Telecommun Syst (2014) 56:33–47
DOI 10.1007/s11235-013-9817-8

Highly available network design and resource management of SINET4

Shigeo Urushidani · Michihiro Aoki · Kensuke Fukuda · Shunji Abe · Motonori Nakamura · Michihiro Koibuchi · Yusheng Ji · Shigeki Yamada

Published online: 17 August 2013

Abstract The Japanese academic backbone network has been providing a variety of multilayer network services to support a wide range of research and education activities for more than 700 universities and research institutions. The new version, called SINET4, was launched in 2011 in order to enhance the service availability and the network bandwidth as well as to expand the service menu. Its enhanced service availability was unexpectedly verified by the disastrous March 11 Great East Japan Earthquake, when the network managed not to stop service operation even after the earthquake. This paper describes the design and implementation of SINET4 in terms of multiple service provision, network resource control and management, and high reliability from physical level to network management level. The impacts of the huge earthquake are also reported.

Keywords Multilayer services · Converged network · High availability · Network resource management

Preliminary results for the contents of this paper have been presented at the international workshop on reliable network design and modeling (RNDM) 2011.

S. Urushidani (✉) · M. Aoki · K. Fukuda · S. Abe · M. Nakamura · M. Koibuchi · Y. Ji · S. Yamada
National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
e-mail: urushi@nii.ac.jp

M. Aoki
e-mail: aoki-m@nii.ac.jp

K. Fukuda
e-mail: kensuke@nii.ac.jp

S. Abe
e-mail: abe@nii.ac.jp

M. Nakamura
e-mail: motonori@nii.ac.jp

M. Koibuchi
e-mail: koibuchi@nii.ac.jp

Y. Ji
e-mail: kei@nii.ac.jp

S. Yamada
e-mail: shigeki@nii.ac.jp

1 Introduction

Many projects in cutting-edge research areas share huge research devices or use special devices and exchange huge amounts of data through academic backbone networks [5, 10, 21]. For example, high-energy physics, nuclear fusion science, and supercomputing projects share huge research devices financed by government [1, 14, 15, 18], astronomical projects link radio telescopes together [13], seismological projects gather data from lots of seismic sensors [12], and high-realistic communication projects transmit high-resolution videos with advanced communication tools [6, 24]. The academic backbone networks must therefore be very high speed enough to transfer such huge amounts of data. These networks must also provide useful tools for collaborative research, such as virtual private networks (VPNs) in multi-layers. In addition, the networks sometimes need to temporarily provide huge network resources for some big-science experiments.

In order to meet these requirements, the Japanese academic backbone network, called the Science Information Network (SINET), has been enhancing its networking capabilities. Advanced multi-layer VPN services and on-demand services, which started provided by the third version, SINET3 [25, 26], have encouraged more collaboration and increased data-intensive applications. As the use of these services has increased nationwide, the network was

Table 1 Network service menu in SINET4

Service menu		Status	Note
Access interface	E/FE/GE (T)	✓	
	GE (LX)	✓	
	10GE (LR)	✓	
Layer-3 service	Commercial Internet access	✓	Via JPIX, JPNAP, Level3, etc.
	IPv6	✓	Basic: dual stack, option: native, tunnel
	IPv4 full-route information	✓	
	IPv4/IPv6 multicast	✓	
	IPv4/IPv6 multicast (QoS)	✓	
	Application-based QoS	✓	
	L3VPN	✓	
	L3VPN (QoS)	✓	
	Multicast in L3VPN	Planned	
Layer-2 service	L2VPN/VPLS	✓	Most popular service for collaboration
	L2VPN/VPLS (QoS)	✓	
	L2VPN/VPLS on demand	Planned	
Layer-1 service	L1 on demand	✓	More than 1,000 dynamic paths so far
Other service	Private cloud support	✓	Rapidly growing service

requested to be higher-speed and more reliable. The network also had to expand its coverage areas to reduce regional disparities in accessibility. The National Institute of Informatics (NII), which has operated SINET, therefore decided to launch a new network, called SINET4, to address these issues in 2011. During the migration from SINET3 to SINET4, designed high-availability functions were unexpectedly verified by the disastrous March 11 Great East Japan Earthquake, when the network managed not to stop service operation even after the huge earthquake.

This paper describes the design and implementation of SINET4 and the impacts of the huge earthquake and is organized as follows. Section 2 describes the required specifications and design concept. Section 3 describes the detailed network design and network resource control and management to shape the design concept. Section 4 details our high-availability functions from physical level to network management level. Section 5 shows the impacts of the Great East Japan Earthquake on SINET4 and its users. Section 6 presents our conclusion.

2 Requirements and network structure

2.1 Requirements on new network

SINET4 needs to provide a variety of network services to support research and education activities for more than 700 universities and research institutions (Table 1). SINET started its operating as an Internet backbone network, and SINET4 continues to provide a commercial Internet access service via major domestic commercial Internet exchange points and contracted global ISPs and provides IPv4 full-route information for BGP users and network researchers. The network supports IPv6 transfer capabilities in the styles of native, dual-stack, and IPv4 tunneling, along with multicast functions. QoS control services can be supported for mission critical applications. The network also provides a variety of VPN services. SINET started to provide L3VPNs in 2003, L2VPNs in April 2007, and virtual private LAN services (VPLSs) in December 2007. Layer-2 based VPN services have been encouraging collaborative research, and the number of VPNs has been growing steadily. For example, a seismic research project uses broadcast capabilities of VPLS to distribute observed data from each collector to all other collectors. SINET4 plans to provide L2VPN/VPLS on-demand (L2OD) services with VLAN-based QoS control, through which users themselves can create experimental environment freely. The network also provides bandwidth-on-demand services in layer 1, called layer-1 on-demand (L1OD) services. SINET started to provide L1OD services in June 2008 and has established and torn down over 1,000 layer-1 paths so far. SINET4 expands the service area and increases the available bandwidth. For example, an e-VLBI project [13] started to use this service with a 2.4-Gbps bandwidth over a STM-16 interface per antenna in 2008 and now uses an 8.4-Gbps bandwidth over a 10 GE interface per antenna.

Because SINET has been used as a lifeline network for many academic organizations, its reliability and stability must be more reinforced than ever before. Especially, installation environments for SINET nodes that accommodate

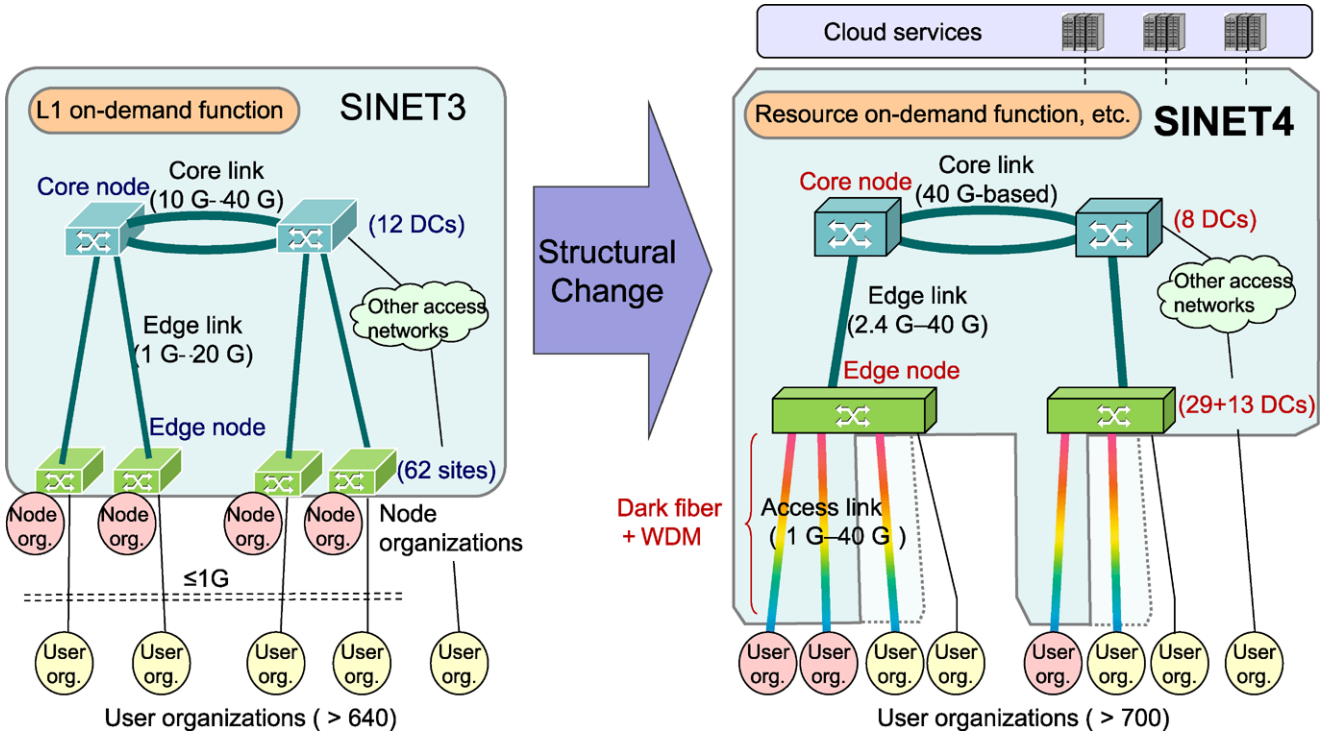

Fig. 1 Structural differences between SINET3 and SINET4

access links of academic institutions must have been reconsidered. In previous SINETs, many nodes were located at selected academic organizations, which led to some problems for the node stability and operability. For example, each organization must receive the annual legal inspection of the power facilities, for which NII needed to send a power-generator truck in order not to stop SINET services during the power outage. Maintenance personnel also suffered from different security policies of each organization when they need an emergency response for a node or link failure. In addition, there was the possibility that this situation cause great perplexity in case of natural disasters such as earthquakes. NII therefore decided to place every SINET node at selected data centers that resolve these problems.

