OSN-IX: A Novel Internet eXchange (IX) Architecture based on Overlaid-Star Networks

Peng He and Gregor v. Bochmann

School of Information Technology and Engineering (SITE)
University of Ottawa, Ottawa, ON, K1N 6N5, Canada
{penghe, bochmann}@site.uottawa.ca

Abstract— We propose a novel Internet Exchange (IX) architecture, namely OSN-IX, which adopts an overlaid-star network (OSN) as an IX. OSN can be considered as a "distributed" switch, which combines the advantages of the network and switch. Compared to other IX architectures, e.g., LAN-based IX, MPLS IX, Photonic IX, etc., OSN-IX has good properties of scalability, resilience, and widely distributed access points. Particularly, for the first time, OSN-IX introduces traffic engineering (TE) into the IX world. Based on the TE framework we developed for OSN-IX, OSN-IX can provide optimized dynamic inter-ISP (Internet Service Provider) routing while requiring no change, hardware or software, on existing traditional IP/MPLS routers. We have shown by simulation that our TE framework outperforms several existing inter-AS TE schemes.

***Index Terms*—Internet Exchange, MPLS, Inter-domain Traffic Engineering, Overlaid-Star Networks.**

I. Introduction

THE INTERNET is a worldwide network of networks. As the Internet grows, the Internet eXchange (IX) plays an important role in supporting the Internet backbone. This is because IX is a place where Internet Service Providers (ISPs) (normally Autonomous IP Systems (AS)) can interconnect their networks and exchange Internet traffic between each other. The exchanging of inter-ISP traffic on an IX is known as "peering" [1].

The direct way to implement peering between two ISPs' networks is to build physical links connecting them. However, this will lead to an n-squared scalability problem if the number of these ISPs, n, is large. By adopting an Internet Exchange in the middle of these ISPs' networks (as shown in Fig. 1), a star topology instead of a full-mesh, the n-squared issue is solved. In addition, the ISPs can set up peering with each other in a free and efficient way through the IX.

There are many large and fast growing Internet Exchanges in the world, either non-profit or commercial, e.g., AMS-IX (Amsterdam Internet Exchange [1]), Japan Internet Exchange [2], Switch and Data (U.S.) [3], etc. In Europe, there are now more than 30 IXes and over 1,600 connected networks to these IXes [4]. In May 2001, Euro-IX (European Internet Exchange Association) was formed with the intention to further develop, strengthen and improve the Internet Exchange community [4].

The majority of IXes opted for a layer 2 switched Ethernet LAN architecture, with only a few IXes using ATM or FDDI. Meanwhile, several new architectures for IX are also proposed, e.g., IPv6 IX [5], MPLS-IX [6], photonic IX [7], etc. We analyze the pros and cons of these architectures in Section II. In Section III, we propose a novel IX architecture: Overlaid-Star Networks (OSN) based Internet Exchange, OSN-IX. OSN-IX has outstanding performance in three key aspects, namely switching capacity, scalability, and resilience. Section IV proposes a traffic engineering (TE) framework for OSN-IX: through introducing the concept of "virtual AS Border Router" (v-ASBR) into the control plane, OSN-IX can provide optimal dynamic inter-ISP (Inter-AS, cross-IX) MPLS routing while requiring no change, hardware or software, on existing traditional IP/MPLS routers in the networks of the participating ISPs. As far as we know, OSN-IX is the first Internet Exchange architecture that considers inter-ISP (Inter-AS) TE. Related simulation results are presented in Section V, and Section VI concludes the paper.

II. Review of Current IX Architectures

Most existing IXes are implemented in an Ethernet LAN architecture consisting of Ethernet switches, shown in Fig. 1.

A. *Ethernet-LAN-based Internet Exchange*

In Ethernet-LAN-based IXes, the Ethernet switches are normally organized into two sets, and each set includes both the access switches and powerful core switches with physical links connecting them. The two sets work either in load-sharing mode or in active/backup mode with VSRP (Virtual Switch Redundancy Protocol) running to define the active set and to automatically switch to the backup set based on pre-defined triggers (e.g., link failure). As shown in Fig. 1, the LAN based IXes just provides a shared switching infrastructure (peering LAN) for the customer ISPs to set up peering with each other. Technically, IXes usually use VLAN (virtual LAN) configuration (port-based and/or tag-based VLAN) to build up the peering among customer ISPs [1,2,3].

Normally, only the BGP routing protocol is allowed in an IX, thus the traffic over the IX is exchanged based on BGP routes. An ISP's local traffic is not allowed to pass through the IX. Beyond a peering LAN, another common service that an Ethernet-LAN-based IX can provide is a route server (RS) [8]. As seen in Fig. 1, a physical server located at the IX site functions as the route server. Instead of maintaining separate eBGP sessions to each of its peers' routers, the customer ISPs just need to peer with the route server and can thus opt to send their reachable IP prefixes to specific other peers of the route server, or all of them. Hence, the route server solves the scalability issue of a complete mesh of peering sessions

978-1-4244-1785-8

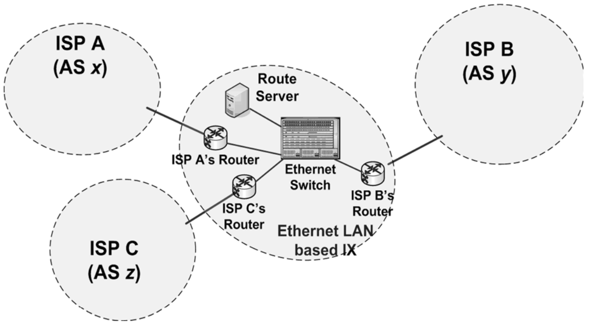

Fig. 1. A conceptual diagram of Ethernet LAN based IX [1,2,3].

and lowers the barrier of entry for new participants on the IX peering platform. Note, however, that the route servers do not forward packets among the routers attached to an IX.

