Data Center Networks and Network Architecture

Hiroshi Esaki
Graduate School of Science and Technology, The University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
Email: hiroshi@wide.ad.jp

ABSTRACT

This paper discusses and proposes the architectural framework, which is for data center networks. The data center networks require new technical challenges, and it would be good opportunity to change the functions, which are not need in current and future networks. Based on the observation and consideration on data center networks, this paper proposes; (i) Broadcast-free layer 2 network (i.e., emulation of broadcast at the end-node), (ii) Full-mesh point-to-point pipes, and (iii) IRIDES (Invitation Routing aDvertisement for path Engineering System).

Keywords: Data Center, Network Architecture, broadcast, routing, switching

1. INTRODUCTION

The Future Internet, which strategically uses the Data Center and cloud computing platform as the core component of smarter city or town, shall contribute the improvement of efficiency and deliver wide spectrum of innovations, regarding all the activity developed and deployed on the Earth.

The growth since the middle of 1980s is with amazing ratio, which has been indicated in more than exponential growth. The growth is not only regarding the number of nodes and networks, is but also regarding the spectrum of user and service categories. In order to achieve the sustainable growth and innovations, the Internet system has been deployed based on the End-to-End architecture paradigm, using the TCP/IP protocol suites. This discipline can be also expressed as “stupid network and intelligent end-nodes”. This means;

(1) Network is transparent
(2) End-node is service consumer, as well as service provider, i.e., P2P (Peer-to-Peer) system rather than client-server system.

For the design, implementation and operation of the Internet, the routing control system of the IP packet is the fundamental and essential technical element. In order to come up with the growth of the Internet system, the routing control system has been always improved and has had some important mutations, in the past.

The static routing had been firstly applied to. With the growth of Internet system, the dynamic routing, that selects the optimized IP packets' forwarding path according to the network status, has been introduced. Even now, we widely use static routing and default routing. However, especially in the core networks, the dynamic routing is generally applied to, in order to improve the quality of connectivity and to achieve robust IP packet forwarding capability. RIP[1], OSPF[2] and IS-IS[3] are used as the IGP(Interior Gateway Protocol), and BGP[4] is used as the EGP(Exterior Gateway Protocol). Each routing protocol applies it's own unique route calculation algorithm and mechanism. However, all routing protocols are based on the same and common architectural paradigm, in order to achieve better scalability. It is the management of routing table based "only" on the destination IP address, and the IP packets are forwarded at every router based on their destination addresses in the receipt IP packets. The routing protocol establishes a global sink-tree, whose root is the destination node, for every destination nodes. By the destination based routing principle, the routing table management complexity in every router can be reduced to order of n, from order of n-square. Here, n is the number of nodes or networks in a given routing domain.

The goal of this paper is;
(1) to grasp the abstraction of existing routing control mechanisms,
(2) to realize the technical requirements against the routing control architecture,
(3) to propose the architecture paradigm to satisfy these requirements.,
while focusing on “Data Center”.

In this paper, the authors propose; (i) Broadcast-free layer 2 network (i.e., emulation of broadcast at the end-node), (ii) Full-mesh point-to-point pipes, and (iii) IRIDES (Invitation Routing aDvertisement for path Engineering System). Here, IRIDES uses the advertisement of "inviting" routing information associated with the destination(s) to be subject for path control. The packet flows, defined by the inviting routing information, are invited to the traffic engineering space, via the nearest invitation point, using the anycast paradigm [5]. In the traffic engineering space, the network operator can determine which traffic engineering technology should be applied to.

In section 2, we describe the issues and requirements of existing Internet routing system, while clarifying the abstraction of existing Internet routing system and data center networks. In section 3, we propose the architecture for data center, satisfying the issues and challenges discussed in section 2. Finally, section 4 gives the conclusion of this paper.

2. ISSUES OF EXISTING ROUTING SYSTEM

2.1 Legacy IP Routing System

In this subsection, we discuss two features of existing Internet routing system, and the requirements for the proposed routing system framework in this paper.

