SPARTA: Scalable Per-Address RouTing Architecture*

Colin Dixon† Brent Stephens‡ Wes Felter† John Carter† Alan Cox‡
† IBM Research—Austin ‡ Rice University

Modern data centers demand network features that current network designs struggle to meet. For example, layer-2 Ethernet networks provide the flexibility and ease of configuration that network operators want, but they scale poorly and make poor use of available bandwidth. Layer-3 IP networks can provide better scalability and bandwidth, but are less flexible and are more difficult to configure and manage.

Concretely, modern data center networks should meet the following four functional requirements [2, 6, 9, 11].

1. **Host mobility:** Hosts, especially virtual hosts, must be able to move without interrupting existing connections.
2. **Exploit available bandwidth:** Flows should not be throughput-limited if there is available bandwidth on other paths.
3. **Self-configuration:** Routers and (v-)switches must be able to forward traffic without manual configuration.
4. **Scalability:** The network should scale to modern data centers without violating the preceding requirements.

In addition to these functional requirements, we limit our design space to architectures that can be implemented and managed efficiently with commodity hardware and software, which leads to three additional design requirements:

1. **No hardware changes:** The architecture must work with commodity networking hardware.
2. **Respects layering:** The architecture must work with unmodified software stacks, e.g., operating systems and hypervisors, and higher layers must not need to understand details of the architecture's implementation.
3. **Topology independent:** The architecture must work with arbitrary topologies, i.e., not just Fat Trees.

Table 1 compares existing data center network architectures and recent academic work against the above requirements. No existing architecture meets all of them. One reason is that the requirements often conflict with one another. For example, Ethernet's distributed control protocol provides host mobility with little or no configuration. However, it does not scale well beyond roughly 1000 hosts as it uses broadcast for name resolution. Further, it makes poor use of available bandwidth because it uses a single spanning tree for packet forwarding—a limitation imposed to avoid forwarding loops.

To address these problems, current large data center networks connect multiple Ethernet LANs using IP routers [3]. The IP routing algorithms allow for shortest path and Equal-Cost Multipath (ECMP) routing, which provide more usable bandwidth than Ethernet's spanning tree. However, the mixed layer-2/layer-3 solution requires significant manual configuration and (typically) limits host mobility to be within a single LAN.

The trend in recent work to address these problems is to introduce special hardware and topologies. For example, PortLand [9] is only implementable on Fat Tree topologies and requires ECMP hardware, which is not available on every Ethernet switch. TRILL [10] introduces a new packet header format and thus requires new hardware and/or firmware features.

We pose the following question: Are special hardware or topologies necessary to implement a data center network that meets our requirements, or can we build such a data center network with only commodity Ethernet hardware?

Surprisingly, we find that we *can* build a data center network that meets all of the requirements using only basic Ethernet switch functionality. Contrary to the suggestions of recent work, special hardware and restricted topologies are *not* necessary.

To prove this point, we present SPARTA, a flat layer-2 data center network architecture that supports full host mobility, high end-to-end bandwidth, autonomous route construction, and tens of thousands of hosts on top of common commodity Ethernet switches. SPARTA satisfies our functional and design requirements as follows. When a host joins the network or migrates, a new spanning tree is installed to carry traffic destined for that host. This spanning tree is implemented using only entries in the large Ethernet (exact match) forwarding table present in commodity switch chips, which allows SPARTA to support as many hosts as there are entries in that table. In aggregate, the trees spread

*This is an extended abstract for previously published work [13].

