Computer Networks 77 (2015) 56–72

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Tradeoff study among traffic egression schemes and member allocation optimization for link aggregation groups in integrated switching systems

Dexiang Wang

Juniper Networks, Inc., Sunnyvale, CA 94089, USA

ARTICLE INFO

Article history:
Received 10 June 2014
Received in revised form 24 October 2014
Accepted 1 December 2014
Available online 5 December 2014

Keywords:
Bandwidth optimization
Designated forwarding
Integrated switching system
Load balance
Local biasing

ABSTRACT

An integrated switching system (ISS) is an integration of individual switching devices with newly developed control and management protocols, in an integration topology that interconnects the switching devices. It scales the capacity of individual switching devices while functions still as a single switching facility to the outside network. In an ISS, a link aggregation group (LAG) has flexibility to allocate its member links across different switching devices, often as an attempt to spread risk of failure to individual switching devices. Due to existence of the integration topology, traffic admitted into the ISS and destined to egress out of the LAG interface may need to travel a number of hops before leaving the ISS. This lays a bandwidth burden on the integration links of the ISS. Local biasing and designated forwarding have been proposed as LAG egression options to relieve the bandwidth pressure on integration links. They bias traffic egression to be from the neighborhood of the switching device where the traffic ingresses, if the traffic is destined to the LAG interface and a LAG member link is present in the neighborhood. Those enhanced LAG egression schemes are in contrast to the regular LAG egression scheme, in which the traffic will be split and evenly distributed across all LAG members, regardless of their topological distance to the switching device where it ingresses. Although the enhanced egression schemes help to reduce the bandwidth demand on the integration links, there comes a price that load balance across LAG members may be sacrificed and eventually stable LAG capacity could be compromised. In this paper, we study such performance tradeoff and investigate impact factors on LAG performance. In the end, we formulate an optimization problem that optimizes LAG member allocation via pursuing the best tradeoff between integration bandwidth utilization and stable LAG capacity. The solution can be treated as a guideline to deploy LAGs in an ISS. The results show, with optimized LAG member allocation, the potential of integration bandwidth saving and stable LAG capacity maintenance can be maximally explored.

1. Introduction

Many applications of modern packet switching systems, such as enterprise networks [1–3] and data center networks [4–6], demand advanced features like large scale, high capacity, low latency, and ease of management. Those features are not affordable by traditional standalone switching devices. Integrated switching systems (ISSs) become a widely accepted solution to address those challenges.[1] It provides flexible integration of individual switching devices with the help of newly developed control

[1] Branded examples of ISSs include Juniper Networks' virtual chassis technology [7], Cisco Systems' switching stacks [8], etc.

E-mail address: camelwdx@ufl.edu

http://dx.doi.org/10.1016/j.comnet.2014.12.002
1389-1286/

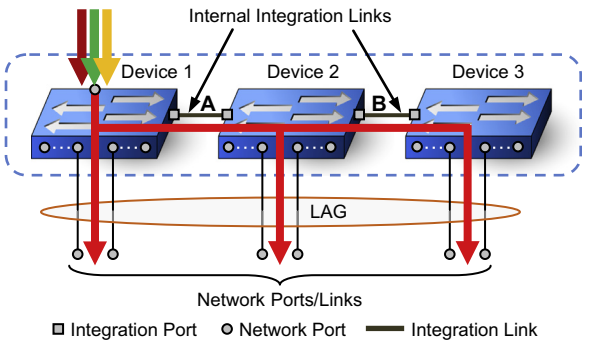

Fig. 1. An integrated switching system (ISS) and a link aggregation group with members across the ISS and with hashing based egression scheme.

and management protocols. A network operator can easily manage large-scale network resources (ports, VLANs, etc.) across different switching devices within the ISS. An illustrative ISS is abstracted in Fig. 1, which consists of three switching devices (numbered 1, 2, and 3). The three devices are connected via internal integration links (A and B), which form a *line* topology (we call such topology the integration topology). The network ports (usually a few dozen per switching device) are configurable to admit traffic from or to switch traffic to the outside of the ISS. Note that by integrating the switching devices together, those devices functionally become one single switching system to the outside of the ISS.