2.1.1 Destination-Based Routing

RIP and OSPF/ISIS is the major routing protocol applied as IGP (Interior Gateway Protocol), and BGP is as EGP (Exterior Gateway Protocol), in the Internet operation. RIP uses a distance vector algorithm, OSPF/ISIS uses a link state algorithm and BGP uses a path vector algorithm. For large scale intra-domain routing, OSPF or ISIS is applied to. And, for inter-domain routing, BGP is applied to. Each routing protocol applies it's own unique route calculation algorithm and mechanism. However, all routing protocols are based on the same and common architectural paradigm, in order to achieve better scalability. It is the management of routing table based "only" on the destination IP address, and the IP packets are forwarded at every router based on their destination IP addresses contained in the receipt IP packets. Every router maintains it's routing table, whose search key is destination IP address and output is the next hop node. In the destination-based routing system, in order to create the routing table, the routing protocol establishes a global sink-tree, whose root is the destination node, for every destination nodes. By the destination based routing principle, the routing table management complexity in every router can be reduced to order of n, from order of n-square. Here, n is the number of nodes or networks in a given routing domain. BGP is of default-free inter-domain routing protocol, creating a global sink-tree by exchanging the full routes with the peering nodes. OSPF and ISIS is of intra-domain routing protocol, creating a sing-tree within it's routing domain. The sink-tree does not have any loop in it, i.e., open tree topology. Since there is no loop, the routing protocol can provide a loop-free routing function.

And the important feature of routing system defined in IP layer is of point-to-point, i.e., does not have broadcasting nor multicasting. The operation only by the point-to-point is one of key technical features so as to achieve and deliver the scalability of IP-based networking, compared to the Ethernet-based networking.

2.1.2 Aggregation of Routing Information

Historically, each organization or campus has their own address space, called as a portable IP address space. This portable IP address spaces were advertised to the Internet. According to the growth of the Internet, it was getting difficult to have and to advertise this individual portable IP addresses from each organization. This is simply because of the increase of number of portable IP address spaces, since every router must maintain all the portable IP address spaces to be advertised. In order to come up with this scaling issue, the Internet system has introduced an "address aggregation". By the introduction of address aggregation, we could decrease the increase of routing entries in routing table at every router. In general, an organization or a campus obtains it's IP address space from the ISP(Internet Service Provider) or from the NIR (National Information Registry), in order to effectively aggregate the IP address spaces. As the results, the IP address structure has been delivered to a hierarchical structure. And, the strong correlation among IP address structure and IP packet forwarding routes has been established.

2.1.3 Challenges on IP Routing System

Internet routing system must come up with the complex technical requirements and with complex network structure and topology. As discussed in the previous subsections, the existing routing system has the following two technical restrictions.

(i) Within the routing domain, the sink-tree, whose root is the destination object (node or network), is shared among all the source objects (nodes and networks).
(ii) In order to achieve effective address aggregation, the hierarchical IP address allocation has been applied to.

Due to these two technical restrictions, the Internet system has experienced some technical issues. These technical challenges, derived from these two technical restrictions, can be categorized into the following two points.

(1) Provision of un-shared routing paths
For traffic engineering, the provision of IP packet forwarding path that is not shared with the other IP packet flows, for some particular IP packet flow is one of key functional components. This is because, for existing routing system, all the IP packet flows, whose destination network/node is the same, must share the same IP packet forwarding path. When we can provide an alternative path or multiple paths to the destination node/network, it will contribute to the improvement of QoS provided by ISP. Here, QoS includes not only naive communication quality, such as latency or available bandwidth, but also the robustness of service. When we apply providing multiple paths to some (campus) network, it corresponds to multi-homing.

(2) Provision of Plug-and-Play
Due to the hierarchical address allocation and address aggregation, conjunction with provider based IP address allocation, the network is required re-numbering and re-configuration of many system parameters, when it changes it's location physically or logically. One of typical example of logical location change is changing the ISP to be connected. Since the user network is allocated it's IP address space from the ISP, the IP address space must be changed, when the user network changes the ISP. This is simply because that the user network does not have a portable IP address space.

In the practical networking, we have many manual configurations to run the system. At the system initiation phase, it is not plug-and-play. However, when the user network does not need any re-configuration of system parameters, it could be said as "plug-and-play" against the change of location where the user network is connected to the Internet. Of course, it is beneficial when the system does not need any re-configuration of system parameter against the location change, for practical network operation.