Architecture	Functional Requirements				Design Requirements		
	Mobility	High BW	Self Config	Scales	No H/W Changes	Respect Layers	Topo Ind
Ethernet with STP	✓	X	✓	X	✓	✓	✓
IP (e.g. OSPF)	X	✓	X	✓	✓	✓	✓
MLAG [7]	✓	✓	✓	X	✓	✓	✓
SPAIN [8]	✓	✓	✓	X	✓	X	✓
PortLand [9]	✓	✓	✓	✓	✓	✓	X
VL2 [4]	X	✓	X	✓	✓	X	X
SEATTLE [6]	✓	X	✓	✓	✓	✓	✓
TRILL [10]	✓	✓	✓	X	X	✓	✓
EthAir [11], VIRO [5]	✓	X	✓	✓	✓	✓	✓
SPARTA	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of data center network architectures.

traffic across all links in the network, so SPARTA provides aggregate bandwidth equal to or greater than layer-3 ECMP routing. SPARTA provides Ethernet semantics and runs on unmodified switches and hosts without appropriating the VLAN ID or other header fields. Finally, SPARTA works on arbitrary network topologies, including HyperX [1] and Jellyfish [12], which can perform as well as or better than Fat Tree [9] topologies at a fraction of the cost.

Our OpenFlow-based SPARTA implementation was crafted carefully to utilize the kinds of match-action rules present in commodity switch hardware, the number of rules per table, and the speed with which rules can be installed. By restricting SPARTA to route solely using destination MAC addresses and VLAN tags, we can use the large layer-2 forwarding table, rather than relying on the more general, but much smaller, TCAM forwarding table, as is done in previous OpenFlow architectures.

We defer a complete evaluation of SPARTA to the previously published work [13], but present a summary here. In broad strokes, SPARTA provides performance equal to or better than the state-of-the-art ECMP routing without requiring layer-3 routing or multi-pathing hardware support. It does so on arbitrary topologies; we show results for Fat Tree, HyperX and Jellyfish topologies. Further, a variant of SPARTA which constructs non-minimal trees outperforms both ECMP and Valiant load balancing for workloads that are hotspot-prone under minimal routing.

Overall, through careful design informed by real commodity switching hardware, SPARTA provides a flat layer-2 data center network architecture that supports full host mobility, high end-to-end bandwidth, self-configuration, and tens of thousands of hosts using Ethernet switches built from today's commodity switch chips.

References

[1] J. H. Ahn, N. Binkert, A. Davis, M. McLaren, and R. S. Schreiber. Hyperx: topology, routing, and packaging of efficient large-scale networks. *SC Conference*, 2009.

[2] I. Gashinsky. SDN in warehouse scale datacenter v2.0. In *Open Networking Summit*, 2012.

[3] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel. The cost of a cloud: Research problems in data center networks. In *ACM CCR*, January 2009.

[4] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta. VL2: A scalable and flexible data center network. In *SIGCOMM*, 2009.

[5] S. Jain, Y. Chen, Z.-L. Zhang, and S. Jain. Viro: A scalable, robust and namespace independent virtual id routing for future networks. In *INFOCOMM*, 2011.

[6] C. Kim, M. Caesar, and J. Rexford. Floodless in SEATTLE: A scalable Ethernet architecture for large enterprises. In *Proceedings of ACM SIGCOMM*, 2008.

[7] MC-LAG. http://en.wikipedia.org/wiki/MC_LAG.

[8] J. Mudigonda, P. Yalagandula, M. Al-Fares, and J. C. Mogul. SPAIN: COTS data-center Ethernet for multipathing over arbitrary topologies. In *NSDI*, 2010.

[9] R. N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat. PortLand: A scalable fault-tolerant layer 2 data center network fabric. In *SIGCOMM*, 2009.

[10] R. Perlman. Rbridges: Transparent Routing. In *INFOCOMM*, 2004.

[11] D. Sampath, S. Agarwal, and J. Gacia-Luna-Aceves. 'Ethernet on AIR' : Scalable Routing in Very Large Ethernet-based Networks. In *ICDCS*, 2010.

[12] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey. Jellyfish: Networking data centers randomly. In *NSDI*, April 2012.

[13] B. Stephens, A. Cox, C. Dixon, W. Felter, and J. Carter. PAST: Scalable Ethernet for Data Centers. In *CoNEXT*, 2012.