10 Network Papers that Changed the World

George Varghese
UCSD

Categories and Subject Descriptors: C.2.6 Internetworking : Routers

General Terms: Algorithms, Design, Network Protocols, Computer Systems

Keywords: Computer Networks, Networking History, Networking Literature

1. INTRODUCTION

In this list, I focus on papers that have had impact — that have changed the networking world. Of course, many commercial systems have done just that, so I also require that each paper has a memorable *idea*. Finally, I am drawn to papers that are well written, and in which the writing and the ideas stand the test of time. I break up papers in classic layered fashion, going bottom up. Because the list spans all levels of abstraction from Data Links to Applications, I hope this list of papers also provides a quick and inspirational overview of the world of networking systems for a beginning student.

2. DATA LINK PROTOCOLS

1. R. Metcalfe and D. Boggs, "Ethernet: distributed packet switching for local computer networks", *Communications of the ACM*, v.19 n.7, p.395-404, July 1976

The original Ethernet paper has a number of beautiful ideas and has had a lasting impact. When Ethernet was first proposed people did not believe in randomized algorithms, and token rings were considered more deterministic and "reliable". There are a number of subtle ideas. For example, the facile description of collision detection is that when two people talk at the same time they collide; in reality, detecting collisions is specific to a point in space and time. Thus different stations detect collisions at different times (shades of Einstein) and it takes care to ensure that there is universality of collision (if one station detects a collision, all stations do). The backoff algorithm has become part of our lexicon, and there is even a performance analysis. Although what is called Gigabit Ethernet only masquerades under the same name, earlier versions of Ethernet (10M and 100M) have had a huge impact on technology. This took place despite early papers denouncing the unreliablity and poor performance of Ethernet (both untrue).

2. J. Saltzer, D. Clark, and K. Pogran. "Why a Ring?", *IEEE Seventh Data Communications Symposium*, October 1981, pp. 211-217.

During the transition to fiber, when fiber could not be tapped as Ethernets require, it was natural to consider rings which use point-to-point links. But rings had a fatal flaw: the failure of one station could fail the ring. In this insightful paper, the authors propose (among other things) separating the logical topology (a ring) from the physical topology (a star shaped ring using central concentrators that could bypass failed stations). This idea was used in all rings including FDDI. The separation of logical and physical topologies continues with hubs and star shaped wiring for local networks.

3. A.G. Fraser, "Towards a Universal Data Transport System", *IEEE Journal on Selected Areas in Communication*, SAC-1, 5, Nov 1983.

Sandy Fraser's paper has two very interesting ideas. First, it provides an unusual interface to end users, that of a bit faucet, which is different from the usual packet interface that most students feel is the only possible one. Second, under the hood, this abstraction is implemented using cells, an unusual LAN, and virtual circuits, foreshadowing ATM technology. Whatever your feelings about ATM, most routers are internally cell-switched, and virtual circuit ideas (e.g., labels in MPLS) abound even in IP networks.

3. ROUTING PROTOCOLS

4. R. Perlman, "An algorithm for distributed computation of a Spanning Tree in an Extended LAN", *Proceedings of the 9th Symposium on Data Communications*, v. 20, n. 7, 1985, pp. 44-52.

This is a beautiful paper and none of the networking textbooks (including Perlman's own text) do justice to this algorithm and its many layers of subprotocols. At the basic centralized level there is a way to build a tree by choosing a root and the bridge that offers the shortest path to the root as a parent. Then there is a simple distributed algorithm to compute the root and distances in a single pass. However, the subtlety comes in with the failure protocol to recover from a lost root (all the textbooks ignore this). Unlike distance vector which counts up to combat failure, this one uses timers and there are so many ways to get this wrong (try designing it yourself!). Added on for good measure are countless little subprotocols such as mechanisms for dealing with wholesale station movement after topology changes by making everyone reduce their timers. While the bridge itself was invented by Mark Kempf (a lovely idea but no paper, only a patent), Spanning Tree has become the standard for a billion dollar bridging market. While people have worked on somewhat inelegant extensions for fast reconfiguration, the original paper stands out for its simplicity, elegance, and thoroughness.

5. R. Perlman, "Fault-Tolerant Broadcast of Routing Information", *Computer Networks*, vol 7, 1983, pp. 395-405.

While a classic paper, this one is admittedly derivative from the original Arpanet paper from McQuilan et al[2] which really introduced Link State Routing. But perhaps it is fair to say that this paper pointed out the issues in the original proposal, and helped make Link State Routing work. There is a nice idea for fault tolerance which involves sending back newer information to people sending you older information (as opposed to merely ignoring the old information) and the distributed consequences (non-trivial to analyze). Although the lollipop-based sequence number was later simplified to a simple linear space, this paper was the basis for OSI Routing, and then (via changes from John Moy) to OSPF. And OSPF is the IGP of choice for most ISPs.