2.2 De-Fact Layer 2, i.e., Ethernet

For local sub-network, layer 2, called as data-link, has been defined and use in the computer networks. In order to accommodate and interconnect sub-networks, network layer, which is layer 3 such as IP (Internet Protocol) has been defined. MAC (Media Access Control) algorithm and protocol are defined so as to deliver the layer 2 datagram to the target interface within the sub-network. When the destination node is not in the same sub-network, the datagram is forwarded to the neighbor sub-networks via gateway (i.e., router) using the layer 3 address of destination node.

We have wide variety of layer 2 technologies. Some are broadcast based architecture (e.g., bus or ring), some are NBMA(Non-Broadcast Multiple Access) based and some are point-to-point.

As a de-facto layer 2 technology, the Ethernet is widely used in the recent computer networks. Here, even the original Ethernet is broadcast-based bus architecture, we observe that Ethernet protocol is applied in the point-to-point data-link.

2.3 Features and Requirements for the Networks in Data Center

2.3.1 Services provided by data center

The data center provides the following services for customer.

(1) Co-location Service
Customer brings in their own equipments into the data center. Data center provides layer 1, i.e., physical cable, connectivity among those equipments, based on customer's requests.

(2) Hosting Service
Data center provides either physical or virtual computer to the customer, with external connectivity.

(3) Cloud Service
Basically, three services are provided to the customer; (i) IaaS (Infrastructure as a Service), (ii) PaaS (Platform as a Service), and SaaS (Service as a Service).

The customer may ask to data center the network topology, which they want to have, with certain qualities. The service quality is defined as SLA, Service Level Agreement. SLA may include available bandwidth or system availability, such as MTBF(Mean Time Between Failure) or SAIDI(System Average Interruption Duration Index).

2.3.2 Emulation of Ethernet

For the data center provider, the following two services frequently require the emulation of Ethernet segment (i.e., LAN emulation).

(1) hosting multiple servers, i.e., server cluster segment
(2) IaaS for the migration of all or part of equipments, which were installed in the customer premises, including the hybrid IaaS that is consisted by Off-the-Premises and On-the-Premises.

In order to satisfy the above requests by the customers, most of data center provides Virtual Ethernet segment, using possible technological solutions, such as VLAN (IEEE802.1Q) or VXLAN [6].

2.3.3 Applicability against VM migration

The other request for the data center networks is migration of VMs (virtual machines) in the data center and among data centers. Even when the virtual machine is changed it's connecting location, the customer wants to preserve the allocated IP addresses to the virtual machines. Even when the virtual machine changes the connecting location while preserving the IP address, the data center must preserve the connectivity and communication quality, such as latency and packet loss.

2.3.4 High Performance Computing

The data center network is expected to provide the platform, which can provide high performance computing (HPC) capability. In the most of recent super computer and HPC system, the processors are interconnected via point-to-point links, e.g., InfiniBand rather than shared media, e.g., Ethernet. This means that the HPC system may want to have full-mesh dedicated point-to-point low layer pipes among all processors in the HPC cluster. HPC system may want to have full-mesh point-to-point pipes, because of (a) low and stable latency among any processors and (b) large and stable bandwidth among any processors. This is also true for data center networks, since many customers desire the same features to data center network.

3. PROPOSED DATA CENTER NETWORK FRAMEWORK

3.1 Broadcast-Free Network

This paper proposes the inherit of successful Internet principle, which is discipline of "end-to-end", and the disciple of "stupid network and intelligent end-node". In the past, a lot of challenges, e.g., LAN Emulation in ATM networks, to introduce intelligence into the network had been failed. The intelligence may include broadcast and multicast service. The TCP/IP, which is the most successful global system, does not provide broadcast and multicast service.

Therefore, this paper proposes that end-node may emulate broadcast and multicast service provided by legacy layer 2 segments, e.g., Ethernet and by some applications in the data center networks.

