override". To enable highly scalable systems, we will investigate design of synchrony-aware coordination primitives, such as barrier synchronization, mutual exclusion, leader election, and causally and totally ordered communication support. The principle guiding the design of the synchrony-aware coordination primitives will be "silent consent". These primitives will improve performance and efficiency over their asynchronous system counterparts by trading timing information gathered from HLC for avoiding explicit communication needed for coordination.

This project has applications to cloud computing, distributed NewSQL databases, and globally distributed web services. Our hybrid clocks have recently been adopted by CockroachDB [5], an opensource clone of Google Spanner multiversion distributed database.

3 Smartsourced sensing and collaboration **(funded by Google Research 2010, NIH/NIEHS 2010-2012, NSF CRI 2012-2015)**

Problem: Smartphones show a lot of promise for solving the large-scale sensing problem and fulfilling the ubiquitous computing vision of pervasive collaboration, however they fall short of their potentials. Consider DARPA's 2009 network grand challenge on accurately finding 10 red weather balloons deployed at arbitrary locations of the US. The winning team managed to solve the challenge in 9 hours, but the team had to prepare, campaign, and publicize aggressively for a month, and employed a multilevel incentive structure that distributed the $40K prize money among participants. To achieve full potential of smartphone based sensing and collaboration, a platform is needed to enable development of smartsourcing apps that solve similar collaboration and coordination problems without requiring month-long campaigns and $40K awards.

Approach: We believe that the reason current apps fail to solve problems like 10 red balloons is the lack of a platform to utilize smartphones for collaboration and coordination. In contrast, providing a platform for publish/subscribe and tasking of these devices would enable any smartphone to utilize the data published by other smartphones in a region and to task other smartphones in the region to acquire new data if needed. In order to utilize smartphones for solving collaboration and coordination problems, we propose an open publish-subscribe middleware at the cloud backend, *Eywa*, which enables smartphones to benefit from data collected by other smartphones and Internet of Things (IoT).

Contributions: In our early work we employed Twitter as a primitive publish-subscribe middleware and developed a crowdsourced weather radar [6] and citywide sensing applications over Twitter [7–9]. To improve the success rate of location-based services, we built a system [10] that categorizes Twitter users based on their familiarity to location types in Foursquare and uses this information to forward more relevant queries to the users.

Our work on mobile user profiling [11–18] aimed to improve the efficiency of our smartsourced sensing system. As a limited scope demonstration of the Eywa vision, we employed the localization capabilities on the smartphones and leveraged on the computational power provided by cloud servers in order to estimate and forecast the wait times at coffee shops; our system, LineKing, became the first crowdsourced line wait-time estimation service [19].

Since aggregating responses from a crowd is a challenge for the Eywa platform, we investigated multiple-choice question answering (MCQA) [20–22]. To this end, we designed a gamified experiment. We developed an Android app to let the crowd answer questions with their smartphones as they watch the Who Wants To Be A Millionaire (WWTBAM) quiz show on a Turkish TV channel. When the show is on air in Turkey, our smartphone app signals the participants to pickup their phones. When a question is read by the show host, we would type the question and the multiple choices, which would be transmitted via Google Cloud Messaging (GCM) to the app users [23]. App users play the game, and enjoy competing with other app users, and we get a chance to collect multi-dimensional data about MCQA dynamics in crowdsourcing. Our WWTBAM app has been downloaded and installed more than 300,000 times and has enabled us to collect more than 3 GB of MCQA data (2000 live quiz-show questions and more than 200,000 answers) over a period of 9 months. We developed a novel PageRank-like algorithm to perform history-based weighing of capable participants to improve the accuracy of aggregation [20, 22]. Our aggregation algorithm raised the accuracy of answers to 90% even for the hard questions, whose accuracy dropped to 40% using a naive majority voting scheme. We also discovered a simple alternative approach, which is as effective as the sophisticated history-based solution. This approach did not use the history of participants, but rather used the interests of participants (obtained from the category of the apps installed in the participants' phones) to weigh their answers. Our paper [21] introducing this method won the outstanding paper award in the Collaboration Technologies and Systems (CTS 2014) conference.

4 Efficient and resilient querying and tracking services for WSNs

(funded by Office of Naval Research 2009-2012)

Problem: The goal of an in-network querying service in WSNs is to answer spatial queries, such as "What is the location of the nearest enemy tank to my coordinates?". To answer such queries, querying and tracking services require continuous maintenance of distributed data structures (trees, paths, and clusters) over a large number of nodes. In order to achieve scalability, only the relevant WSN nodes should be involved in the execution of the query. That is, these services should implement local operations over these global structures. Locality is also needed in handling of the faults. In the absence of a local healing mechanism, faults/inconsistencies in one part of the system may contaminate the entire system and hence may result in a high-cost, system-wide correction.