The current Ethernet LAN based IX architecture has the following main problems: (1) The VLAN configuration is static, can not be adaptive to traffic change. (2) No traffic engineering is considered in the IX, the ISP networks are independent of IX, can only do over-provisioning most of time. (3) No tunnel technology in the Ethernet shared switching infrastructure, no point-to-point connection; can not guarantee hard/strict QoS. (4) Not-easy to extend the IX's capability: If the number of core switches in an IX is increased, this will lead to complex internal connection issues among these cores; if a more powerful switch is adopted, it will take higher risk under equipment failure, or terrorism attack. Meanwhile, O-E-O conversion at the core Ethernet switches will also limit the switching speed and capacity.

B. *Several Proposed New IX Architectures*

1) An IPv6 Internet Exchange Model [5]

The authors in [5] presented a possible architectural Internet Exchange model that explores the benefits of the IPv6 hierarchical and aggregatable routing and addressing model. The IX functionalities might be enhanced due to the IX directly assigning addresses to its customers' networks.

2) MPLS-IX [6]

In [6], MPLS technology was used in the IX architecture to interconnect ISPs, namely MPLS-IX. MPLS-IX has the advantages of the independence of data-link mediums and widely distributed access points [6]. Similarly, GMPLS, the generalized version of MPLS, can also be introduced into the IX, see GMPLS-Based Exchange Points [9].

3) Photonic IX [7]

Photonic IX [7] employs a flat and simple OXCs (optical cross-connects) network (e.g., ring) as IX core with GMPLS control. The ISPs' ASBRs use the optical user network interface (OUNI) signaling between the ASBRs and OXCs for (ASBR-IX-ASBR) optical path set-up/tear-down.

Generally speaking, there are two common potential drawbacks in the above proposed IX architectures:

(1) Complex IX internal routing: large number of core routers/switches in an IX will greatly increase the internal complexity of an IX, e.g., routing issues among these cores.

(2) Lack of TE capability: none of the above IX architectures considers inter-ISP traffic engineering.

III. Internet Exchange Based on Overlaid-Star Networks

In this section, we propose a new architecture for Internet Exchange, namely OSN-IX (Overlaid-Star Network based IX). We first introduce the OSN, then describe our OSN-IX architecture. After that, we present a traffic engineering framework based on the concept of "virtual AS Border Router" (v-ASBR). This TE framework can provide inter-ISP (ISP—OSN-IX—ISP) optimal dynamic routing. We assume that the ISP networks are IP/MPLS networks; this is reasonable since IP/MPLS has been deployed widely in most ISP core networks and has already become a common inter-networking technology.

A. *Overlaid-Star Networks(OSN)*

An overlaid-star network (OSN) is a network that comprises edge nodes interconnected by core nodes that function independently from each other to form an overlaid-star (also called composite-star) topology. Three reprehensive examples of overlaid-star networks are Agile All-Photonic Networks [10, 11], Ethernet Star Networks with TE Capability, and PetaWeb [12].

1) Agile All-Photonic Networks (AAPN)

As shown in Fig. 2, an Agile All-Photonic Network (AAPN) consists of a number of hybrid photonic/electronic edge nodes connected together via several (at least two) load-balancing core nodes and optical fibers to form an overlaid star topology. By introducing concentrator devices, AAPN can support a large number of edge nodes (up to 1024) [10]. Each core node contains a stack of bufferless transparent photonic space switches (one for each wavelength). In order to avoid optical memory and optical header recognition (hence no E-O-E conversion within AAPN), also as required by certain forms of burst switching, AAPN adopted synchronous slot-by-slot switching which uses fixed-size timeslots (e.g., 10μs per timeslot). Due to the simple star topology, the only synchronization requirement is that an edge node transmits the next slot at such a time that it arrives at a core switch in time for the beginning of a slot period. This can be realized by a relatively simple synchronization protocol (as compared with the case in the meshed optical networks) between the edge node and the core node.

A scheduler at each core node is used to dynamically allocate timeslots over the various wavelengths to each edge node. Each edge node contains a separate buffer for the traffic

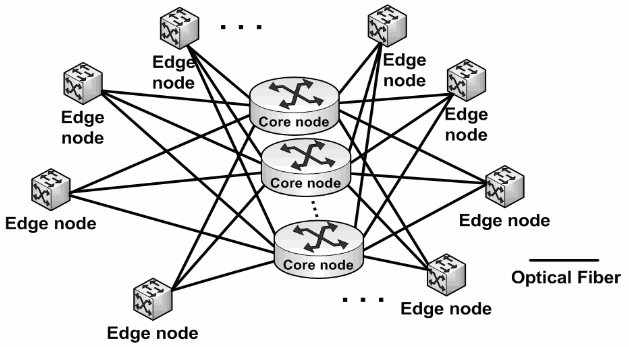

Fig. 2. Agile All-Photonic Network (AAPN) (OSN with several core nodes).

destined to each of the other edge nodes. In these buffers, packets are collected together in fixed-size slots that are then transmitted as single units across the AAPN via optical links. At the destination edge node the slots are partitioned, with reassembly as necessary, into the original packets that are sent to the outside routers. The term "agility" in AAPN describes its ability to deploy bandwidth on demand at fine granularity (timeslot instead of a whole wavelength), which radically increases network efficiency and brings to the user much higher performance at reduced cost.

2) *Ethernet Overlaid-Star Networks with TE Capability*

It is also possible to implement the overlaid-star network by Ethernet switches with TE capability. Here we define a TE capable Ethernet switch including the following aspects:

- PBB-TE (Provider Backbone Bridging – Traffic Engineering) technology enabled, and
- Routing protocol (e.g., OSPF-TE [13]) and signaling protocol (e.g., RSVP-TE [14]) support.