SINET also needed to expand its coverage areas and reduce regional disparities in accessibility in response to regional universities' requests. In previous SINETs, 13 of 47 prefectures did not have SINET nodes, and regional universities in these prefectures had to connect their access links to the SINET nodes placed in other prefectures. This raised their link costs or reduced their link bandwidths. NII therefore decided to place SINET nodes in all prefectures after optimizing the network topology.

In addition, the advent of cloud computing services brings the need for us to support user organizations to build the "private cloud" infrastructure, where their virtual servers and storages are placed at commercial data centers and are used in closed environments, as well as to utilize upper layer applications such as e-mail. NII therefore decided to allow the cloud computing service providers who support academic organizations' activities to connect directly to SINET4 as SINET's service providing organizations.

In summary, the new network was requested to achieve higher speed cost-effectively; to support multilayer network services nationwide; to be more reliable and stable than before; to expand the coverage areas; and to facilitate the movement to private cloud infrastructure.

2.2 Structural design of SINET4

Structural features of SINET4 as well as those of SINET3 are shown in Fig. 1. Both networks have edge nodes which accommodate access links of SINET's user organizations and core nodes which exchange the traffic among the edge nodes. SINET3 had 62 edge nodes located at selected user organizations, called node organizations, and 12 core nodes co-located at telecom carriers' buildings. Edge links between the edge and core nodes had a speed of 1 to 20 Gbps, and core links between the core nodes had a speed of 10 to 40 Gbps. The core links between Tokyo, Nagoya, and Osaka were the Japan's first 40 Gbps (STM256) links [25].

In SINET4, all edge and core nodes are placed at selected data centers, and each core node includes edge node functions. By taking into account distances between nodes as well as between nodes and user organizations, we consolidated SINET3's edge nodes into 29 and the core nodes

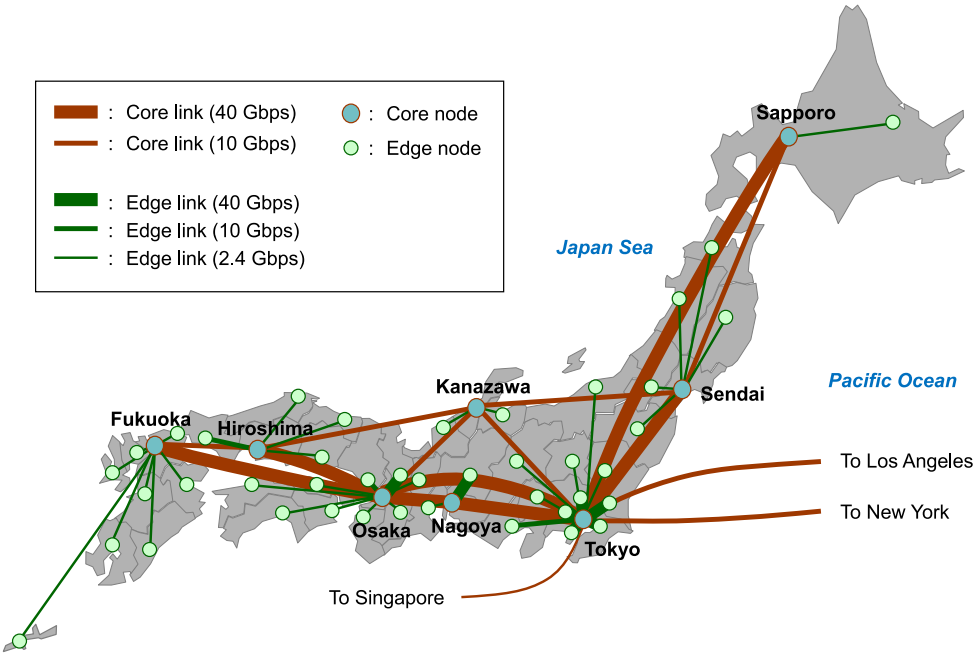

Fig. 2 Node location and network topology

Table 2 Criteria for selecting data centers

Item	Criteria
Neutrality	Neutrality to telecommunications carriers' lines
	Neutrality to system vendors' equipment
Secure power supply	No planned power outages
	Emergency power supply for at least ten hours without refueling
Natural disaster resistance	Resistance to earthquakes equivalent to the Great Hanshin Awaji Earthquake in 1995
	5 meters or higher above sea level for seaboard cities
Security	Secure accessibility 24/7/365
	Admission within 2 hours in emergency situations
Location	Closeness to previous node organizations for WDM-based access links

into 8, and instead enhanced the link bandwidths between nodes. We also decided to add 13 edge nodes in order to resolve regional disparities. We installed four edge nodes in 2011 and nine edge nodes in 2012. Eventually every prefecture have at least one edge/core node, and the network have 50 edge/core nodes in 47 prefectures as of March 2012. SINET4 forms a nationwide 40-Gbps (STM-256) backbone from Sapporo to Fukuoka and has more bandwidth between Tokyo and Osaka (Fig. 2). Core links form six loops to create redundant routes for high service availability. Edge links have a speed of 2.4 Gbps (STM16) to 40 Gbps (STM256) depending on the expected traffic volume. Here, every core/edge link is a dispersed duplexed link.

SINET4 introduced WDM devices for access links between the previous node organizations and the data centers to attain a maximum capacity of 40 Gbps with four 10 gigabit Ethernet (10GE) interfaces. The WDM devices can transmit optical signals up to about 40 km without amplifiers, which we took into account to select the data centers. We also performed joint procurement of WDM-based access links for other user organizations to obtain faster access links at reasonable costs.

SINET4 deployed similar node architecture to that of SINET3 and focuses on expanding resource-on-demand services. SINET4 also supports private cloud computing services by using L2VPNs in collaboration with commercial cloud computing service providers.

2.3 Node installation environment

SINET4 places every node at commercial data centers which meet the following criteria to improve the availability and operability (Table 2). The data centers must be able to accommodate any telecommunications carriers' lines as well as any system vendors' equipment. They must be able to supply electric power to our equipment by emergency power

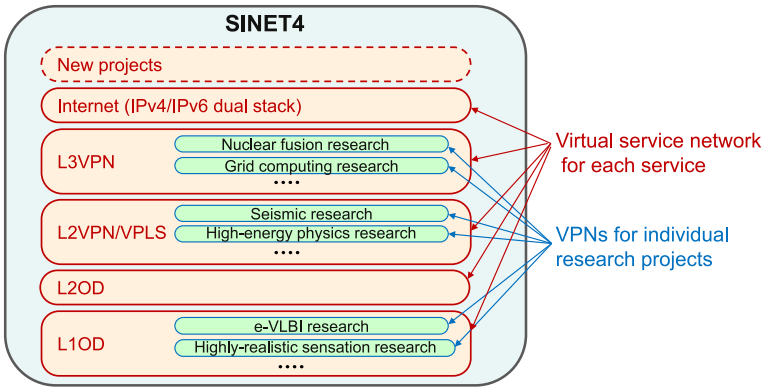

Fig. 3 Service provision using virtual service networks

supplies for at least ten hours without refueling them in case of blackouts. They must be resistant to earthquakes equivalent in intensity to the Great Hanshin Awaji Earthquake in 1995. They must be securely accessible 24/7/365 and give us admittance to our spaces within 2 hours in emergency situations. They were also requested to be moderately close to the previous node organizations so as to be able to use the WDM-based access links without amplifiers.

