Architectures and Algorithms for Internet-Scale (p2p) Data Management

Joseph M. Hellerstein

EECS Computer Science Division, UC Berkeley
Intel Research, Berkeley

The database community prides itself on scalable data management solutions. In recent years, a new set of scalability challenges have arisen in the context of Internet-scale peer-to-peer (p2p) systems, in which the scaling metric is the number of participating computers, rather than the number of bytes stored. This is new and intriguing territory for the design of data management algorithms and systems.

The best-known application of p2p technology to date has been filesharing, which despite its sometimes unsavory use has been a vibrant technology driver. In addition to filesharing, there are compelling new application agendas for p2p systems including Internet monitoring, content distribution, distributed storage, multi-user games and next-generation Internet routing. The energy behind p2p technology has led to an academic renaissance in the distributed algorithms and distributed systems communities, much of which directly addresses issues in massively distributed data management.

Internet-scale systems present numerous unique technical challenges, including steady-state "churn" (nodes joining and leaving), the need to evolve and scale without reconfiguration, an absence of ongoing system administration, and adversarial participants in the processing. These challenges are not unique to what we commonly think of as p2p deployment scenarios. Hence many "p2p" techniques have relevance for any large distributed system in which the scale and distribution of the infrastructure makes traditional administrative models untenable.

In this tutorial we will focus on key data management building blocks including:

- Architectures for popular filesharing systems
- Indirection in time and space
- Structured overlay networks such as Distributed Hash Tables (DHTs), and their relationship to Interconnection Networks in parallel computers
- Embeddings of computations and communication in structured networks
- Persistence models for the Internet, including soft state and stronger guarantees
- Federated resource allocation
- Challenges in security and trust

We will also discuss motivations for the use of p2p technologies in both the popular conception of the term, and in related scenarios.

We will ground the presentation in experiences from deployed systems, including popular filesharing systems (Gnutella, KaZaA, BitTorrent), DHTs (Chord [7], Bamboo [6], Kademlia [5]), storage systems (LOCKSS [4], OceanStore [3]), and general-purpose query engines (PIER [2]). We will also discuss the PlanetLab [1] infrastructure for prototyping and deploying distributed systems.

References

[1] A. Bavier, L. Peterson, M. Wawrzoniak, S. Karlin, T. Spalink, T. Roscoe, D. Culler, B. Chun, and M. Bowman. Operating systems support for planetary-scale network services. In *Proc. 1st Symposium on Networked Systems Design and Implementation (NSDI)*, San Francisco, CA, Mar. 2004.

[2] R. Huebsch, J. M. Hellerstein, N. Lanham, B. T. Loo, S. Shenker, and I. Stoica. Querying the Internet with PIER. In *Proc. 19th VLDB*, Sep 2003.

[3] J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton, D. Geels, R. Gummadi, S. Rhea, H. Weatherspoon, W. Weimer, C. Wells, and B. Zhao. OceanStore: An Architecture for Global-Scale Persistent Storage. In *Proc. 9th ASPLOS*, Nov. 2000.

[4] P. Maniatis, M. Roussopoulos, T. Giuli, D. S. H. Rosenthal, M. Baker, and Y. Muliadi. Preserving peer replicas by rate-limited sampled voting. In *Proc. 19th ACM SOSP*, Oct. 2003.

[5] P. Maymounkov and D. Mazières. Kademlia: A peer-to-peer information system based on the XOR metric. In *Proc. 1st International Workshop on Peer-to-Peer Systems (IPTPS)*, Mar. 2002.

[6] S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz. Handling Churn in a DHT. In *Proc. USENIX*, June 2004.

[7] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications. In *Proc. ACM SIGCOMM*, 2001.

**Proceedings of the 30th VLDB Conference,
Toronto, Canada, 2004**