Chinese Science Bulletin

November 2012 Vol.57 No.32: 4244
doi: 10.1007/s11434-012-9912-3

A new high performance and fault-tolerant datacenter network for modular datacenters proposed by researchers from NUDT

A modular datacenter network (MDCN) is the key component in building mega-datacenters. Huang Feng, Li Dongsheng, and their group from the National Key Laboratory of Parallel and Distributed Processing, School of Computers, National University of Defense Technology present a novel hybrid intra-container network for a modular datacenter (MDC), called SCautz, together with a suite of routing protocols. SCautz is able to provide high network throughput for various traffic patterns, and achieves graceful performance degradation when failures occur and increase. Their work, entitled "SCautz: a high performance and fault-tolerant datacenter network for modular datacenters", has been published in *Science China Information Sciences*, 2012, Vol. 55 (7).

To construct mega-datacenters, the MDC first packages thousands of servers, which have been set up and wired, into one shipping container, and then connects the containers together by viewing them as large pluggable building blocks. Once connected to power, a cooling infrastructure, and the Internet, the MDC can provide services at any location in the world.

Traditional datacenter networks (DCN) interconnect large numbers of servers directly, making these difficult to realize and maintain. An MDCN consists of intra- and inter-container networks, which largely simplify the design and implementation. Typically, a standard 20 or 40 foot shipping container is equipped with 1200–2500 servers, with the number of servers in a container fixed during its lifetime. A container's moderate scale relaxes the restriction on the scalability of DCNs. So, intra-container networks can adopt more complex topologies. Although the Kautz graph achieves near-optimal tradeoff between node degree and diameter, and has better bisection width and bottleneck degree, it is considered to be unsuitable for traditional DCNs, due to the difficulty of not violating the original structure in its incremental deployment. For an MDC, on the other hand, the number of servers in a single container is fixed, and the inner network does not change throughout its lifespan. Thus, SCautz was designed by modeling the Kautz graph.

Containerization lowers the total cost of ownership for cloud providers, and allows operators to manage the MDC using a particular "service-free" model. In other words, a container as a whole is never repaired during its deployment lifespan (typically 3–5 years). As long as the performance of the entire container meets an engineered minimum criterion, there is no continuous component repair. However, server failures not only decrease the MDC's computation and storage capacity, but also destroy the MDCN's structure. For instance, BCube is an excellent network architecture for current MDCs. However, due to its incomplete structure, the throughput for one-to-x traffic patterns drops noticeably, while the ABT (aggregate bottleneck throughput) for all-to-all traffic degrades faster than the computation and storage. Furthermore, switch failures decrease BCube's performance much more significantly, since its ABT deteriorates by more than 50% in the presence of 20% switch failures. The fact that network performance degrades faster than computation or storage capacity is the MDCN's ultimate weakness, since this causes the container's overall performance to decrease below the threshold criterion and end its lifespan prematurely.

Therefore, based on the "scale out" principle, SCautz presents a hybrid structure for dealing with server and switch faults. This comprises a base physical Kautz topology, which is built by interconnecting the servers' NIC ports, and a small number of redundant commercial off-the-shelf (COTS) switches. In SCautz, each switch and a specific number of servers form a "cluster". Since switches are divided into two types according to their identifiers, "clusters" are also divided into two types. If viewed as logical nodes, the two types of "clusters" construct two higher-level logical Kautz structures, respectively. SCautz's hybrid structure has the following advantages. First, SCautz can run in different modes with switches on or off. Its base topology provides high capacity for processing different types of traffic patterns, and behaves as well as BCube. Moreover SCautz's complete structure is able to improve performance twofold through the use of switches, thereby providing the ability to handle bursts of network flows effectively without lowering the quality of bandwidth-intensive applications. Second, SCautz improves the fault tolerance of an MDCN by using redundant switches. When a server fails, SCautz finds a peer server in the same cluster to bypass the failed one. Thus, it can maintain the throughput for one-to-x traffic (e.g., one-to-one, one-to-all), and reduce the ABT loss by about a half so that network performance degrades much more slowly than the MDC's computation or storage capacity. Third, the extra cost of the redundant design is very low. Theoretical analysis shows that a typical SCautz-based container with 1280 servers only needs 160 COTS switches.

The design and implementation of SCautz was a collaborative effort involving many researchers. It is an important breakthrough in modular datacenter network architectures. The researchers suggest that their work needs to be implemented and examined in a larger production datacenter. This work will have significant impact on datacenter construction and cloud computing.

This work was supported by the National Basic Research Program of China (*2011CB302600*) *and the National Natural Science Foundation of China* (*60903205*).

See the article: Huang F, Lu X C, Li D S, et al. SCautz: a high performance and fault-tolerant datacenter network for modular datacenters. Sci China Inf Sci, 2012, 55: 1493–1508