PBB-TE or what was formerly known as Provider Backbone Transport (PBT) [15] is a new technology concept that enables "connection-oriented" end-to-end Ethernet tunnels to be created in order to make the Ethernet an ISP-class transport network. Technically, PBB-TE uses the existing Ethernet technologies of VLAN tagging (IEEE 802.1Q), Q-in-Q (IEEE 802.1ad) and MAC-in-MAC (IEEE 802.1ah), but disables flooding/broadcasting (MAC learning) and the spanning tree protocol (STP). The packets are forwarded based on VLAN ID (VID) and destination MAC address. The PBB-TE tunnels are set up either manually by a management plane or dynamically through a full or partial implementation of the GMPLS control plane [16]. PBB-TE is thus intended to be used in connection-oriented network applications. PBB-TE is now undergoing ratification in the standards bodies.

Replacing the edge nodes and core nodes in Fig. 2 with TE capable Ethernet switches, we get an overlaid star topology: core Ethernet switches surrounded by edge Ethernet switches with 10Gbit/s Ethernet over optical fibers. There is E-O-E conversion at the core switches, in opposition to the AAPN. Core and edge Ethernet switches use RSVP-TE to set up on-demand PBB-TE tunnels (edge-core-edge).

The Ethernet overlaid-star network is much more scalable than the traditional Ethernet-LAN-based IXes architecture (Section II.A.). When the number of core switches exceeds two in the traditional architecture, people have to build complex switch clusters at the core. This would lead to both severe scalability issues and complex internal routing among the switches in the cluster. But this will not happen in the overlaid-star architecture, since the core Ethernet switches are fully independent (no direct physical connection among them) and all the intelligence is distributed among the edge nodes.

3) *PetaWeb*

The PetaWeb architecture, proposed by Nortel Networks [12], is another example of overlaid-star networks. PetaWeb is designed to scale to a capacity of several Petabits per second, as well as to thousands of edge nodes with a global geographic coverage so as to be a candidate for future Internet infrastructure. It is based on the use of a variety of adaptive switching cores and universal edge switches. PetaWeb can be thought as a generalized version of AAPN, since its core nodes may operate in different modes including wavelength switching, TDM switching, or burst switching. The edge nodes must be adapted to interact with various core nodes to support both the connectionless and connection-oriented services.

Generally speaking, an OSN can be viewed as a distributed switch with potentially large geographical coverage. It contains three key ingredients: (1) switching core: rapidly reconfigurable switching of the core, (2) intelligent edge: control and routing functionality concentrated at the edge nodes that surround the switching core, (3) Overlaid star topology for reliability and increased bandwidth.

B. *OSN based Internet Exchange: OSN-IX*

Given OSN as a distributed switch, we find it is very suitable to be deployed as an Internet Exchange. See Fig. 3 for our implementation of OSN-IX. Particularly, OSN-IX has the following advantages compared to other IX architectures:

1) *Flexible and Distributed Access for ISPs*

As shown in Fig. 3, due to the simple and geographically distributed topology of OSN, OSN-IX can provide access (through its edge nodes) to customer ISPs at exactly their local locations. Furthermore, an ISP can have several access points to the OSN-IX, one per OSN-IX's edge node. In this way, the inter-ISP traffic can be distributed widely and balanced (avoid the "bottleneck") within the ISP's network, which also increases the reliability of inter-ISP peering.

2) *Near-Unlimited Capacity with Good Scalability*

We derive Equation (1) to calculate the switching capacity or throughput of an OSN-IX, C_{OSN-IX}, as the follows:

$$C_{OSN-IX} = B \times w \times p \times c\text{, where} \tag{1}$$

- B is the bandwidth of a wavelength over an OSN internal fiber. Typically, B is 10Gbit/s.
- w is the number of wavelengths per OSN internal fiber.
- p is the number of switch ports (fibers) per OSN core node.
- c is the number of core nodes in the OSN.

Let w be 20, p be 64 (maximal value [10]), c be 5, C_{OSN-IX} of such an OSN-IX is 64Terabit/s. Actually, OSN-IX can scale gracefully and continuously in capacity (almost no upper bound according to Equation (1)) to keep pace with the growth in demand of customer ISPs by increasing the number of core nodes and/or wavelengths per fiber. Furthermore, adding one or more edge nodes in a customer ISP's network can be done independently and has no impact on other ISPs.

3) *Good Resilience*

We consider OSN as a switch only conceptually; it does not mean we can use a powerful single physical switch to replace OSN. This is because adopting a single core switch will concentrate all the risks of an IX into a single point. The more powerful the switch, the higher is the risk. In the OSN-IX architecture, things are just on the contrary: the higher the capacity (more core nodes), the more reliable is the OSN-IX. This is due to the fact that in the OSN-IX architecture, the risk is distributed among several independent (not-so-powerful)

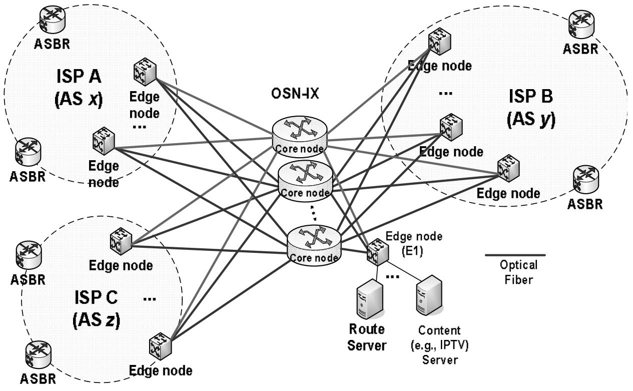

Fig. 3. OSN-IX implementation. There could be content servers (e.g., IPTV server) attached closely to core nodes to distribute contents efficiently).

core nodes, and the overlaid star topology lets them back-up each other. Besides, OSN-IX can increase its capacity by installing extra core nodes in a graceful and natural way, while introducing no complex internal routing, like in other IX architectures.

