OCN Experience to Handle the Internet Growth and the Future

Takeshi Tomochika <takeshi.tomochika(at)ntt.com>
Chika Yoshimura <chika.yoshimura(at)ntt.com>

NTT Communications, OCN

Background

- Internet traffic / full routes are growing more and more
- One of the most important missions of ISPs
  - to carry the traffic with stability and without any congestion
- Making the backbone robust
- We will talk about:
  - current traffic situation in Japan
  - issues at OCN when designing the backbone network
  - future visions

Agenda

1. Current situation of Internet traffic in Japan

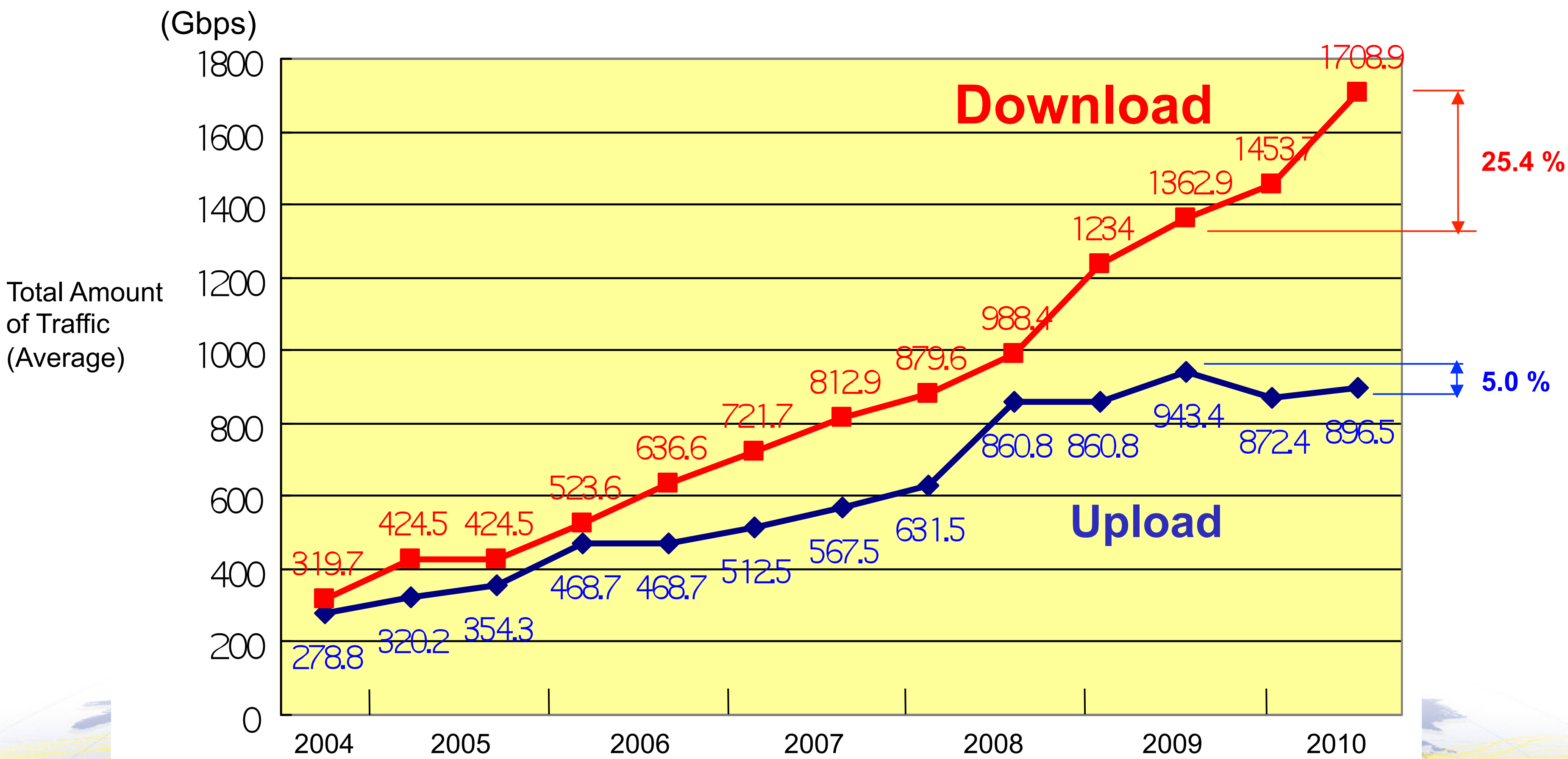
2. What is OCN?