Approach: Our approach in achieving locality and scalability of WSN services is to exploit geometric ideas and techniques while devising distributed algorithms. In contrast to Internet, where the topology is logical and arbitrary graph models are used, WSNs are deployed in physical spaces and the use of geometric networks are warranted for modeling WSNs. We find that when the problem domain is constrained to geometric networks it is possible to devise simpler and more efficient algorithms than those designed for arbitrary graph topologies. Especially for reasoning about locality of solutions in WSN (where communication cost is the biggest constraint on design) geometric methods are a good fit. Furthermore, by exploiting the geometry of the network, we propose efficient fault-containment techniques to enable locality in handling of faults.

Contributions: For in-network querying, we proposed a *distributed quad-tree (DQT) structure* [24] that achieved a querying cost of $2\sqrt{2} * d$, and showed graceful resilience to the face of failures. In *Glance* [25], we introduced a geometry-based elegant algorithm which ensured that a query invoked within d distance of an event intercepts the event's advertisement within $d * s$ distance, where s is a "stretch-factor" tunable by the user.

For tracking in WSNs, in *Stalk* [26], we employed hierarchical partitioning to maintain a tracking structure over a small number of nodes and with accuracy proportional to the distance from the evader. In Stalk: 1) Operations to find the mobile object distance d away take $O(d)$ time and communication to complete, 2) Updates to the tracking structure after the object has moved a total of m distance take $O(m * logm)$ amortized time and communication to complete, 3) The tracked object may relocate without waiting for completion of the updates resulting from prior moves, and 4) The mobile object can move while a find is in progress. For achieving local healing of the tracking structure in Stalk, we used *containment waves*, which ensured that contamination due to faults is restricted to an area proportional to the perturbation size, and the tracking path stabilizes within work proportional to the perturbation size instead of the network size.

Continuing in this vein, in *Trail* [27], we presented a tracking protocol that achieves the same linear costs for find and update in Stalk *without* requiring a hierarchical partitioning of the network. We also applied our geometric ideas and techniques for in-network querying/tracking algorithms in order to devise energy-efficient and low-latency data collection mechanisms for WSNs using network-controlled mobile basestations [28–30].

5 An in-network collaboration and coordination framework for wireless sensor/actor networks (funded by NSF-CAREER 2008-2013)

Problem: As WSNs get increasingly more integrated with actuation capabilities, consistency and timeliness guarantees become significant issues. However, in the presence of unreliable communication channels—as is the case in WSNs— even the most basic consensus problem of getting nodes agree on a binary decision is unsolvable (due to the Coordinated Attack impossibility result). Therefore, effectively managing concurrent execution ranks as one of the biggest challenges for future wireless sensor-actor networks (WSANs).

Approach: Our key insight in this project is to observe that singlehop wireless broadcast has many useful features for facilitating collaboration and coordination. Firstly, broadcasting is atomic (i.e., for all the recipients of a broadcast, the reception occurs simultaneously), which is useful for synchronizing the nodes in singlehop for building a structured operation. Secondly, broadcast messages share the same medium enabling collision detection and snooping, which are useful for implementing collaboration and coordination in a decentralized manner. Leveraging these properties, we develop a framework that provides simple programming abstractions to cope with the consistent coordination challenges of WSANs while retaining efficiency of execution. Our framework consists of two components: (1) a singlehop communication primitive for fast robust feedback collection, and (2) a transactional abstraction for robust computing in WSANs.

Contributions: In [31] we have shown, for the first time, that it is possible to solve consensus efficiently in WSNs using a novel receiver-side collision detection (RCD) technique. The idea here was to provide a dedicated round for communicating negative feedback, and hence conveying information even when the negative feedback messages collide. While the transmitter cannot detect collisions in WSNs, there is no barrier against the RCD. In our algorithm, a collision detected in the veto round indicates the existence of at least one veto message and that the consensus should be deferred for a later round. We have also given a classification of RCD with respect to its completeness (ability to detect collisions) and accuracy (ability to avoid false positives) and identified the lower-bounds for solving consensus for each class. We developed reliable implementations and quantitative evaluations of RCD on TinyOS and mote platforms in [32]. Using RCD, we built a primitive "pollcast" for quick and robust singlehop feedback collection from a singlehop neighborhood. We have analyzed the theoretical lowerbounds and upperbounds associated with querying with pollcast in [33]. In [34], we developed and showcased applications of RCD for singlehop collaboration protocols in WSNs, including a quick and robust threshold querying primitive *tcast*. We also applied these principles to the reliable broadcast problem in [35,36], which is another relevant and significant problem for single-hop communication.