3 Network design for multi-layer network services

3.1 Virtual separation of network services

To support a variety of network services and enable each network service to grow independently and stably, we introduced virtual service networks, each of which is dedicated to each network service group. Each network service group is formed by network services which need similar networking functions, such as routing, signaling, and forwarding functions. As of March 2012, SINET4 has five virtual service networks for the following service groups: IPv4 and IPv6 transfer (IPv4/IPv6 dual stack); L3VPN; L2VPN and VPLS; L2OD; and L1OD services (Fig. 3). Here, each VPN is created in the corresponding virtual service network.

3.2 Node architecture for multilayer network services

We needed to combine different equipment in order to create the virtual service networks on a single network. Through procurements in 2010, we decided to combine layer-1 switches (NEC's UN5000), layer-2 multiplexers (Alaxala's AX6600), and IP routers (Juniper Networks' MX960) (Fig. 4).

Each edge node is composed of a layer-1 switch and a layer-2 multiplexer and accommodates user access links with Ethernet-family interfaces. Each layer-2 multiplexer receives layer-2/3 service packets, inserts internal VLAN tags corresponding to each user organization or research project into the packets, and sends the tagged packets to the layer-1 switch. The layer-1 switch accommodates the received packets through 10 GE interfaces into layer-1 paths for layer-2/3 services in a SDH-based (STM256/64/16) link by using the generic framing procedure (GFP) [7] and virtual concatenation (VCAT) [9] technologies. Each core node is composed of layer-1 switches and an IP router, and the layer-2/3 service packets over the layer-1 paths are transferred to the IP router via 10 GE interfaces. The IP router distributes the packets to its logical systems corresponding to each virtual service network. Here, each IP router has four logical systems, indicated by "IPv4/IPv6" for IPv4 and IPv6 transfer services, "L3VPN" for L3VPN services, "L2VPN" for L2VPN and VPLS services, and "L2OD" for L2OD services, as shown in Fig. 4. The VLAN tags of each service packet are used for this distribution, and the VLAN tags of IPv4/IPv6 and L3VPN service packets are removed at the logical systems. Except the logical system of IPv4/IPv6, each logical system encapsulates the service packets with multi-protocol label switching (MPLS) tags. Then, each logical system inserts another VLAN tag corresponding to the virtual service network into the service packets and sends them to the layer-1 switch. The layer-1 switch accommodates the received packets through 10 GE interfaces into a layer-1 path for layer-2/3 services in a SDH-based line by GFP and VCAT.

Each L1OD service user is dynamically assigned a layer-1 path between layer-1 switches. We usually obtain the network bandwidth for L1OD services by changing the bandwidths of layer-1 paths for layer-2/3 services. This bandwidth change is done with a VC-4 (about 150 Mbps) granularity without any packet loss by using the link capacity adjustment scheme (LCAS) [8]. We developed an L1OD server for this dynamic layer-1 paths setup/release and bandwidth change. The L1OD server receives user requests, such as destinations, bandwidths, and durations, via simple Web screens, calculates the best routes, controls the layer-1 switches through their operation system via CORBA interface [23], and manages the network resources [26]. Upon receipt of path setup/release orders from the L1OD server,

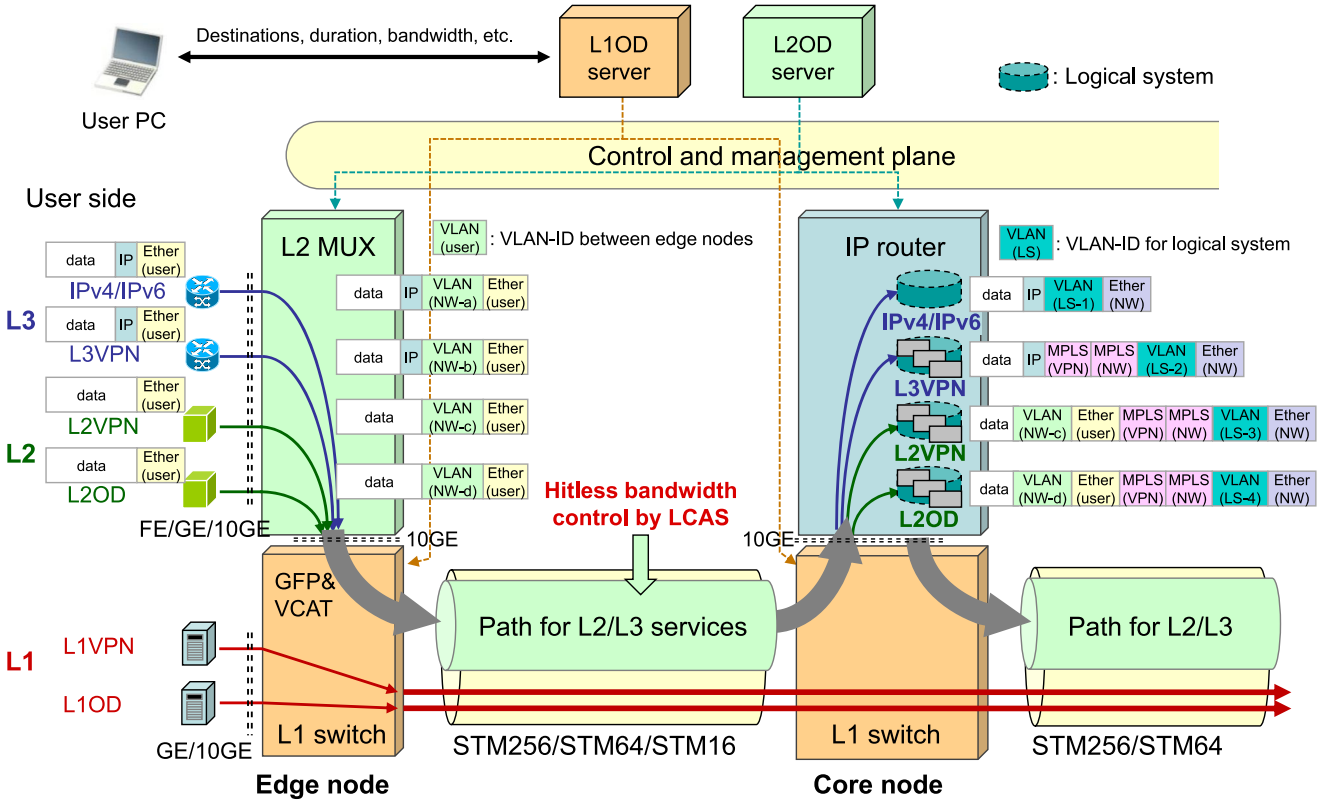

Fig. 4 Node architecture and networking technologies

the layer-1 switches exchange the generalized MPLS (GMPLS) protocols [2, 16] to set up and release these layer-1 paths. The layer-1 switches also change the bandwidths of the layer-1 paths for layer-2/3 services by using LCAS upon receipt of path bandwidth change orders from the L1OD server.

We plan to provide L2OD services by using an L2OD server that receives user requests via simple Web screens similar to those of the L1OD services and controls the layer-2 multiplexers and the IP routers via NETCONF interface [4] to set up and release layer-2 paths with QoS control along assigned routes.

3.3 Resource assignment for multilayer services

We accommodate both layer-1 paths for layer-2/3 services and those for L1OD services into STM256/64/16 links and dynamically change their assigned bandwidths. The following shows how we assign the network resources for these services. For a STM64 (or STM16) link, a layer-1 switch accommodates layer-2/3 service packets received from a 10 GE interface of an IP router or a layer-2 multiplexer into the assigned time slots of the STM link by GFP. As we use the granularity of VC-4, a STM64 (or STM16) link has 64 (or 16) VC-4 time slots numbered from 1 to 64 (or 16). We assign the time slots numbered from 1 to T ($1 \leq T \leq 64$ (or 16)) for layer-2/3 services and those numbered from 64 (or 16) to $T + 1$ in reverse order for L1OD services when needed.

For a STM256 link, a layer-1 switch accommodates layer-2/3 service packets received from each of four 10 GE interfaces of an IP router or a layer-2 multiplexer into each assigned time slot group of the STM256 link. We divide 256 time slots of the STM256 link into four time slot groups (1–64, 65–128, 129–192, and 193–256) and assign these time slots of each group to each 10 GE interface when there are no L1OD services (Fig. 5(a)). In the figure, we express N VC-4s as simply Nv. Next, when we need the bandwidth for L1OD services, we obtain the bandwidth from time slots of four time slot groups evenly, in order of older to younger time slots, in a round-robin fashion. For example when we obtain a full bandwidth of a Gigabit Ethernet interface, which needs 7v, we assign seven time slots: 256–255, 192–191, 128–127, and 64. Note that the VCAT technology allows us to use arbitrary time slots to obtain required bandwidth for a layer-1 path. When we need another 7v, we assign seven time slots: 254–253, 190–189, 126, and 63–62. Figure 5(b) shows the case in which we assign 56 time slots (256–243, 192–179, 128–115, and 64–51) to obtain 8.4 Gbps for e-VLBI. The remaining time slots (1–50, 65–114, 129–178, and 193–242) are assigned to four 10 GE interfaces of the IP router.

