A selective and biased choice of techniques for building a distributed data store

Paweł T. Wojciechowski
Poznan University of Technology
Poland
Pawel.T.Wojciechowski@cs.put.edu.pl

ABSTRACT

Single-machine data stores cannot support the scale and ubiquity of data today. The Internet applications and services must process a huge number of concurrent requests and events per second. So, they use distributed (or replicated) data stores which store and process data on multiple machines, offering key advantages in performance, scalability, and reliability.

The purpose of the talk is to present a selective and biased choice of techniques and results which can be used for building an efficient distributed data store. Biased, because I only present solutions and results developed within a research project that I did with my PhD students. Selective, because an exhaustive description would be too exhausting to fit into a single talk. Therefore I will be discussing in detail just the design of our novel database index for key-value data store systems, and only skim our other contributions that are directly related to distributed systems.

The index, called Jiffy, has been designed with performance and scalability in mind. Therefore it has been designed as a lock-free concurrent data structure, which can dynamically adapt to the changing workload. It achieves superior performance despite built-in atomic operations (batch updates, snapshots, and range scans). During the talk I will be presenting Jiffy's architecture, the algorithms for inserting and looking up the key-value pairs, and the operations used for resizing the data structure dynamically.

The other contributions of our project include: efficient support for replica state recovery after failures, either by extending the classic Paxos consensus algorithm, or through the use of persistent memory, and some surprising theoretical results which are applicable to distributed data store systems that compromise consistency in favour of high availability and speed, but also support operations ensuring strong consistency (which requires consensus among replicas).

CCS CONCEPTS

• **Information systems** → **Distributed storage**; **Remote replication**; • **Theory of computation** → **Concurrent algorithms**.

KEYWORDS

Paxos, replica state recovery, mixed consistency, concurrent data structures, ordered index, lock-free skip list

ACM Reference Format:
Paweł T. Wojciechowski. 2023. A selective and biased choice of techniques for building a distributed data store. In *The 17th ACM International Conference on Distributed and Event-based Systems (DEBS '23), June 27–30, 2023, Neuchatel, Switzerland.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3583678.3603441

1 BIOGRAPHY

Paweł T. Wojciechowski received the Habilitation degree from Poznan University of Technology, Poland, in 2008, and the Ph.D. degree in computer science from the University of Cambridge, in 2000. He was a postdoctoral researcher with the School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, from 2001 to 2005, and with the University of Cambridge, in 2001. He is currently an Associate Professor with the Institute of Computing Science, Poznan University of Technology, where he leads the Concurrent Systems group. His research interests span topics in concurrency, distributed computing, and programming languages. He has published his research in leading journals, such as IEEE Transactions on Parallel and Distributed Systems, IEEE Transactions on Dependable and Secure Computing, Journal of Parallel and Distributed Computing, Distributed Computing, ACM Transactions on Programming Languages and Systems, and International Journal of Parallel Programming.

DEBS '23, June 27–30, 2023, Neuchatel, Switzerland

ACM ISBN 979-8-4007-0122-1/23/06.
https://doi.org/10.1145/3583678.3603441

Acknowledgments

This work was supported by the Foundation for Polish Science, within the TEAM programme co-financed by the European Union under the European Regional Development Fund, grant No. POIR.04.04.00-00-5C5B/17-00. We thank Intel Corporation and Intel Technology Poland for providing us with computing resources equipped with the Intel® OptaneTM DC persistent memory technology.