Sometimes, broadcast is used for the auto-configuration so as to learn the routing ID to deliver the datagram to the destination node. We can achieve this requirement using Rendezvous framework with anycast framework. Typical implementation in very large global scale network environment is the root-DNS system. We have logical 13 root servers over the global Internet. A logical root server is consist by multiple servers, which are geographically distributed over the global Internet, while having and advertizing the same IP address into the global IP routing system, i.e., anycasting. The address of anycasting for root servers are provided by off-line fashion {and be able to update on-line fashion}, as a hint-file, that contains IP addresses of root servers. Every DNS server refers the hint-file, so as to contact root servers, without broadcast function. In other words, every DNS server automatically accesses the nearest root server, without broadcast platform.

Applying this framework to data center network, the data center networks may not need to provide broadcast nor multicast, but may only provide unicast (point-to-point). This framework achieves simple or simpler implementation for data center switched and routers.

3.2 Full-mesh Point-to-Pint Pipes

Based on the discussion in subsection 2.3.4, this paper also proposes that we may not need packet switching function by switches in the data center networks.

IP packets can be recognized by end-node, and switch may not need to recognize IP packets. Then, the switches provide only point-to-point transparent pipes among end-nodes.

This is related with the discussion in subsection 3.1. No intelligence in the network, and intelligence in the end-node. When the end-node can take care of services, which expect the broadcast or multicast function of layer 2, while applying the framework discussed in subsection 3.1, we do not need to ask the broadcast nor multicast function. Actually, the de-facto layer 2 technology, Ethernet can work well with point-to-point link. As we recognized, the provision of multicast and broadcast in very high speed network is not easy, is but difficult. This is because the switches must securely manage spanning tree, so as to avoid multicast or broadcast storm.

3.3 Accommodation of legacy system and VM migration

In order to accommodate legacy systems and VM migration, this subsection proposes the IRIDES framework[7], which uses the advertisement of invitation routing information from some border routers called as IIR (Ingress IRIDES Router). IRIDES represents Invitation Routing Information aDvertisement for path Engineering, which has a lot of similarity with LISP (Location and Identifier Separation Protocol) [8] and the proposed architecture in [9]. IRIDES framework can solve two technical issues discussed in the previous section, i.e., provision of alternate path against sharing the same destination-based sink-tree and provision of plug-and-play against system reconfiguration. IIRs interconnect two internetworks. One is legacy layer 3 networks, and the other is some other layer 3 networks, called as Traffic Engineering Plane. Traffic engineering plane would have some new traffic engineering capability, such as GMPLS. Also, traffic engineering plane can be either overlaying network above legacy layer 3 networks (e.g., IPv6 tunneling network or virtual network defined by MPLS's LDPs) or physically independent networks (e.g., lambda network using GMPLS). Multiple IIRs interconnect these two networks.

Figure 1 shows the IRIDES abstracted configuration in legacy internet system. This is the case where we install two IIR routers. IIRs advertise the invitation routing information to the legacy internet system. And, IIRs can advertise exactly the same routing information (i.e., anycast routing), to achieve load balancing and system redundancy against the outage of IIR. IP packets are automatically forwarded to the closer (closest) IIR, by the nature of routing protocol capability.

Figure 2 shows the IP packet transmission in IRIDES system, where the IP packets are transmitted from different three source nodes (SH1, SH2 and SH3) in the legacy internet system to the destination node (DH1) in the traffic engineering plane. We have three IIRs in this example, i.e., IIR1, IIR2 and IIR3. The IP packets, that should be transmitted to the destination node (DH1) in traffic engineering plane, are forwarded to the best (i.e., nearest) IIR for each source node (SH1, SH2 and SH3). IIR defines the appropriate identifier, which is used for the receipt IP packet forwarding in the traffic engineering plane, using the destination IP address (and other information to define that particular IP packet flow) in the receipt IP packet. The identifier used for routing in the traffic engineering plane depends on the technology used in the traffic engineering plane and can be, for example, MPLS label, VLAN tag, outer IP address for IP-in-IP tunneling path, lambda label for GMPLS or IPv6 address.