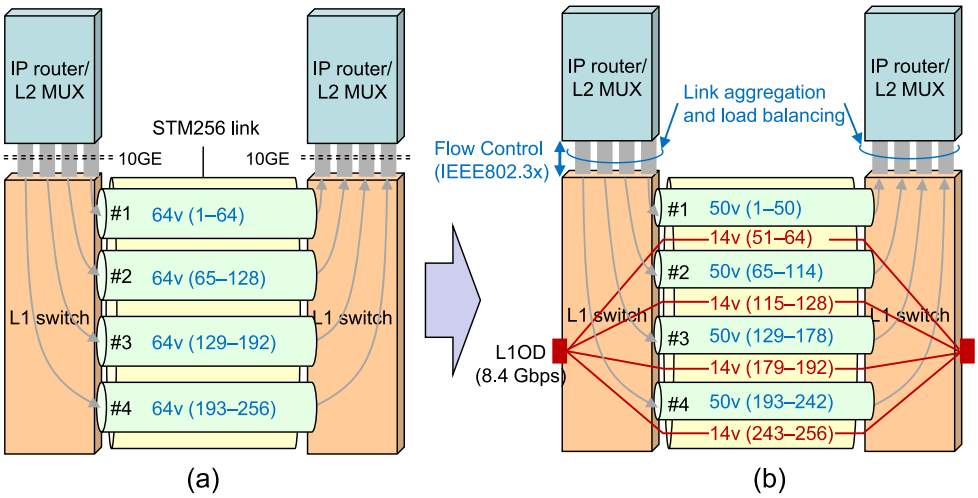

Fig. 5 Time slot assignment for multilayer services

The reason we assign the time slots evenly is as follows. We treat four 10 GE interfaces for layer-2/3 services as one virtual interface by using the link aggregation technique of IP routers or layer-2 multiplexers. In addition, we distribute layer-2/3 service packets to four interfaces as evenly as possible by using the load balancing technique. The assigned bandwidth (the number of time slots) to each interface, therefore, should be as even as possible to maximize the total link utilization. As for effective load balancing, we use different hash keys for different services. For IPv4/IPv6 services, we use the contents of multiple header fields of each IP packet for load balancing, i.e. source IP address, destination IP address, protocol ID, and destination TCP or UDP port number. Source TCP or UDP port number is excluded for load balancing in order to prevent the packet arrival disorder in some applications. For L3VPN services, we use at most three MPLS labels for load balancing, i.e. two MPLS labels for identifying network route and VPN as a default and one more MPLS label for fast reroute in case of trouble as described in Sect. 4.5. As for L2VPN/VPLS and L2OD services, we use destination MAC address, source MAC address, and even the contents of the IP header fields in addition to three MPLS labels.

If an IP router or layer-2 multiplexer detects a failure of one of the aggregated interfaces, it departs the interface and evenly loads the layer-2/3 service traffic to the remaining interfaces in accordance with the load balancing rules described above. If the interfaces of the layer-1 switch have a shortage of buffers due to heavy traffic, the flow control complied with IEEE 802.3x is done over each interface between the layer-1 switch and the IP router or the layer-2 switch in order to avoid packet loss.

3.4 Network resource management

As we provide a variety of services on the single network, we give different service availability between layer-2/3 and L1OD services. The service availability of L1OD services especially depends on the time and destinations, because we assign the network resources to the L1OD services by taking into account the traffic volume of layer 2/3 services which many users use. We manage the resource assignment by setting the available bandwidth of each link for L1OD services. The available bandwidth of each link for L1OD services is defined by a 4-tuple, $\{D, A_1, A_2, A_3\}$. "D" is the available bandwidth for any time without reservation. "A_1" and "A_2" are the available bandwidth from 8:00 to 22:00 and that from 22:00 to 08:00 on weekdays respectively when reservation requests are made by the day before the use. "A_3" is the available bandwidth for any time on weekends when reservation requests are made by the day before the use. Here, these values for each link are set in the L1OD server by our network operators via simple Web screens. We usually assign the network bandwidth for L1OD services on a reservation basis, but some link bandwidths can be used without reservation for some experimental use. Figure 6 shows the available bandwidth of each link as of March 2012. The available bandwidths of the other links not shown in the figure are set to $\{0, 0, 0, 0\}$. For example, $\{0, 56, 56, 56\}$ means that a user has to make a reservation by the day before the use in order to use L1OD services and can reserve the link bandwidth of up to 56v (=8.4 Gbps). The National Institute for Astronomical Observatory of Japan, which leads eVLBI projects and is located near Chofu data center, can use L1OD services only on a reservation basis, and can use a maximum of 112v for daytime on weekdays and 168v for nighttime on weekdays and on weekends on the Chofu–Tokyo link. The L1OD server calculates an appropriate route of each layer-1 path by taking account into the required end-to-end bandwidth and the available bandwidth of each link. If a link fails, L1OD can recalculate the routes for reserved paths by removing the failed link, but we currently do not prepare redundant resource for L1OD services in order to save layer-2/3 services first.

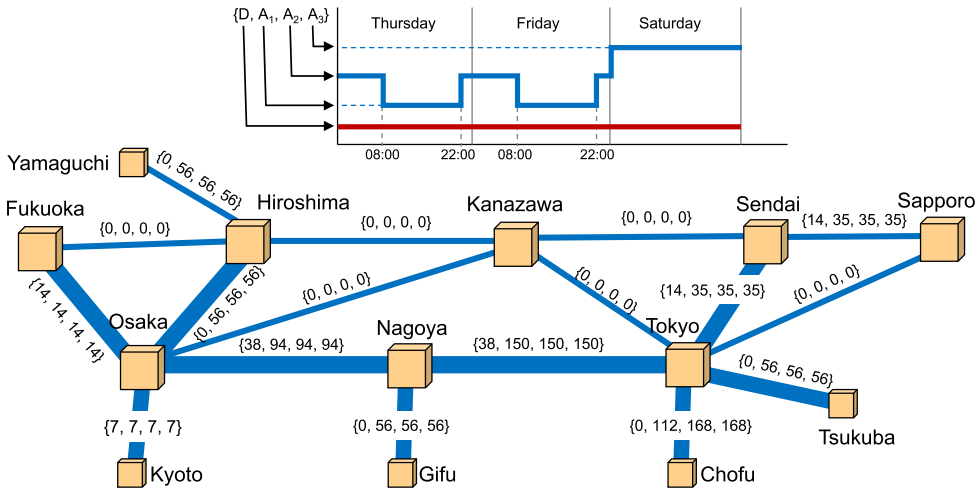

Fig. 6 Available link bandwidth for L1OD services

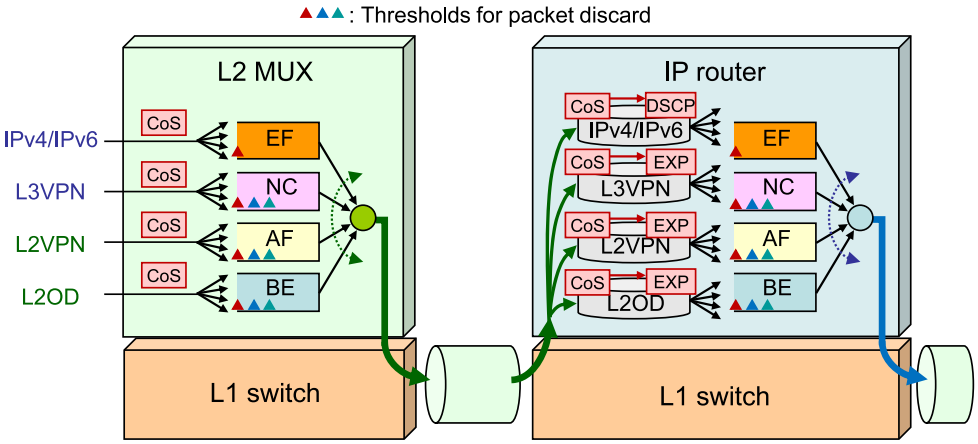

Fig. 7 Prioritized packet forwarding at L2 multiplexer and IP router

As for the layer-2/3 services, SINET4 has quality of service (QoS) control functions in order to transfer control protocol packets and performance-sensitive packets with high priority even if link congestions occur. Each IP router or layer-2 multiplexer has, in order of priority, the following four forwarding queues: expedited forwarding (EF), network control (NC), assured forwarding (AF), and best effort (BE) queues (Fig. 7). We currently use the AF and BE queues for user data packets and the NC queue for control protocol packets, and each queue has two kinds of drop precedence, low and high, each of which starts to drop packets at different thresholds. Here, the EF queue is reserved for mission-critical applications. Control packets for routing and signaling, such as RIP, OSPF, BGP, RSVP-TE, and packets for urgent informing, such as TRAP, are assigned low drop precedence, and packets for network monitoring such as SNMP are assigned high drop precedence because we collect large amounts of traffic information of every interface by SNMP. For QoS control, we mark the QoS identifier of each packet: CoS (or user priority) bits of each Ethernet packet at layer-2 multiplexers, DSCP bits of each IP packet, and EXP bits of each MPLS packet at IP routers. As for packets that are transferred from a layer-2 multiplexer to an IP router, the CoS bits of packets are copied to the DSCP bits for IPv4/IPv6 service packets or to the EXP bits for L3VPN/L2VPN/VPLS/L2OD service packets. The QoS is controlled at each device depending on the value of the corresponding QoS identifier.