IV. OSN-IX'S TE FRAMEWORK

One of the most attractive features of OSN-IX is its traffic engineering capability. OSN-IX, for the first time, introduces traffic engineering into the Internet exchange world. Through the TE framework we developed (Fig. 4), OSN-IX can provide inter-ISP (Inter-AS) optimal dynamic routing without maintaining the global TE information and without changing the existing routers, software or hardware, in the ISP networks. We assume that each ISP network is MPLS enabled with TE capability (e.g., deploying OSPF-TE [13] as IGP routing protocol, RSVP-TE [14] as signaling protocol) and has one single AS number. An inter-ISP connection or LSP (Label Switched Path) starts in an ISP network, traverses the IX, and terminates in another ISP network. The OSN-IX TE framework consists of three main components, namely the routing-info, path computation and signaling components.

A. The Routing-info Component

This component is responsible for the discovery and export of the TE topology of OSN-IX. The idea is to export part of the OSN-IX's TE information into each ISP network. First, due to the symmetric architecture of OSN-IX (see Figures 2,3), we use the "bundle" concept to reduce the overhead traffic to the outside. That is, all the links from one edge node to the different core nodes are exported as one single TE link. Second, similarly to above, the overlaid core nodes of the OSN-IX are exported as one single core node, named as "the core" (see Fig. 4). Third, as seen in Fig. 4, from the TE point of view, we expand each ISP AS a little so that there is an overlap between the OSN-IX and each expanded AS. Then the OSN-IX edge nodes located in the overlap, together with their direct TE links to the core and the associated part of the core, belong to both the OSN-IX and an ISP AS. In this scenario, legacy routers (e.g., R1-R5, or R6-R8 in Fig. 4) in an ISP AS see the related OSN-IX edge nodes as normal internal IP/MPLS routers, see the TE links to the core as normal

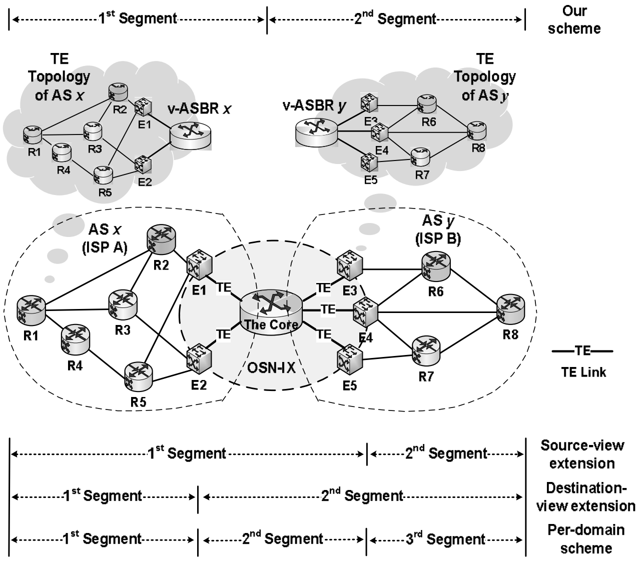

Fig. 4. OSN-IX's traffic engineering framework with example topologies of ISP A and ISP B, and three other inter-domain routing schemes (bottom).

internal links, and see the associated part of the core as an ASBR of its network. In other words, a legacy router sees what it can see in its network about the core as an ASBR, which we call a *v-ASBR* (virtual-ASBR). For the legacy routers in an expanded AS, the exchange and distribution of routing/TE information is just like in any other standard AS with TE capability.

B. The Path Computation Component

In the OSN-IX TE framework, an inter-ISP LSP can be considered consisting of two segments (see the top of Fig. 4): the first one in the head-end (expanded) AS and the second in the tail-end (expanded) AS. The core connects these two segments to form a complete inter-ISP LSP.

The interesting aspect of this architecture is that local routing optimization performed by each of these two sub-LSPs can lead naturally to a globally-optimized inter-ISP LSP (see upper part of Fig. 4). This is due to the particular star topology of the OSN-IX architecture and the load-sharing core nodes that can be exported as one single core to the outside MPLS world. The local routing optimization in the head-end AS can be performed by the source LSR (Label Switch Router), which takes the TE topology and LSP constraints into account. While in the tail-end AS, local routing optimization is done by one of the OSN-IX edge nodes in the AS. Obviously, dynamic inter-ISP routing can be implemented in the OSN-IX TE framework.

C. The Signalling Component

This component is responsible for the establishment of the LSP along the computed path. In Fig. 4, consider the case that a source LSR (e.g., R1) wants to set up a LSP to a destination LSR (e.g., R8). R1 must first compute an optimized path to the v-ASBR of AS *x* through CSPF (constrained shortest path first), and then signal this establishment request to the network. Shown in Fig. 5, R1 starts the signaling process by creating a RSVP Path message with two objects inserted, namely LABEL_REQUEST Object (LRO) to request a label binding

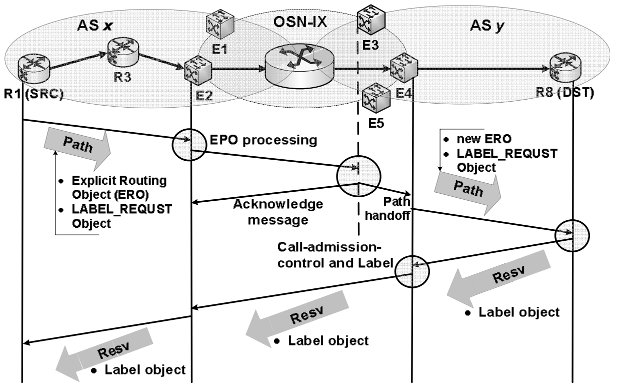

Fig. 5. Inter-ISP LSP signaling process that is fully based on RSVP-TE signaling protocol.

for the path, and EXPLICIT_ROUTE object (ERO) to indicate the computed explicit path (with one sub-object per hop). However, R1 has to use the loose ERO sub-objects for the hops outside AS *x*. In Fig. 5, the ERO specifies the explicit path as R1->R3->E2->v-ASBR *x*->R8, where R8 is a loose ERO sub-object. Then, R1 sends the Path message to the next hop defined in the ERO, which is R3.