As complementary singlehop communication primitive for fast robust feedback collection, we designed and developed the TRANSACT framework [37,38], which provides an efficient and lightweight implementation of transaction primitive in a distributed manner. In contrast to database systems, in distributed WSANs there is no central database repository or an arbiter; the control and sensor variables, on which the transactions operate, are maintained distributedly over several nodes. As such, it is infeasible to impose control over scheduling of transactions at different nodes, and also challenging to evaluate whether distributed transactions are conflicting. However, by exploiting the properties of broadcast communication inherent in WSANs, TRANSACT overcomes this chal-

lenge and provides a lightweight implementation of transaction processing. A major contribution of TRANSACT is to simplify the reasoning and verification of a distributed WSANs program. Building blocks for process control and coordination programs (such as, leader election, mutual exclusion, cluster construction, neighborhood discovery, recovery actions, and consensus) are easy to denote using TRANSACT.

References

[1] S. Tasci and M. Demirbas. Panopticon: A lock broker architecture for scalable transactions in the datacenter. *IEEE International Conference on Big Data*, 2015.

[2] S. Kulkarni, M. Demirbas, D. Madappa, B. Avva, and M. Leone. Logical physical clocks. In *Principles of Distributed Systems*, pages 17–32. Springer, 2014.

[3] M. Demirbas and S. Kulkarni. Beyond truetime: Using augmentedtime for improving google spanner. *LADIS '13: 7th Workshop on Large-Scale Distributed Systems and Middleware*, 2013.

[4] M. Demirbas and S. Kulkarni. Highly auditable distributed systems. *USENIX HotCloud*, 2015.

[5] Cockroachdb: A scalable, transactional, geo-replicated data store. http://cockroachdb.org/.

[6] M. Demirbas, M. A. Bayir, C. G. Akcora, Y. Yilmaz, and H. Ferhatosmanoglu. Crowd-sourced sensing and collaboration using twitter. *11th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2010.

[7] C. G. Akcora, M. A. Bayir, M. Demirbas, and H. Ferhatosmanoglu. Identifying breakpoints in public opinion. *Workshop on Social Media Analytics*, 2010.

[8] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in twitter to improve information filtering. *33rd ACM SIGIR Conference*, 2010.

[9] `http://ubicomp.cse.buffalo.edu/citypulse`.

[10] M. Demirbas M. F. Bulut, Y. S. Yilmaz. Crowdsourcing location-based queries. *IEEE Workshop on Pervasive Collaboration and Social Networking (PerCol)*, 2011.

[11] M. A. Bayir, M. Demirbas, and N. Eagle. Discovering spatiotemporal mobility profiles of cellphone users. *Pervasive and Mobile Computing Journal, Special Issue on Human Behaviour in Ubiquitous Environments*, 6(4):435–454, 2010.

[12] M. A. Bayir, M. Demirbas, and A. Cosar. A web based personalized mobility service for smartphone applications. *The Computer Journal*, 54(5):800–814, 2011.

[13] M. A. Bayir and M. Demirbas. Pro: A profile based routing algorithm for pocket switched networks. *IEEE Global Communications Conference (Globecom)*, 2010.

[14] M. Demirbas, C. Rudra, A. Rudra, and M. A. Bayir. imap: Indirect measurement of air pollution with cellphones. *PerCOM*, pages 537–542, 2009.

[15] M Glasgow, L Mu, P Nayak, C Crabtree-Ide, M Demirbas, E Yoo, A Szpiro, A Rudra, J Merriman, J Wactawski-Wende, et al. Smartphone technology for improving air pollution exposure estimates. In *American Journal of Epidemiology*, volume 175, pages S9–S9, 2012.

[16] M. L. Glasgow, C. B Rudra, E-H Yoo, M. Demirbas, J. Merriman, P. Nayak, C. Crabtree-Ide, A. A. Szpiro, A. Rudra, J. Wactawski-Wende, et al. Using smartphones to collect time–activity data for long-term personal-level air pollution exposure assessment. *Journal of Exposure Science and Environmental Epidemiology*, 2014.