3.5 Network management

To manage the backbone network composed of IP routers, layer-2 multiplexers, and layer-1 switches, we have a network management center, and for WDM devices we have an access link management center. For robust network management, we have three ways to reach the devices through monitor routers: data plane (i.e. in an in-band fashion), control and management plane (i.e. in an out-of-band fashion), and integrated services digital network (ISDN) (Fig. 8).

As for the IP routers and layer-2 multiplexers, we primarily monitor them through the data plane because we usually obtain a large volume of management data including traffic information of every interface at one-minute intervals. If we cannot access the devices though the data plane, we

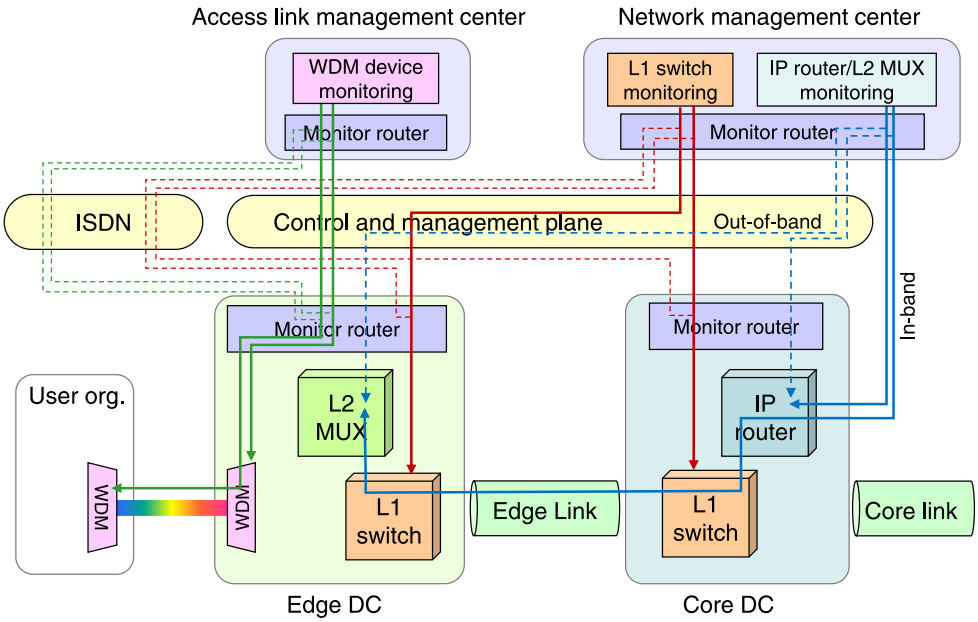

Fig. 8 Access routes to each device for control and management

use the control and management plane, but this is used only for command line interface (CLI) operation due to the limited link bandwidth, which is 5 Mbps for each edge node and 10 Mbps for each core node and is shared with layer-1 switches. The layer-1 switches need the control and management plane in order to exchange GMPLS protocols for obtaining link state information and setting up/releasing layer-1 paths and also to be monitored outside. When we cannot access the layer-1 switches through the control and management plane, we use ISDN links to monitor them.

As for WDM devices, we always use Ethernet operations, administration, and maintenance (Ethernet OAM) frames between WDM devices to monitor the device status as well as operate them. Statistical results and emergency messages are sent from the WDM devices located at each data center to the access link management center through the control and management plane. ISDN links are used when the control and management plane is down.

Table 3 summarizes the primary and backup networks through which the two management centers manage the related devices.

Table 3 Primary and backup routes for control and management

	Data plane	Control and management plane	ISDN
IP router	Primary	Backup (CLI only)	
L2 MUX	Primary	Backup (CLI only)	
L1 switch		Primary	Backup (limited operation)
WDM device		Primary	Backup (limited operation)

4 Network design for high availability

To attain a highly available network, we have to consider the availability from diversified viewpoints [3, 22]. This section describes our entire network design for high availability from seven different standpoints in detail (Fig. 9).

4.1 Node placement at data centers

All of the nodes are placed in the selected data centers in order to attain stable node operation even for strong earthquakes and sudden blackouts, as described in Sect. 2.3. As there are many earthquakes nationwide in Japan, we have prepared for possible strong earthquakes with a seismic intensity of 7. As for blackouts, daily electric power supply in Japan was very stable thanks to electric power companies' efforts until March 11 2011, but the terrible nuclear plant disaster suddenly changed the situation, and people now worry about possible electric power shortages and rolling blackouts. We are therefore glad we decided to move the node location to data centers that have sufficient emergency power supply capabilities.

4.2 Duplexed edge and core links

Every link builds up with a pair of primary and secondary circuits (Fig. 10). We decided each circuit pair in order that the secondary circuit goes through geographically different routes from the primary circuit. For example, for the core link between Tokyo and Sapporo, the primary circuit goes

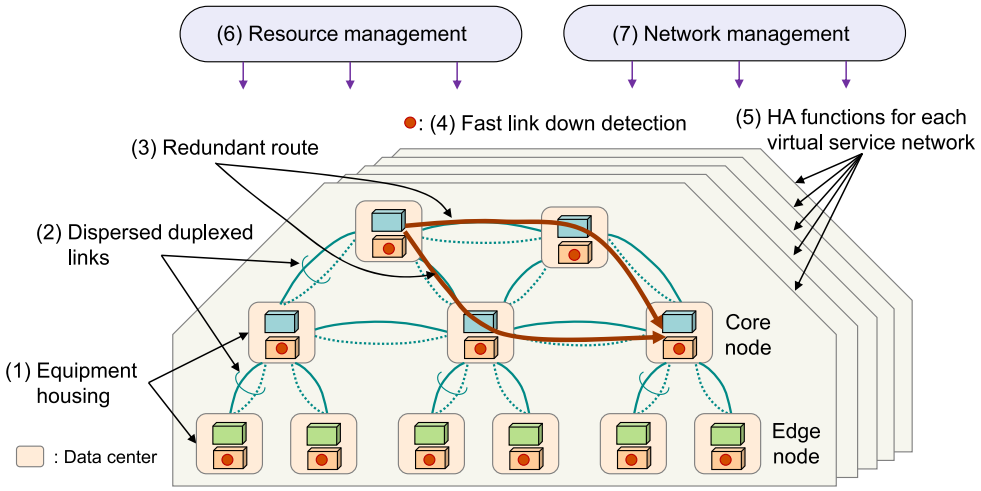

Fig. 9 Network design for high availability

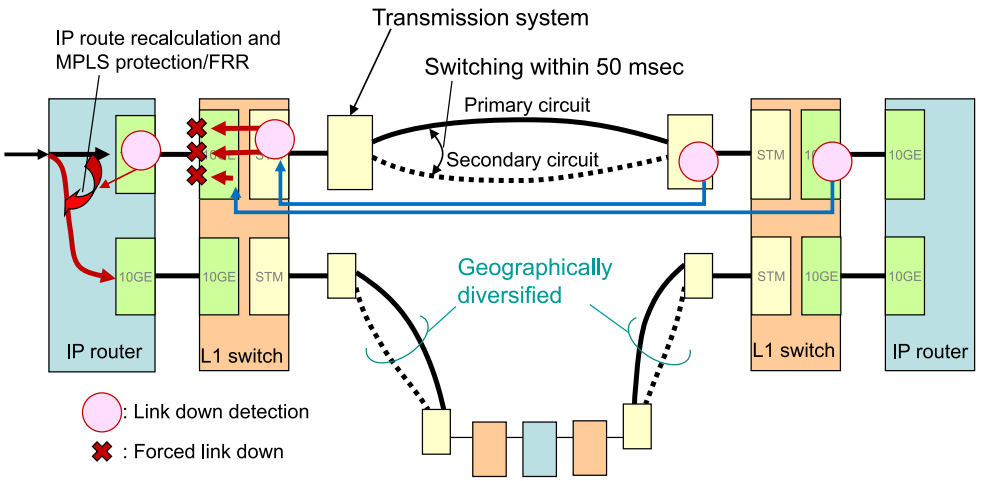

Fig. 10 Duplexed links and fast link down detection

through the Pacific Ocean side, and the secondary circuit goes through the Japan Sea side. We use the shorter delay route as a primary circuit and the other as a secondary circuit. The primary circuit is usually active, and if it fails the secondary circuit automatically becomes active within 50 ms. Only if both the circuits fail does the link come down.