R3 receives the Path message and processes it as follows: (1) checks the message format to make sure everything is OK, (2) performs admission control to check the required bandwidth, (3) stores the "path state" from the Path message in its local Path State Block (PSB) to be used by the reverse-routing function, and (4) if successful, deletes the 1st sub-object (itself) in the ERO and forwards the Path message according to the new 1st sub-object (next hop) in the ERO, in our case, E2.

E2, an OSN-IX edge node, receives the Path message from R3 and checks the contained ERO. If E2 finds that the IP address of the 2nd sub-object in the ERO is v-ASBR *x* and the 3rd sub-object (with the loose attribute) is beyond AS *x*, then E2 has the task of resolving the loose sub-object into strict ones. In our case, there is one loose sub-object, R8, which represents the destination of the requested LSP. Although E2 can not find a strict path from v-ASBR *x* to R8 by itself, it knows who can. First, by checking the inter-AS reachability information and internal parameters, E2 finds out which group of edge nodes (also which associated v-ASBR) are located in the same AS as R8. In Fig. 4, these are E3, E4 and E5 (v-ASBR *y*). Second, it selects an edge node among them randomly, e.g., E3. In the third step, E2 removes the first two sub-objects (itself and v-ASBR *x*) from the ERO of the original received Path message, and inserts v-ASBR *y* at the top, then forwards the modified Path message to E3.

When E3 receives the Path message and finds the 1st sub-object in the received ERO is v-ASBR *y*, together with a loose second sub-object, R8, it knows that it should find an explicit path between these two sub-objects. As shown in Fig. 5, E3 is capable to do the resolving work because E3 and R8 reside in the same expanded AS, AS *y*. E3 finds the optimized explicit path: v-ASBR *y*->E4->R8. E3 then replaces the ERO object in the received Path message with a new ERO object that stores the resolved explicit route (E4->R8). Finally, E3 forwards the new modified Path message to E4 as if it were forwarded from E2 by using E2's data (IP address, etc.). We call this process a Path message handoff. At the same time, E3 also sends an acknowledge message (containing the resolved path) to E2 (Fig. 5). From the above handoff process, we can see that only the AS-specific reachability (not TE) information needs to be exchanged among different AS (see Section IV). In our framework, TE information is organized within each AS independently. Edge node E4 receives the Path message and believes it is from E2. Since all the sub-objects in the received ERO are strict, E4 processes this Path message in a standard way, just as R3 did in AS *x*, and then forwards the processed Path message to R8.

When the destination, R8, gets the Path message, it responds to this establishment request by sending a RSVP Resv message. The purpose of this response is to have all routers along the path perform the Call Admission Control (CAC), make the necessary bandwidth reservations and distribute the label binding to the upstream router. The label is distributed using the Label Object in the Resv message. The Resv message makes its way upstream (Fig. 5), hop by hop, and when it reaches the source LSR, R1, the inter-ISP path is setup: R1->R3->E2->v-ASBR *x*->v-ASBR *y*->E4->R8. Now, a globally-optimized inter-ISP LSP is set-up. It can be maintained or torn-down just as any normal intra-AS LSP.

D. Route Service in OSN-IX

OSN-IX adopts an efficient way to build up BGP sessions internally (iBGP sessions within each ISP AS) and externally (eBGP sessions among the ISP ASes) to exchange inter-ISP reachability information. As illustrated in Fig. 6, there is a physical route server connected to an edge node (e.g., E1) that is co-located with core nodes. Each individual ISP AS has one route reflector (RR) [17] instance running in the route server to offer an alternative to the logical full-mesh requirement of iBGP sessions within the ISP ASes. The RR instance acts as a concentrated focal point; multiple BGP routers of an AS can thus peer with it rather than peer with every other router in a full mesh style. This is similar to the idea of the RS (Route Server). For eBGP sessions among peered ASes, the exchange of the reachable IP prefixes offered by each AS is done within the route server all by software. There is no explicit eBGP sessions among the ASes connected by an OSN-IX. In other words, the router server implements both the functions of RR and RS in an integrated way. Compared with the methods in current IXes, this is more efficient and manageable. Furthermore, our method matches well with our OSN-IX TE

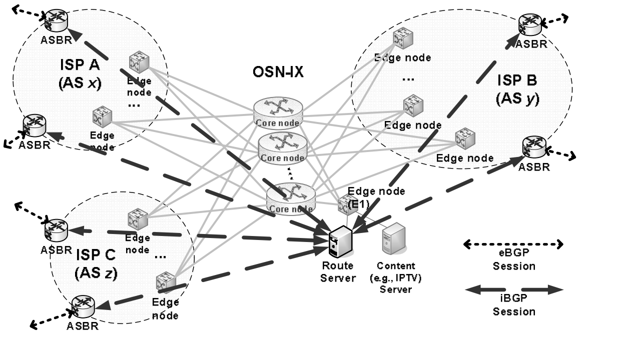

Fig. 6. Route Server in OSN-IX to solve the scalability issue of BGP sessions.

framework: TE info exporting on behalf of the v-ASBR is done by different edge nodes, the exchange of reachable prefixes is done by the route server. Note that for the prefix exchange with other ASes that are not connected to the OSN-IX, the standard BGP procedures must be followed, e.g., eBGP sessions build up as shown in Fig. 6.