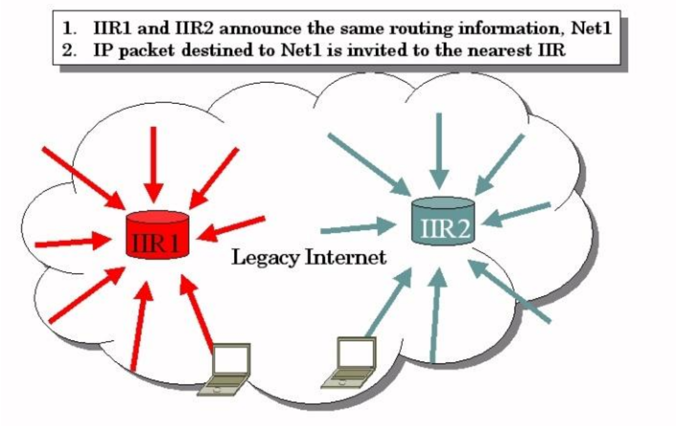
1. IIR1 and IIR2 announce the same routing information, Net1
2. IP packet destined to Net1 is invited to the nearest IIR
IIR1
IIR2
Legacy Internet

Figure 1. IRIDES Configuration in Legacy Internet Space

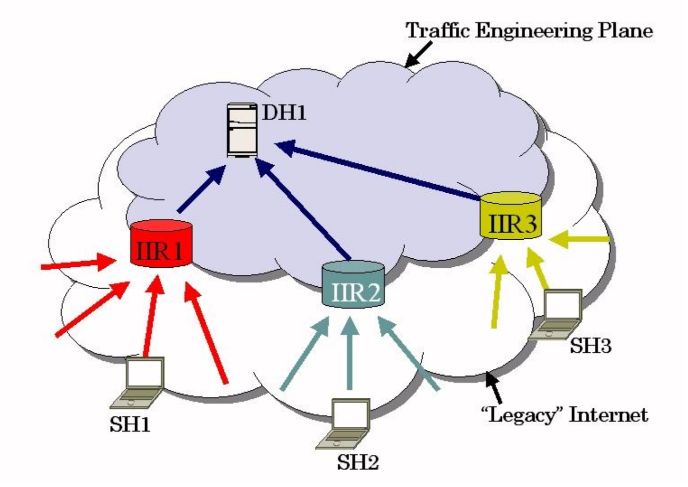

Figure 2. IP Packet Transmission in IRIDES System

Figure 3 shows the case where the IREDES system provides multiple paths to the destination node/network, i.e., multi-homing. In this example, two paths are defined from each IIR to the destination. Therefore, in this case, four of redundant paths are defined between source node and destination node. Here, an important operational point would be this configuration and operation is independent from source node, which is in the legacy internet system. This means that we do not need any modification for the nodes in the legacy internet system. This is a beneficial technical feature for reducing the deployment difficulty of IRIDES architecture framework.

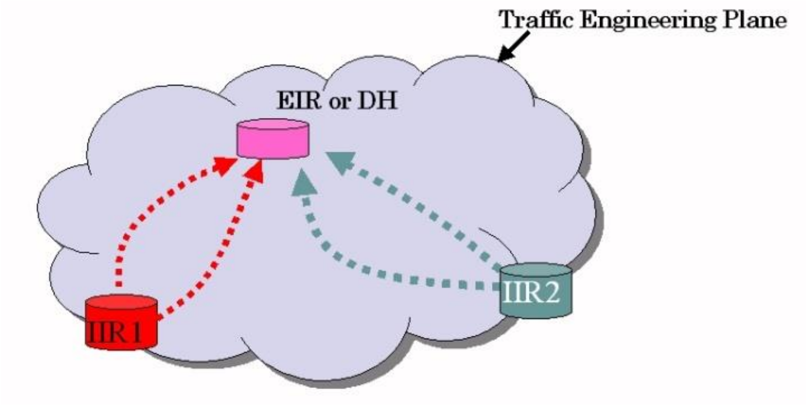

Figure 3. Multi-Paths Data Transmission in Traffic Engineering Plane

3.4 Technical Features of IRIDES Architecture

3.4.1. Distributed Installation of IRRs with Anycast Routing

(1) Scalability
Using the anycast routing, we can achieve load balancing among IIRs, which can be (globally) distributed, via the distribution of forwarding paths to the destination.