4.3 Redundant routes between core nodes

Core links between core nodes have longer distances than edge links and go through many transit points to interconnect optical fibers and insert amplifiers, so they have larger possibilities of failures than edge links. We therefore built sufficient redundant routes between core nodes, as shown in Fig. 2, to divert the service traffic to different directions. Although we have not prepared abundant bandwidth for diverted traffic in case of failures due to the limited budget, control protocol packets are transferred with high priority even for heavy congestion.

4.4 Link down detection by layer-1 switches

SINET4 provides every service on the top of layer-1 switches that accommodate edge and core links. The layer-1 switches can detect link down including opposite interface down quickly by using link monitoring and link down transfer functions and can quickly inform the IP routers and layer-2 multiplexers of the detected link down by forced link down (Fig. 10). Triggered by forced link down, the IP routers can quickly divert the service packets to other routes by functions described in Sect. 4.5 or detach the failed interface from aggregated interfaces. Unlike with normal failure detection between IP routers, which need about 40 s by OSPF, we can remove the failed link or interface very quickly and reduce the packet loss. The guard time for layer-1 switches to inform IP routers or layer-2 multiplexers of the detected link down is set to 100 ms by taking into account the switching time of 50 ms between duplexed links.

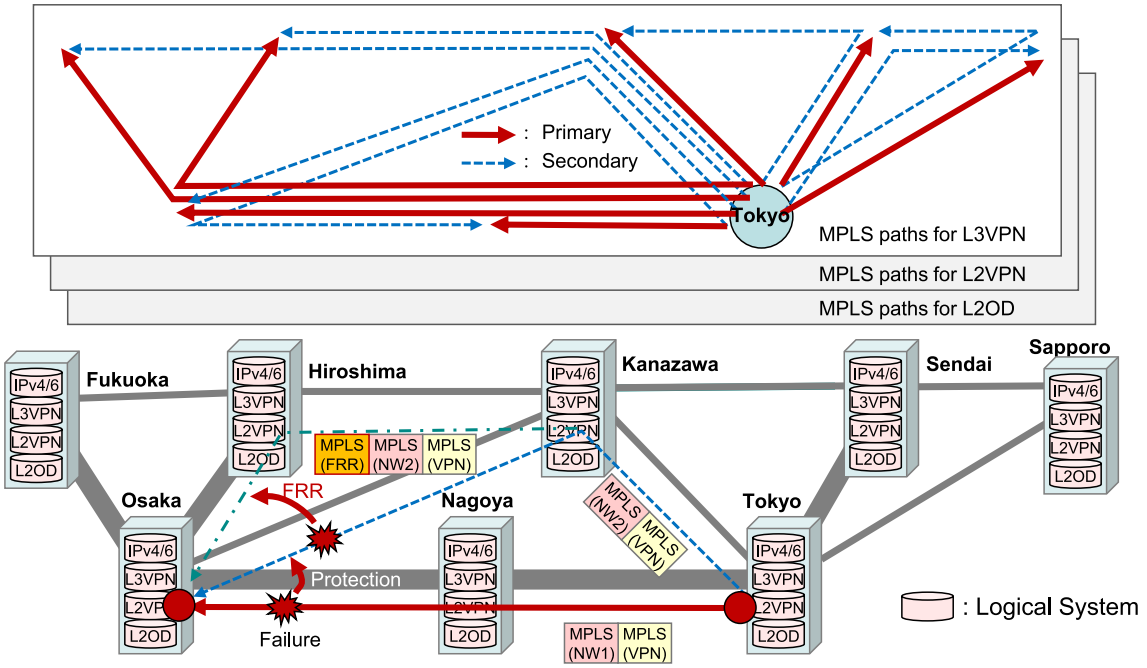

Fig. 11 MPLS-based high-availability functions for VPN services

4.5 High availability functions for multilayer services

High-availability functions are implemented in each virtual service network. The virtual service network for IPv4/IPv6 dual stack services uses the OSPFv2/v3 protocols to decide the shortest routes and quickly recalculates alternative routes by using them in case of failures. The virtual service networks for L3VPN, L2VPN/VPLS, and L2OD services use MPLS technologies to transfer VPN packets and divert the packets to other routes by using both MPLS protection and fast reroute (FRR) techniques [11, 19, 20] in case of failures. We set up disjoint primary and secondary MPLS paths between arbitrary logical systems for stable service recovery, and each primary MPLS path goes though the smallest-delay route in the network (Fig. 11). We also use FRR functions which find alternative routes for quick recovery and divert the packets to the routes by using additional MPLS labels. When a failure occurs on a primary MPLS path, the logical system detecting the failure (Nagoya's in Fig. 11) uses FRR for partial recovery and then informs the ingress logical system (Tokyo's in Fig. 11) of the failure with an RSVP PathErr message which triggers the MPLS protection. If another failure occurs on the secondary MPLS path, the logical system detecting the failure (Kanazawa's in Fig. 11) performs FRR for partial recovery. The virtual service network for L1OD services uses the L1OD server to calculate the best routes and to assign the time slots, recalculates the assigned network resources by the L1OD server if a failure occurs before the layer-1 path setup, and can use GMPLS LSP rerouting functions [17] of layer-1 switches for mission critical applications if a failure occurs during the service.

4.6 Bandwidth management and QoS control

To steadily carry the layer-2/3 service traffic whose volume has a very analogous daily pattern on each link in our network, we manage the availability for L1OD services by setting the available bandwidth of each link in response to the traffic pattern of layer-2/3 services, as described in Sect. 3.4. Even when link congestions occur due to burst traffic or detoured traffic by a link failure, the network can maintain the stable connection state between IP routers by definitely transferring control protocol packets with QoS control functions. For higher availability for layer-1 services, we can obtain the required bandwidth from the multiple routes between core nodes by using VCAT and Edmonds-Karp algorithm [26]. For example, we can obtain the required bandwidth between Tokyo and Osaka nodes from two routes, Tokyo–Nagoya–Osaka and Tokyo–Kanazawa–Osaka routes. In this case, even when a failure occurs on either route we can keep the connectivity with reduced bandwidth by using LCAS. We will use this method for mission-critical applications, although currently we do not set the available bandwidth for this purpose.

4.7 Reliable network management

We have primary and backup routes to control and manage all our devices, i.e. IP routers, layer-2 multiplexers, layer-1

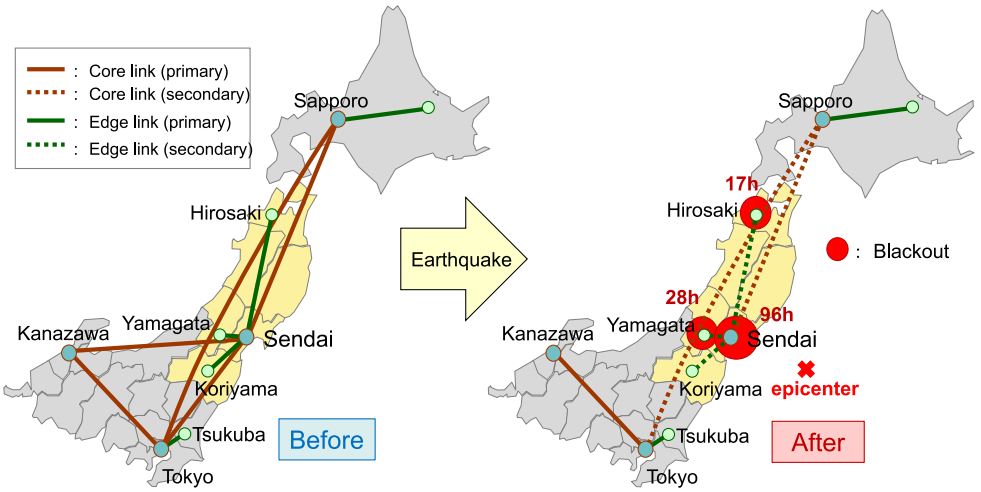

Fig. 12 Impacts of the Great East Japan Earthquake

switches, and WDM devices, for reliable network management. By taking into account different management styles of each device, we prepare three different planes: data plane, control and management plane, and ISDN, as described in Sect. 3.5.