A link aggregation group (LAG) is a bundle of network links that are treated as a single logical link [9]. Fundamentally, it addresses two problems with Ethernet connections: bandwidth limitation and lack of resiliency. Once a member link incurs a failure, the traffic that it carries can be automatically switched over to other active members in the LAG, with assistance of the link aggregation control protocol (LACP) [10]. In an ISS, a LAG configuration has flexibility to allocate its member links across different switching devices as an attempt to spread risk of failure to individual switching devices. As shown in Fig. 1, six links are allocated to be members of a LAG. The traffic flows that ingress from another network port (as shown on device 1), if destined to the LAG, will be switched out through the LAG member links. The selection of LAG members to carry the packets in the traffic is decided by a hashing algorithm, which takes a set of hashing parameters associated with the traffic flows (distinguished by different colors[2] in Fig. 1) to form the hash key. Such hashing parameters can be MAC addresses, IP addresses, VLAN ID, ingress port number, etc. A perfect hashing algorithm will evenly spread the keys out such that the traffic can be evenly distributed across the LAG members. The order of packets within a specific flow can be maintained after hashing since all the packets in that flow are of the same hash key and hence will be hashed toward the same LAG member link. We call the device from which the traffic is admitted to the ISS the *ingress device*, while call the device from which the traffic leaves the ISS the *egress device*. In the example shown in Fig. 1, device 1 is the *ingress device* for the three traffic flows. Devices 1, 2 and 3 are the *egress devices* for those flows, depending on to which device the flow is hashed. Within the integration topology, the packet flow travels along the shortest path from its *ingress device* toward its *egress device*. For example in Fig. 1, if the flow is hashed to egress from a LAG member on device 3, since the shortest path from device 1 to device 3 is through device 2, the packets in that flow have to take bandwidth on integration links A and B.

[2] For interpretation of color in Fig. 1, the reader is referred to the web version of this article.

1.1. Regular LAG egression scheme

With perfect hashing, each of the three devices in Fig. 1 will be responsible for switching out one third of the traffic that is admitted into the ISS on device 1, as shown in Fig. 1, since all three devices host the same number of LAG members. As a result, one third of the traffic will egress out of device 1 without utilizing integration links, one third of the traffic will egress out of device 2 by passing through integration link A, and other one third of the traffic will egress out of device 3 by passing through both integration links A and B. Therefore, statistically, per unit of traffic intensity (say 1 bit/s) offered by the ingress traffic, 2/3 units of bandwidth (in bits/s) is taken on integration link A, while 1/3 units of bandwidth is taken on integration link B. In total, such traffic forwarding distribution generates $(2/3) + (1/3) = 1$ unit of bandwidth demand on the entire integration network, per unit of the ingress traffic intensity. We call this egress distribution scheme the *regular LAG egression scheme*.

1.2. Local biasing (LB)

In order to save stringent integration bandwidth (the bandwidth provision offered by the integration network), local biasing (LB) can be applied to the LAG interface, which forces traffic to egress out of and get balanced across the LAG member links on its *ingress device* if the *ingress device* hosts local member(s) of the LAG to which the traffic is destined. An example is demonstrated in Fig. 2(a). The same LAG of six member links as in Fig. 1 is formed. However, the traffic destined to the LAG interface will not be evenly spread across the three devices but egresses out locally from its ingress device (device 1) and gets balanced across the two local LAG members of device 1. The resulting benefit is that no integration bandwidth is demanded, as compared with the *regular LAG egression scheme* as shown in Fig. 1. The saved integration bandwidth can be used to accommodate other traffic and hence to improve the overall capacity of the ISS. If the traffic ingresses on a device that does not host LAG members, the traffic will be hashed as it is in the *regular LAG egression scheme* (i.e., the traffic will be evenly distributed across all LAG members). This is illustrated by a LAG of four member links in Fig. 2(b). We call this egress distribution scheme the *local biasing LAG egression scheme*. Note that the traffic destined to the LAG interface may possibly ingress on any switching devices (not just device 1 in Fig. 2). The egression scheme as described above holds for all the traffic.