3. Current issues we are facing
   - Router Forwarding Table
   - Link Aggregation

4. Future Plan

Internet Traffic Trend in Japan

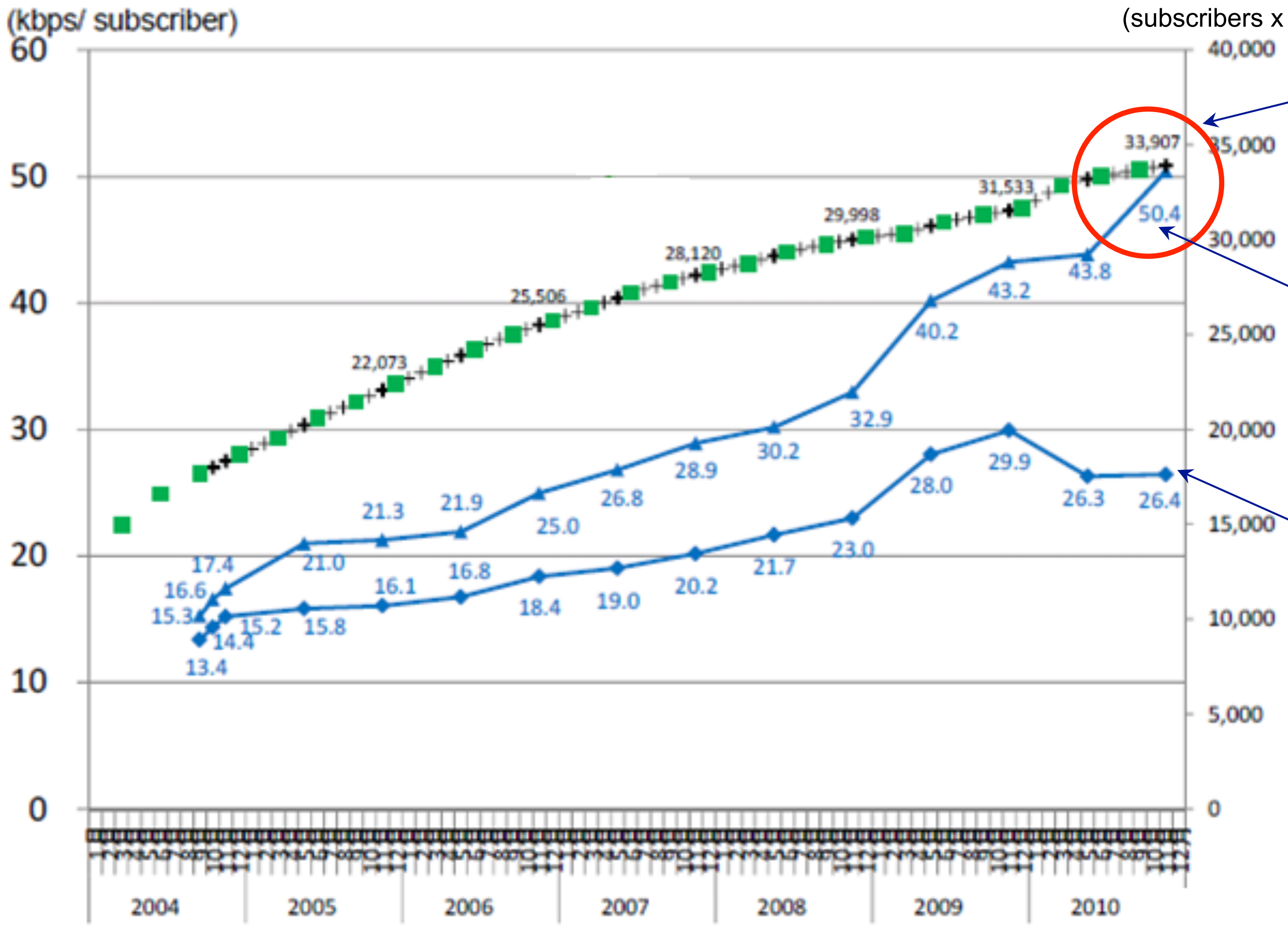
- Total amount of broadband traffic is <u>1.7Tbps</u> (Download)
  - 25.4% growth compared to last year
- Upload traffic decreased over the last year (896Gbps)



source: **Internet Traffic Trends in Japan (** Ministry of Internal Affairs and Communications)
http://www.soumu.go.jp/menu_news/s-news/01kiban04_01000006.html (Japanese)

Internet Traffic Trend in Japan (cont.)