5 Impacts of the great east Japan earthquake

We constructed the initial version of SINET4 by January 2011, started the migration from SINET3 to SINET4 in early February 2011, and moved the access links of user organizations to SINET4 by the end of March 2011. When the earthquake struck the Tohoku area on March 11, we fortunately had almost completed the migration there and managed to continue the service operation. This section reports the impacts of the earthquake from the viewpoints of SINET4 and user organizations.

5.1 Impacts on our backbone

Figure 12 shows the network situation before and after the earthquake in the Tohoku area, where we had the core node at Sendai city and the edge nodes at Koriyama, Yamagata, and Hirosaki cities as of March 2011. Because we placed all nodes in selected data centers, no equipment was damaged by the earthquake despite its seismic intensity of 7 due to the earthquake resistant design of the data centers. After the big shakes, the blackout took place at many places and lasted for a long period, for example about 96, 28.5, and 17 hours around data centers at Sendai, Yamagata, and Hirosaki cities, respectively. The blackout also lasted 12.5-hours around the data center at Chofu city in Tokyo. Equipment in these data centers continued to work though emergency power supply systems, which were refueled until commercial power sources were recovered.

On the other hand, backbone links were severely affected. In Fig. 12, solid lines indicate the primary circuits of the links, and dashed lines indicate the secondary circuits that became active after the earthquake. The two core links between Sendai and Tokyo and between Sendai and Kanazawa went down due to the damage of both the primary and secondary circuits. Nevertheless, the Tohoku and Hokkaido areas could keep the connectivity to other areas through the secondary circuits between Sendai and Sapporo and between Sapporo and Tokyo. All three edge links in the Tohoku area also survived though the secondary circuits. Therefore, none of the areas were isolated. As for two down core links, the secondary circuits were repaired after about 57 hours, but the primary circuits needed more than one month to be repaired. The affected primary circuits of the other links were repaired in a couple of days.

The surviving nodes and links successfully diverted the layer-2/3 service traffic to other routes by OSPFv2/v3, MPLS protection, and FRR functions. Although this diversion increased the delay for the communication from and to the Tohoku area, no packet losses were observed in IP routers. Just in case of an emergency, we set the available bandwidth for L1OD services between Sendai and Sapporo to {0, 0, 0, 0} until the secondary circuit between Sendai and Tokyo was repaired.

5.2 Impacts on user organizations

WDM access links between previous node organizations and data centers were fortunately not affected by the earthquake but the blackout around the organizations made them lose connectivity to SINET4. The disconnected times were 46.5 and 26 hours for Tohoku University at Sendai city and Hirosaki University at Hirosaki city, respectively. Other previous node organizations, such as KEK and Tsukuba University at Tsukuba city and Keio University at Yokohama

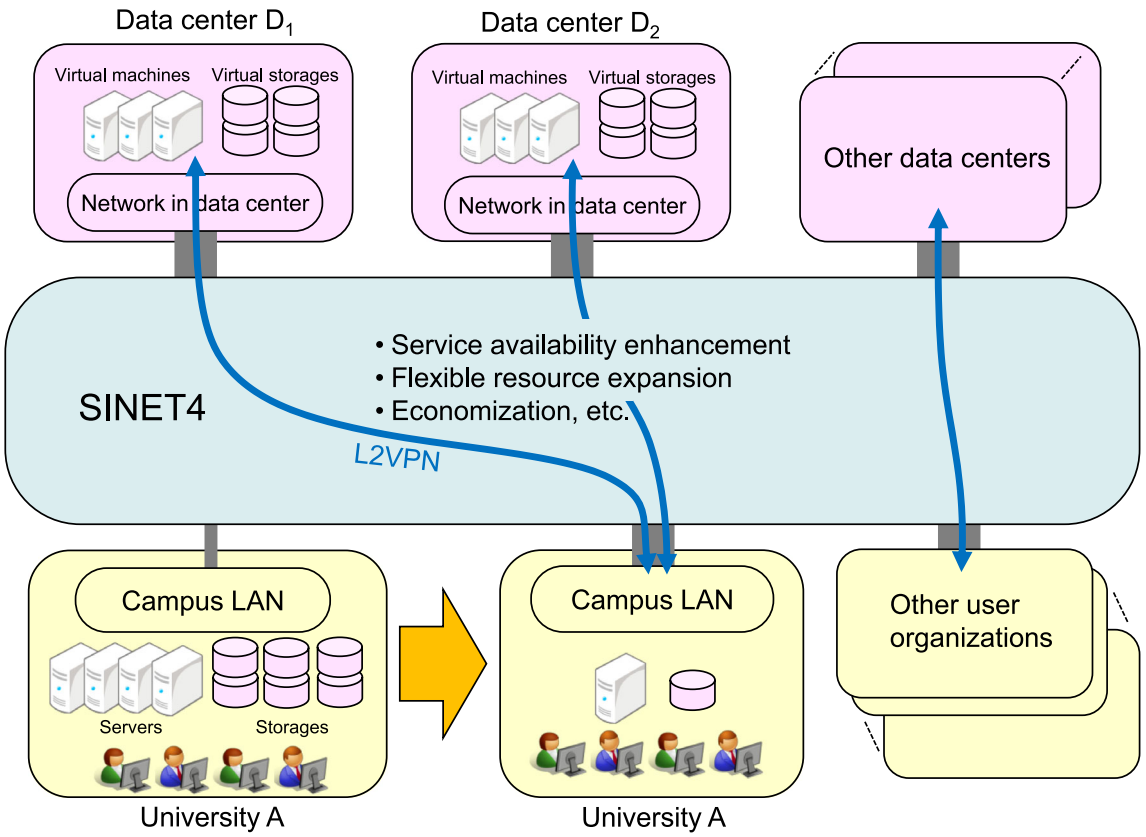

Fig. 13 Movement to private cloud infrastructure

city also lost their connectivity for 68, 1.5, and 8 hours, respectively. The user organizations that were connected to the previous node organizations and had not moved their connection to data centers were severely affected by the disconnectivity even after they recovered. We therefore have encouraged user organizations to move the connection from the previous node organizations to data centers as soon as possible in preparation for future possible disasters.

The impacts of the earthquake did not end with the above-mentioned effects. Because nuclear power plants at Fukushima city, which had supplied electric power to Tokyo and other areas, stopped operating due to the huge *tsunami*, people there were suddenly forced to suffer from rolling blackouts in order to avoid a total electric power shortage. The target areas were divided into five groups, and the rolling blackouts were scheduled in periods such as 9:30–12:10, 11:30–14:10, 13:30–16:10, 15:30–18:10, and 17:30–20:00 on a regular basis. In preparation for the rolling blackouts, user organizations needed to switch off their communications devices as well as servers and storages in order to avoid possible device failures. In addition, user organizations were required to reduce electric power consumption by more than 15 % until September 2011. This situation very negatively affected research and education activities, while it made user organizations accelerate placing their research resources at commercial datacenters in a geographically distributed fashion and also use cloud computing services of cloud service providers through SINET4 (Fig. 13). As we allow the cloud computing service providers that support user organizations' activities to directly connect to SINET4, they can build the "private cloud" infrastructure economically. As some organizations tend to use two or more data centers, we have recently become expected to enhance the service availability more in collaboration with these data centers.

6 Conclusion

This paper described the required specifications, structural design, network components and applied technologies, network resource management, and network management of the new SINET4. The paper also clarified the entire design for a highly available network based on our architecture. It also reported on the impacts of the Great East Japan Earthquake of March 11, 2011, from the viewpoints of SINET4 and user organizations and showed that our highly available network design could keep connectivity even after the earthquake.

Acknowledgements We wish to thank all the members of the Organization of Science Network Operations and Coordination for their

support of SINET4. We are also grateful to Mr. Yasuhiro Kimura of NTT Communications, Mr. Takuro Sono of IIJ, Mr. Takeshi Mizumoto of NTT East, and Mr. Akihiro Sato of NTT ME, Mr. Takumi Mori of Juniper Networks, and Prof. Jun Adachi, Mr. Toyomi Takekawa, Mr. Suguru Sato, Mr. Shinji Takano, Mr. Ken-ichiro Minomo, Mr. Junichi Sayama, Mr. Akitoshi Morishima, Mr. Takaaki Hirabara, and Mr. Yoshihiro Kubota of NII, for their continuous cooperation and support. This work was supported by KAKENHI (23240011).