6. Y. Dalal and R. Metcalfe, "Reverse path forwarding of broadcast packets", *Communications of the ACM*, v.21, n.12, Dec 1978, pp.1040-1048.

Was it Jacobi who said "you should always invert a problem"? The idea of Dalal and Metcalfe inverts the problem of delivering multicast *to* a set of receivers to accepting multicast *from* a source. This allows an elegant reduction of the multicast routing tree computation to shortest path routing rooted at the source of the multicast. After all these years, this paper still delights me with its simplicity. Of course, this paper influenced Deering [1] who added a number of ideas like pruning to get to DVMRP and IP multicast. And IP multicast, dare we say it, is making a comeback.

4. TRANSPORT PROTOCOLS

7. R. Tomlinson, "Selecting Sequence Numbers", *Proceedings of the ACM SIGCOMM/SIGOPS Interprocess Communications Workshop*, Santa Monica, CA, March 24, 1975.

While the TCP papers are deservedly classic, and Cerf and Kahn clearly deserved their Turing Award, an elegant and enduring idea is the idea of a 3-way handshake. If a client sends a sequence number in the past, the simplest idea is for the server to remember all past sequence numbers to guard against delayed duplicates. This can greatly increase the memory at the server. Tomlinson's elegant way out, used in TCP, is to have the server send a number never used recently (nonce) and only believe the client if it echoes back the number. This is almost like a challenge-response sequence but it guards against duplicates and not attackers. While the first proposal had some warts, and further refinements were added by Yogen Dalal, this paper is still the genesis of a memorable idea in networking, one that is used a zillion times, every time a user starts a TCP connection.

8. Richard W. Watson, "Timer-Based Mechanisms in Reliable Transport Protocol Connection Management", *Computer Networks* 5, 1981, pp. 47-56.

In contrast to the TCP method for connection establishment (which trades off latency in opening a connection for reduced storage of past sequence numbers), Watson's paper carefully thinks through the alternative to 3-way handshakes. I feel strongly that students need to study alternative mechanisms to even classic solutions such as 3-way handshakes. Besides, timer-based mechansism have influenced a number of transports such as those in VMTP and those underlying many RPC protocols.

5. NETWORK ABSTRACTIONS

9. A. Birrell and B. Nelson, "Implementing Remote Procedure Calls", *ACM Transactions on Computer Systems*, vol. 2, No, 1, October 1984, p. 39-58

This paper is mentioned in at least two of the previous CCR lists. This is fine because I do believe this is a must-read paper. However, perhaps the reasons I like it are different from that of other reviewers. I like it most of all because it provides an alternative API to TCP's socket queue interface. More fundamentally, it suggests a line of thinking of extending other IPC mechanisms *within* hosts (e.g., Shared Memory) to similar mechanisms *across* hosts (e.g., Distributed Shared Memory). Beyond the innovative abstraction, one has to solve new problems (e.g., procedure calls within a host do not have to deal with the issue of partial failure). The implementation has also to be fast if people are to use it. Although a throwaway point, the idea of active messages[3] is at least foreshadowed in this paper by the suggestion that the RPC message carry the address of the interrupt handler.

10. N. Kronenberg, H. Levy and W. Strecker, "VAX clusters: A closely-coupled distributed system," *ACM Transactions on Computer Systems*, vol. 4, no. 2, May 1986, pp. 130-146.

Legend has it that Bill Strecker of the VAX Architecture group asked to learn about networking and was pointed to a networking text. A few months later, he and his colleagues had created VAX Clusters, a way of hooking up a bunch of computers to storage via a new LAN. Their twist on the Ethernet style Local Area Network is interesting. But the most elegant idea (and one that only a computer architect would think of) is the idea of extending Direct Memory Access (DMA) to RDMA (Remote DMA across the network). Added for gravy is a very elegant Distributed Lock Manager implementation. Many networking students are unaware of a billion dollar industry in Storage Area Networks (SANs). Modern day incarnations of the RDMA idea can be found in things like Fiber Channel, Infiniband, and iSCSI.

6. ACKNOWLEDGEMENTS

Thanks to Jim Kurose who persuaded me to do this and then provided very useful comments.

7. REFERENCES

[1] S. Deering and D. Cheriton, "Multicast Routing in Datagram Internetworks and Extended LANs," *ACM Transactions on Computer Systems*, vol. 8, No. 2, May 1990, pp. 85-110

[2] J. McQuillan, I. Richer, and Eric C. Rosen, "The New Routing Algorithm for the ARPANet", *IEEE Transactions on Communications*, 28(5), 1980, pp. 711-719.

[3] T. von Eicken, D. Culler, S. Goldstein, and K. Schauser. "Active messages: A Mechanism for integrated communication and computation", *Proceedings 19th International Symposium on Computer Architecture*, 1992, pp. 256-266.