(2) Route Optimization
The IP packet can be forwarded to the nearest (i.e., optimal) IIR, from the viewpoint of routing matrices.

(3) Possibility of Traffic Control at Ingress Point
IIR would be able to act as the control point for many purposes. These would be for DDOS defending, congestion control or policy control.

3.4.2 Co-existence with Legacy Internet

(1) When we define the edge/access network in existing internet infrastructure as the legacy internet system of IRIDES system, and define the core network as the traffic engineering plane of IRIDES system, we can accommodate any existing networks, without any re-configuration for every node in the existing internet.

(2) The IP packet transmission technology applied in the traffic engineering plane is not pre-defined single technology, but can be any transmission technology.

(3) As shown in figures 2 and 3, the IRIDES system re-defines a sink-tree from the destination node to any leaves in the legacy internet system. This sink-tree can have less number of shared links, compared with the corresponding sink-tree defined only by legacy internet system.

(4) The path calculation in the traffic engineering plane can be independent from that in the legacy internet system. Therefore, it is easy to co-exist with the existing internet system and the traffic engineering plane.

(5) We can define the network topology in the traffic engineering plane, which is independent from that of legacy internet system. Also, even over the heterogeneous network infrastructure, we can define homogeneous virtual overlaying network using the existing internet system. In the practical network, even when we design the network with only a single vendor, the functions provided by the equipments are frequently not identical.

3.4.3. Provision of Portable IP Address Space for Customer Networks

In the IRIDES system, we separates the identifier into (1) IP address for end-to-end communication, and (2) label/address for IP packet forwarding in the traffic engineering plane. By this separation of identifier, every end-object (node or network) in the legacy internet system can obtain a portable IP address(es). By the provision of portable IP address space to the end-object (i.e., customer networks in the legacy internet), the end-object can be re-configuration-free network, against both the physical and logical location change.

In the legacy internet system, the IP address has two semantics. One is the identifier for the interface of node, and the other is the identifier for the routing of IP packet. IRIDES architecture separates these semantics and define the different identifier for each semantics, at least within the traffic engineering plane.

4. CONCLUSION

This paper discusses and proposes the architectural framework, which is for data center networks. The data center networks require new technical challenges and it would be good opportunity to change the functions, which are not need in current and future networks. Based on the observation and consideration on data center networks, this paper proposes; (i) Broadcast-free layer 2 network (i.e., emulation of broadcast at the end-node), (ii) Full-mesh point-to-point pipes, and (iii) IRIDES (Invitation Routing aDvertisement for path Engineering System). The author may think the network architecture framework discussed in this paper could apply to the other networks, in general.

REFERENCES

[1] G.Malkin, "RIP Version 2", IETF RFC2453, Nov.1998.

[2] J.Moy,"OSPF Version 2", IETF RFC2328, April 1998.

[3] D.Oran, "OSI IS-IS Inter-Domain Routing Protocol", IETF RFC1142, Feb.1990.

[4] Y.Rehkter, T.Li, "A Border Gateway Protocol 4 (BGP-4)", IETF RFC1771, March 1995.

[5] C.Partridge, T.Mendez, W.Milliken, "Host Anycasting Service", IETF RFC1546, Nov.1993.

[6] M. Mahalingam, and others, " VXLAN: A Framework for Overlaying Virtualized Layer Networks over Layer 3 Networks, IETF, draft-mahalingam-dutt-dcops-vxlan-06.txt, November 2013

[7] Hiroshi Esaki, "10G Wideband Global Testbed", Part 35th, WIDE project annual report 2004, March 2005.

[8] D.Farinacci, V.Fuller, D.Mayer, D.Lewis, "The Locator/ID Separation Protocol (LISP)", IETF RFC6830, January 2013.

[9] Ryo Nakamura, Yuji Sekiya and Hiroshi Esaki, "Implementation and Operation of User Defined Network on IaaS Clouds using Layer 3 Overlay", 3rd International Conference on Cloud Cimputing and Service Science (CLOSER 2013), Poster Session, Aachen, Germany, May 2013.