References

1. *ATLAS at Large Hadron Collider (LHC)*. http://www.atlas.ch/.
2. Berger, L. (2003). *GMPLS signaling resource reservation protocol: traffic engineering*. RFC3473.
3. Çetinkaya, E. K., Broyles, D., Dandekar, A., Srinivasan, S., & Sterbenz, J. P. G. (2013). Modelling communication network challenges for future Internet resilience, survivability, and disruption tolerance: a simulation-based approach. *Telecommunications Systems*, *52*(2), 751–766.
4. Enns, R. (2006). *NETCONF configuration protocol*. RFC4741.
5. GÉANT. http://www.geant.net/.
6. Harada, K., Kawano, T., Zaima, K., Hatta, S., & Meno, S. (2003). Uncompressed HDTV over IP transmission system using ultra-high-speed IP streaming technology. *NTT Technical Review*, *1*(1), 84–89.
7. International Telecommunication Union (2005). *Generic framing procedure (GFP)*. ITU-T recommendation G.7041.
8. International Telecommunication Union (2006). *Link capacity adjustment scheme (LCAS) for virtual concatenated signals*. ITU-T recommendation G.7042.
9. International Telecommunication Union (2003). *Network node interface for the synchronous digital hierarchy (SDH)*. ITU-T recommendation G.707.
10. Internet2. http://www.internet2.edu/.
11. Jarry, A. (2013). Fast reroute paths algorithms. *Telecommunication Systems*, *52*(2), 881–888.
12. JDXnet. http://www.sinet.ad.jp/case-examples/eri.
13. Kawaguchi, N. (2008). Trial on the efficient use of trunk communication lines for VLBI in Japan. In *Proceedings of the 7th international eVLBI workshop*.
14. K computer. http://www.nsc.riken.jp/project-eng.html.
15. Large Helical Device (LHD). e.g. http://www.sinet.ad.jp/case-examples/nifs.
16. Mannie, E., & Papadimitriou, D. (2004). *GMPLS extensions for SONET and SDH control*. RFC3946.
17. Mannie, E., & Papadimitriou, D. (2006). *Recovery (protection and restoration) terminology for generalized multi-protocol label switching*. RFC4427.
18. Nagayama, Y., Emoto, M., Kozaki, Y., Nakanishi, H., Sudo, S., Yamamoto, T., Hiraki, K., & Urushidani, S. (2010). A proposal for the ITER remote participation system in Japan. *Fusion Engineering and Design*, *2010*(85), 535–539.
19. Pan, P., Swallow, G., & Atlas, A. (2005). *Fast reroute extensions to RSVP-TE for LSP tunnels*. RFC4090.
20. Sharma, V., & Hellstrand, F. (2003). *Framework for multi-protocol label switching (MPLS)-based recovery*. RFC3469.
21. SINET4. http://www.sinet.ad.jp/index_en.html?lang=english.
22. Sterbenz, J. P. G., Çetinkaya, E. K., Hameed, M. A., Jabbar, A., Qian, S., & Rohrer, J. P. (2013). Evaluation of network resilience, survivability, and disruption, tolerance: analysis, topology generation, simulation, and experimentation. *Telecommunication systems*, *52*(2), 705–736.
23. TM FORUM (2002). *MTNM implementation statement template and guidelines: NML-EML interface for management of SONET/SDH/WDM/ATM transport networks*. TM FORUM 814A version 2.1.
24. t-Room. http://www.mirainodenwa.com/.
25. Urushidani, S., Abe, S., Ji, Y., Fukuda, K., Koibuchi, M., Nakamura, M., Yamada, S., Hayashi, R., Inoue, I., & Shiomoto, K. (2009). Design of versatile academic infrastructure for multilayer network services. *IEEE Journal on Selected Areas in Communications*, *27*(3), 253–267.
26. Urushidani, S., Shimizu, K., Hayashi, R., Tanuma, H., Fukuda, K., Ji, Y., Koibuchi, M., Abe, S., Nakamura, M., Yamada, S., Inoue, I., & Shiomoto, K. (2009). Implementation and evaluation of layer-1 bandwidth-on-demand capabilities in SINET3. In *Proceedings of IEEE international conference on communications (ICC2009)*.

Shigeo Urushidani is a professor and director at the Research Center for Academic Networks of the National Institute of Informatics (NII). He received B.E. and M.E. degrees from Kobe University in 1983 and 1985, respectively, and received a Ph.D. from the University of Tokyo in 2002. He worked for NTT from 1985 to 2006, where he was engaged in the research, development, and deployment of high-performance network service systems, including ATM, AIN, high-speed IP/MPLS, and GMPLS-based optical systems. He moved to NII in 2006 and is currently involved in the design and implementation of the Japanese academic backbone network, called SINET. His current research interests include network architecture and system architecture for ultra-high-speed green networks. He received the Best Paper Award in 1988, the Young Investigators Award in 1990, and the Communications Society Best Tutorial Paper Award in 2009 from IEICE. He is a member of IEEE.

Michihiro Aoki received B.E., M.E., and Ph.D. degrees in Electronic Engineering from Chiba University in 1981, 1983, and 2008, respectively. In 1983, he joined the Electrical Communication laboratory of Nippon Telegraph and Telephone (NTT) Corporation, where he was engaged in the research and development of high-reliability and high-performance switching systems and IP-routers. He moved to the National Institute of Informatics (NII) in 2009 and is currently a research professor in the Research Center for Academic Networks of NII. He is involved in the design and research of the Japanese academic backbone networks.

Kensuke Fukuda is an associate professor at the National Institute of Informatics (NII). He received his Ph.D. degree in computer science from Keio University at 1999. He worked in NTT laboratories from 1999 to 2005, and joined NII in 2006. In 2002, he was a visiting scholar at Boston University. Concurrently, he is a researcher of PRESTO JST (Sakigake) since 2008. His current research interests are Internet traffic measurement and analysis, intelligent network control architectures, and the scientific aspects of networks.

Shunji Abe received B.E. and M.E. degrees from Toyohashi University of Technology, Japan, in 1980 and 1982, respectively. He received a Ph.D. from the University of Tokyo in 1996. In 1982 he joined Fujitsu Laboratories Ltd., where he engaged in research on broadband circuit switching system, ATM switching system, ATM traffic control, and network performance evaluation. He worked at the National Center for Science Information Systems, Japan (NACSIS) from 1995 to 1999. Since 2000 he has worked at the National Institute of Informatics of Japan as an associate professor, and he is now promoting SINET (Science Information Network) use. He is also an associate professor of the Graduate University for Advanced Studies (SOKENDAI). He is currently interested in the Internet traffic analysis, network performance evaluation, and mobile IP system architecture. He is a member of IEICE, and also a member of IEEE.

Motonori Nakamura graduated from Kyoto University, Japan, where he received B.E., M.E., and Ph.D. degrees in engineering in 1989, 1991, and 1996, respectively. From 1995, he was an associate professor at Kyoto University. Currently he is a professor at National Institute of Informatics, Japan (NII) and the Graduate University for Advanced Studies (SOKENDAI). His research interests are message transport network, network communications, next generation internet and Identity & Access Management. He is a member of IEEE, IEICE, IPSJ and JSSST.

Michihiro Koibuchi received the BE, ME, and PhD degrees from Keio University, Yokohama, Japan, in 2000, 2002, and 2003, respectively. He was a visiting researcher at the Technical University of Valencia, Spain, in 2004 and a visiting scholar at the University of Southern California, in 2006. He is currently an associate professor in the Information Systems Architecture Research Division, National Institute of Informatics, Tokyo, and the Graduate University for Advanced Studies, Japan. His research interests include the areas of high-performance computing and interconnection networks. He is a member of the IEEE, IPSJ and IEICE.

Yusheng Ji received B.E., M.E., and D.E. degrees in electrical engineering from the University of Tokyo in 1984, 1986, and 1989, respectively. She joined the National Center for Science Information Systems, Japan (NACSIS) in 1990. Currently, she is a professor at the National Institute of Informatics, Japan (NII), and the Graduate University for Advanced Studies (SOKENDAI). Her research interests include network architecture, traffic control, and performance analysis for quality of service provisioning in communication networks. She is a member of IEEE, IEICE and IPSJ.

Shigeki Yamada is currently a Professor and Director with Principles of Informatics Research Division, National Institute of Informatics, Japan. He received B.E., M.E., and Ph.D. degrees in Electronic Engineering from Hokkaido University, Japan in 1972, 1974, and 1991, respectively. He worked in the NTT (Nippon Telegraph and Telephone Corporation) laboratories from 1974 to 1999, where he was involved in research and development on digital switching systems and information and communication networks. He moved to NII in 2000. From 1981 to 1982, he was a visiting scientist in Computer Science Department, University of California, Los Angeles. His current research interest includes future Internet technologies, clean slate designed network architectures, mobile and wireless networks, and privacy enhancing technologies. He is a senior member of the IEEE, and a member of the IEICE and IPSJ.