E. Related Work on Inter-domain TE

There are several other inter-domain routing schemes proposed in the literature. One class of schemes, e.g., [18], uses a two-step approach to compute an inter-AS route: find out a "loose inter-AS route" first through topology aggregation/ abstraction, then resolve the loose route into a strict path, AS by AS. Actually, this two-step approach would lead to sub-optimal resource utilization and a precise topology aggregation/abstraction always needs very frequent updates which further raise scalability issues. Hence this class of schemes is generally considered unpractical. Another class of schemes [19] claim that they can provide global optimization, but it is at the cost of building up an independent PCE (Path Computation Element) overlay network covering all the ASes, which might somehow decrease the efficiency and increase the cost. In Section V, we will compare our TE framework with the following inter-AS routing schemes:

(1) *Per-domain approach*, which computes the inter-AS path in a AS-by-AS fashion starting from the head-end AS (three segments in total, one by one, see bottom of Fig. 4).

(2) *Source-view-extension [20]*. As shown in the bottom of Fig. 4, this scheme extends the source node's TE visibility so that it can view its own AS and the whole OSN-IX in order to compute the first segment of an inter-AS path. Then the second segment is computed by an ingress gateway(edge) node of the tail-end AS.

(3) *Destination-view-extension [21]*. Still in the bottom of Fig. 4, this scheme defines the two segments in the opposite way to the above approach. The source node can only view its own AS to compute the first segment. The gateway (edge) nodes in the tail-end AS can view its own AS and the whole OSN-IX to compute the second segment.

(4) *Global Knowledge (ideal case)*, in which each inter-AS routing decision is made on the basis of global knowledge of real-time TE information. We use this case as a benchmark.

V. SIMULATION RESULTS

Simulation experiments were conducted on a 27-node ladder-like network, which is the extended version of the topology adopted in [21]. As seen in Fig. 7, two ISP ASes are interconnected through an OSN-IX (AAPN as an OSN). We suppose that the call requests arrive at the network following a Poisson process, and the call holding time is exponentially distributed. We further assume that all the inter-AS source-destination node pairs have the same traffic load, and all the intra-AS node pairs. We assume that 60% of the overall network traffic as inter-AS traffic. We call the links within each ISP AS the normal links and assign to all the same capacity. We use a time-slot (e.g., 100Mbps) as the basic capacity unit on each normal link and the OSN-IX fiber links, and the bandwidth requirement of a call connection is a single time-slot. Least-cost routing is adopted for path selection, where the cost of a path is defined as the sum of the costs of all the links along the path, and the link cost is defined as the inverse of the residual bandwidth of the link. A call is accepted only when there exists one path (single path routing) or two disjoint paths (diverse path routing) with enough available bandwidth. We take the overall (intra- and inter-AS) call blocking probability as our performance metric. All the simulation results have narrow 95% confidence intervals guaranteed by long simulation time.

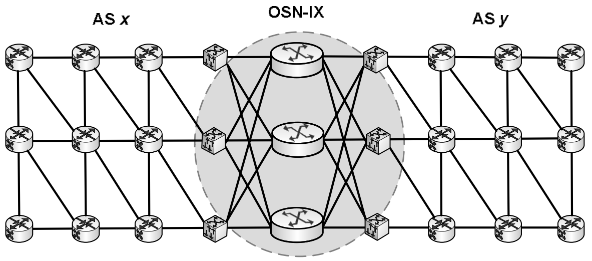

Fig. 7. Network topology used for simulation (27 nodes and 120 directional links in total) (we adopt AAPN as OSN).

A. Single-Path Routing

In the first experiment (Fig. 8a), we fix the bandwidth of the OSN-IX optical links (100 time-slots) while varying the bandwidth of the normal links in the two ISP ASes. This is for the purpose of exploring the effective range of each compared TE scheme. We increase the capacities of the normal links gradually (Fig. 8a) to simulate the phenomenon that the blocking of inter-AS calls is mainly due to lack of capacity in the OSN-IX. Similarly, by decreasing the capacities of normal links, we simulate the phenomenon of inter-AS blocking caused by the lack of capacity in the head-end and/or tail-end ASes (Note that many situations could lead to capacity lack in the real world, e.g., dynamic change of intra/inter-AS traffic, link failures, or not well-engineered network capacity, etc.). Since we use the ideal case as the benchmark for performance comparison, we further adjust the traffic amount so that the blocking probability of the ideal case is kept at around 1% for each given normal link capacity. That's why the blocking curve of the ideal case is a flat line in Fig. 8a.

The per-domain scheme performs worst among the compared schemes. In Fig. 8a, we notice the blocking curve of the per-domain scheme is a near-flat line with the highest values of blocking probability among all the schemes. This is due to its path computation mechanism: domain by domain, sequentially. Referring to the bottom of Fig. 4, each of the three segments of an inter-AS path is computed only on the basis of its own domain's TE information. The starting node of the second or third segment is thus determined by the previous segment "blindly". If there was a "bottleneck" in the OSN-IX or tail-end AS, this scheme could not avoid it.

Our scheme performs best (except for the ideal case) among the compared schemes. As seen in Fig. 8a, the blocking curve of our scheme is flat and very close to the curve of the ideal case in the full value range of normal link capacity. This also shows the robustness (wide effective range) property of our scheme to the change of network environment, e.g., dynamic traffic change, link failure, etc. The small performance difference to the ideal curve is due to the use of approximated (e.g., abstracted/ aggregated) TE information in the OSN-IX.