[17] A. Nandugudi, A. Maiti, F. Bulut, S. Batra, Y Ki, G. Challen, M. Demirbas, S. Ko, T. Kosar, and C. Qiao. Participant behavior in phonelab. *Third International Conference on the Analysis of Mobile Phone Datasets (NetMob)*, 2013.

[18] M. F. Bulut and M. Demirbas. A holistic approach for energy efficient proximity alert on android. In *Global Communications Conference (GLOBECOM), 2013 IEEE*, pages 2816–2821. IEEE, 2013.

[19] M.F. Bulut, Y.S. Yilmaz, M. Demirbas, H. Ferhatosmanoglu, and N. Ferhatosmanoglu. Lineking: Crowdsourced line wait-time estimation using smartphones. In *MobiCASE: Fourth International Conference on Mobile Computing, Applications and Services*, 2012.

[20] B. Aydin, Y. Yilmaz, Y. Li, Q. Li, J. Gao, B. Aydin, and M. Demirbas. Crowdsourcing for multiple-choice question answering. *Twenty-Sixth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-14)*, 2014.

[21] Y. S. Yilmaz, B. Aydin, and M. Demirbas. Targeted question answering on smartphones utilizing app based user classification. *International Conference on Collaboration Technologies and Systems (CTS)*, 2014.

[22] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8(4), 2014.

[23] Y. S. Yilmaz, B. I. Aydin, and M. Demirbas. Google cloud messaging (gcm): An evaluation. In *Global Communications Conference (GLOBECOM), 2014 IEEE*, pages 2807–2812. IEEE, 2014.

[24] M. Demirbas, X. Lu, and P. Singla. An in-network querying framework for wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems*, 2009.

[25] M. Demirbas, A. Arora, and V. Kulathumani. A lightweight querying service for wireless sensor networks. *Theoretical Computer Science Journal, Elsevier*, 2009. To appear.

[26] M. Demirbas, A. Arora, T. Nolte, and N. Lynch. A hierarchy-based fault-local stabilizing algorithm for tracking in sensor networks. *8th International Conference on Principles of Distributed Systems (OPODIS)*, pages 299–315, 2004.

[27] V. Kulathumani, M. Demirbas, A. Arora, and M. Sridharan. Trail: A distance sensitive network service for distributed object tracking. *ACM Transactions on Sensor Networks*, 2009. To appear.

[28] M. Demirbas, O. Soysal, and A. S. Tosun. Data salmon: A greedy mobile basestation protocol for efficient data collection in wireless sensor networks. *IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 267–280, 2007.

[29] O. Soysal and M. Demirbas. Data spider: A resilient mobile basestation protocol for efficient data collection in wireless sensor networks. *The 6th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2010.

[30] S.K. Yoon, O. Soysal, M. Demirbas, and C. Qiao. Coordinated locomotion of mobile sensor networks. *Fifth Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, 2008.

[31] G. Chockler, M. Demirbas, S. Gilbert, N. A. Lynch, C. C. Newport, and T. Nolte. Consensus and collision detectors in radio networks. *Distributed Computing*, 21(1):55–84, 2008.

[32] M. Demirbas, O. Soysal, and M. Hussain. Singlehop collaborative feedback primitives for wireless sensor networks. *INFOCOM*, pages 2047–2055, 2008.

[33] J. Aspnes, E. Blais, M. Demirbas, R. O'Donnell, A. Rudra, and S. Uurtamo. k+ decision trees. *International Workshop on Algorithms for Sensor Systems, Wireless Ad Hoc Networks and Autonomous Mobile Entities (Algosensors)*, 2010.

[34] M. Demirbas, S. Tasci, H. Gunes, and A. Rudra. Singlehop collaborative feedback primitives for threshold querying in wireless sensor networks. *25th IEEE International Parallel & Distributed Processing Symposium (IPDPS)*, 2011.

[35] M. Demirbas and M. Hussain. A mac layer protocol for priority-based reliable broadcast in wireless ad hoc networks. *BroadNets*, 2006.

[36] M. Demirbas and S. Balachandran. Robcast: A singlehop reliable broadcast protocol for wireless sensor networks. *The Sixth International Workshop on Assurance in Distributed Systems and Networks (ADSN)*, 2007.

[37] M. Demirbas, O. Soysal, and M. Hussain. Transact: A transactional programming framework for wireless sensor/actor networks. *IEEE/ACM International Conference on Information Processing in Sensor Networks (IPSN)*, pages 295–306, 2008.

[38] O. Soysal, B. I. Aydin, and M. Demirbas. Optimistic concurrency control for multihop sensor networks. *IWCMC*, pages 89–94, 2011.