As illustrated clearly in Fig. 8a, the source-view-extension and destination-view-extension schemes have opposite behaviors /effective ranges under varying normal link capacities. When the normal links have the major contribution (lower capacities) to the inter-AS call blocking, the destination-view-extension scheme performs better than the source-view-extension scheme. Referring to the bottom of Fig. 4, this is because the computation of the first segment of an inter-AS path in the source-view-extension scheme does not consider any TE information of the destination domains. For a blindly-given ingress border node of the tail-end AS, it is not easy, sometimes impossible, to work around the bottleneck links in the tail-end AS. While for the destination-view-extension scheme, although the first segment is computed without any information of the tail-end AS, the ingress border node of the tail-end AS can be chosen to a large extent freely. This is because the second segment in the destination-view-extension scheme includes the OSN-IX and the OSN-IX can connect any egress border node of the head-end AS to any ingress border node of the tail-end AS if the capacity permits. For the same reason, in an extreme case where almost all the inter-AS blocking is due to the head/tail-end AS (the very left end of Fig. 8a), we observe that (1) the blocking curves of the destination-view-extension scheme, the ideal case, and our scheme merge; (2) the blocking curves of the source-view-extension scheme and per-domain scheme merge.

When increasing the capacities of the normal links (right-hand-side of Fig. 8a), the OSN-IX links become the major contributors to the inter-AS call blocking. Then the source-view-extension scheme starts to work better than the destination-view-extension scheme. Similar as above, this is due to the fact that the computation of the first segment of an inter-AS path in the source-view-extension scheme considers the TE information of the OSN-IX so that the "bottleneck" in the OSN-IX can be avoided, while the destination-view-extension scheme can not do that. Again, in another extreme scenario where the inter-AS call blocking is fully due to the OSN-IX (the very right end of Fig. 8a), the blocking curves of the ideal case and the source-view-extension merge. While the curve of our scheme is still close to the ideal one since our scheme uses approximated TE information of the OSN-IX to compute the inter-AS path.

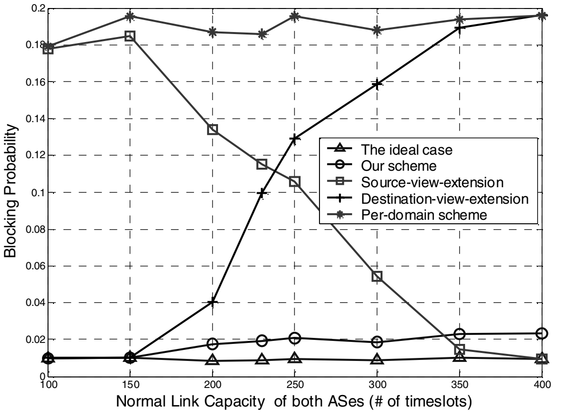

Fig.8a: Overall network blocking probability of single path routing under varying normal link capacities.

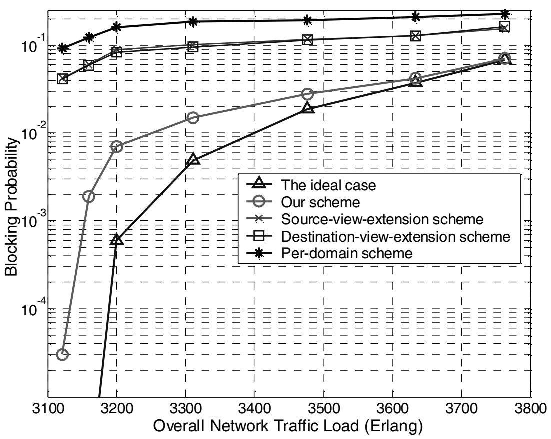

Fig. 8b: Overall network blocking probability of single path routing under various traffic loads with normal link capacity in both domains as 240 timeslots

In our second experiment, we study the blocking performance of the compared schemes under varying traffic loads (Fig. 8b). For this experiment, we wanted to choose a capacity for the normal links to be suitable for both the source- and destination-view-extension schemes. Hence we chose the capacity which corresponds to the point where the blocking curves of the two schemes have a cross-point in Fig. 8a (namely 240 time-slots).

As seen in Fig. 8b, when the traffic grows, the blocking probability increases in all the compared schemes. The source- and destination-view-extension schemes have very close blocking curves along the various traffic loads since we use the crosspoint value of the normal link capacity. This coincides with Fig. 8a. Our proposed TE scheme still works very well as the traffic loads change and remains close to the ideal case. The per-domain scheme performs worst in Fig. 8b, which is expected and mainly due to its path computation mechanism.

B. Diverse Routing

The purpose of diverse routing is load-sharing or end-to-end protection. We define the path diversity as link disjointness in the head-end/tail-end ISP domains, and (edge-, core-) node disjointness in OSN-IX. Each call requires two diverse paths with the same bandwidth requirements (one timeslot for each), and the call is accepted only when both the two diverse paths are available. We still adopt least-cost routing in which the path cost is the sum of the costs of the two diverse paths of a call. The link cost is the same as before. The simulation results of diverse routing are presented in Fig. 9, which is very similar to Fig. 8a. We noticed that the performance of our scheme is very close to the ideal case. It shows that the OSN-IX's TE

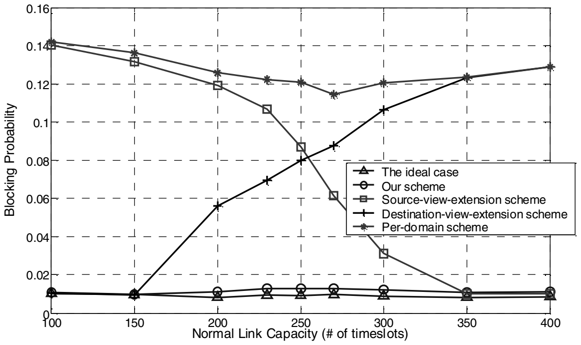

Fig.9: Overall network blocking probability of *diverse* path routing under varying normal link capacities.

framework works very well not only for single-path routing, but also for diverse-path routing.

Regarding establishing protection paths for data flows that require high reliability, we have proposed a protection approach using shared segment protection in OSPF multi-area network scenario [22]. The shared segment protection has the benefits of fast recovery (segmented) and high resource efficiency (sharing) compared to path and link protection. It is interesting to note that this approach can also be adapted to the OSN-IX case and fully take advantages of OSN-IX's TE framework. This would be our future work.

We now analyze the associated complexity of the compared schemes. We first define the following notation:

- *D*: number of ISP ASes.
- *E*: number of edge nodes in each ISP AS (assume each ISP AS has the same number of edge nodes).

In order to compute inter-AS paths (single path or diverse paths), the amount of additional TE information that a normal router in an ISP AS has to maintain either in the source-view-extension scheme or in the destination-view-extension scheme is of size $O(D\times c\times E)$ (where c is the number of core nodes in OSN-IX as defined before); while it is $O(E)$ in our TE scheme, which is much smaller and independent of the number of ASes and core nodes in the OSN-IX. This shows that our scheme has good scalability.

VI. Conclusion

We propose a novel Internet Exchange (IX) architecture, namely OSN-IX, which deploys an overlaid-star network as an IX. Compared to other IX architectures, e.g., Ethernet-LAN-based IX, MPLS IX, Photonic IX, etc., OSN-IX has good properties of scalability and resilience. In addition, OSN-IX has another important and outstanding feature: MPLS TE capability. For the first time, OSN-IX introduces inter-ISP traffic engineering into the IX world. We developed a TE framework for OSN-IX that aims to implement inter-ISP (inter-AS) MPLS traffic engineering while keeping the independence and confidentiality of each ISP network's TE information (only inter-ISP reachability information needs to be exchanged among ISPs). OSN-IX is thus not a TE burden as other IX architectures. In addition, OSN-IX's TE framework has three distinguishing characteristics:

- It provides optimized dynamic inter-ISP routing.
- There will be no change, hardware or software, on existing traditional IP/MPLS routers to implement our framework.
- Simulation results show that our TE framework not only outperforms several existing inter-AS TE schemes and performs very close to the ideal case both in single-path and diverse-path routing, but also has a much wider effective range just as the ideal case .

Indeed, OSN-IX can be considered a highly attractive architectural candidate to the next generation of Internet exchanges.

Acknowledgment

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and industrial and government partners, through the Agile All Photonic Networks (AAPN) Research Network.

References

[1] Amsterdam Internet Exchange, Holland, http://www.ams-ix.net

[2] Japan Internet Exchange, Japan, http://www.jpix.ad.jp

[3] Switch and Data, United States, http://www.switchanddata.com

[4] European Internet Exchange Association, http://www.euro-ix.net

[5] Mario Morelli et al., "An IPv6 Internet Exchange Model. Lessons from Euro6IX project", Proceedings of the 2005 Symposium on Applications and the Internet Workshops (SAINT-W'05).

[6] Nakagawa, I.; Esaki, H.; Nagami, K., "A design of a next generation IX using MPLS technology", Proceedings of the 2002 Symposium on Applications and the Internet (SAINT 2002), Page(s):238 – 245.

[7] I. Shake, et al., Tsukishima, and W. Imajuku "Experiments on optical link capacity adjustment for photonic IX", 31st European Conference on Optical Communications (ECOC2005), Tu.3.4.3.

[8] RFC 1863: A BGP/IDRP Route Server alternative to a full mesh routing.

[9] Slobodanka Tomic, Admela Jukan, "GMPLS-Based Exchange Points: Architecture and Functionality", Workshop on High Performance Switching and Routing, June, 2003, page(s): 245- 249.

[10] L.G. Mason, A. Vinokurov, N. Zhao and D. Plant, "Topological design and dimensioning of Agile All- Photonic Networks", Computer Networks 50 (2006) pp. 268–287.

[11] Gregor v. Bochmann, "Design of an agile all-photonic network", Asia-Pacific Optical Communications (APON2007), Wuhan, China, November, 2007.

[12] R. Vickers and M. Beshai, "PetaWeb architecture," 9th Int Telecom. Netw. Planning Symp., Toronto, Canada, 2000.

[13] RFC 3630: Traffic Engineering Extensions to OSPF Version 2.

[14] RFC 3209 : RSVP-TE: Extensions to RSVP for LSP Tunnels.

[15] Nortel PBT (Provider Backbone Transport) White Paper, http://www.nortel.com/solutions/collateral/nn115500.pdf.

[16] Don Fedyk, Lou Berger, and Loa Andersson, "GMPLS Ethernet Label Switching Architecture and Framework", Internet draft, draft-gmpls-ethernet-arch-00.txt, November, 2007.

[17] RFC 1966: BGP-4 Route Reflectors.

[18] Aihua Guo, et al., "Interdomain traffic engineering in ASON/GMPLS controlled multilayer optical networks", Journal of Optical Networking, vol.6., No.6, June, 2007.

[19] RFC 4655: A Path Computation Element (PCE)-Based Architecture.

[20] Mach Chen, Renhai Zhang, "OSPF-TE Extensions in Support inter-AS MPLS/GMPLS Traffic Engineering", draft-ietf-ccamp-ospf-interas-te-extension-02.txt, November, 2007.

[21] T. Miyamura, T. Kurimoto, M. Aoki and A. Misawa, "An Inter-area SRLG-disjoint Routing Algorithm for Multi-segment Protection in GMPLS Networks," Proc. ICBN 2004, 2004.

[22] Peng He, Gregor v. Bochmann, "Inter-Area Shared Segment Protection of MPLS Flows Over Agile All-Photonic Star Networks", IEEE Globecom, November, 2007.