- The number of broadband subscribers and the traffic volume per subscriber are growing



The number of broadband subscribers in Japan : **33,907,000** in Nov 2010

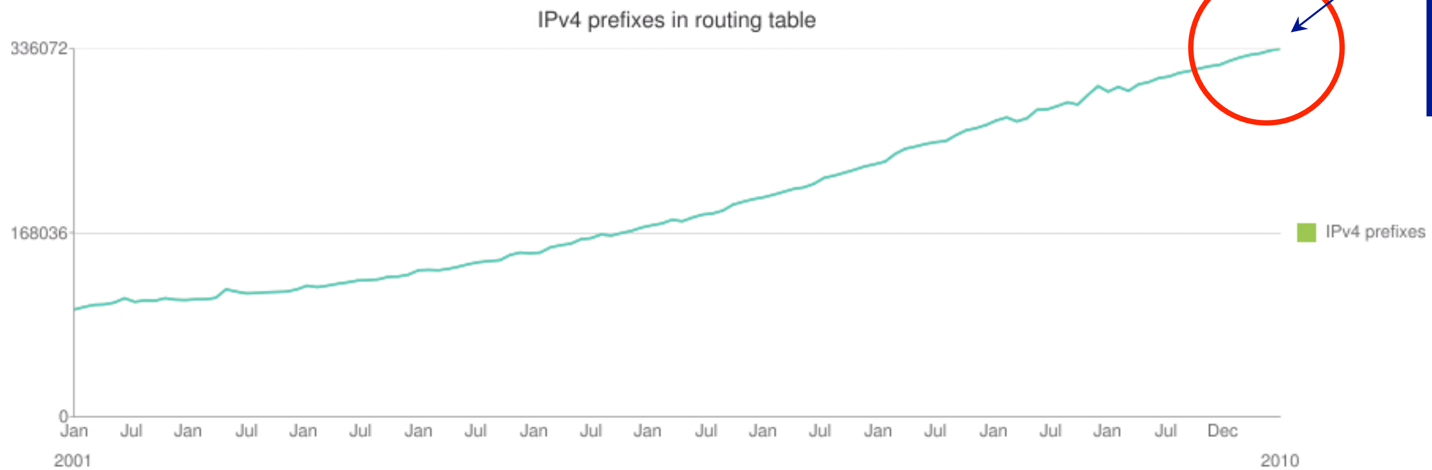
The download traffic per subscriber : **50kbps** in Nov 2010

The upload traffic per subscriber : **26kbps** in Nov 2010

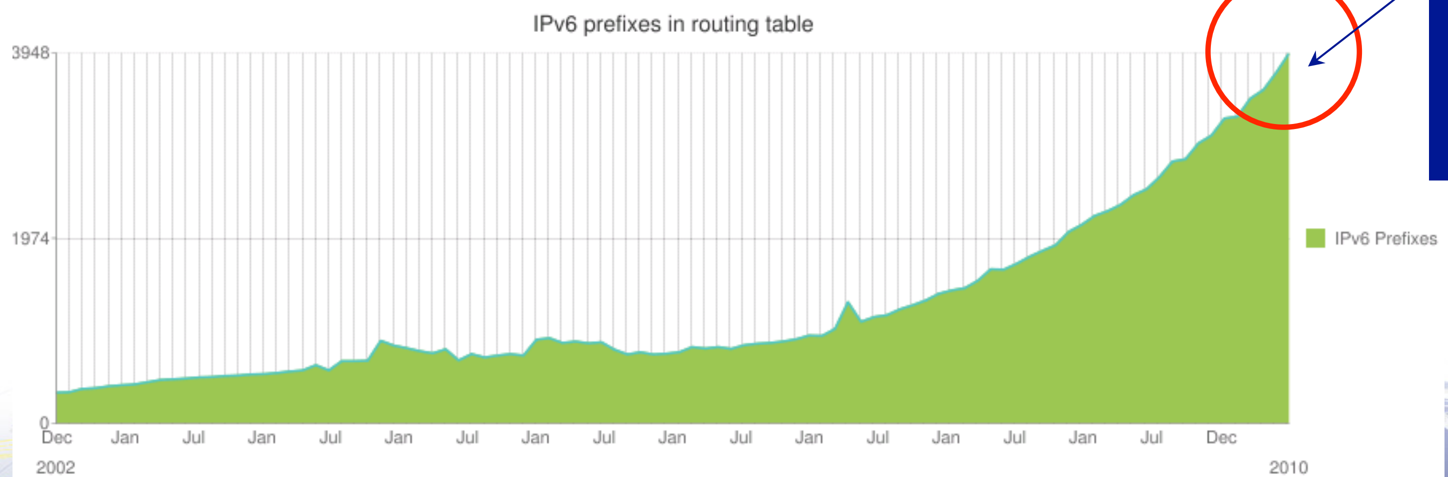
source: Internet Traffic Trends in Japan (Ministry of Internal Affairs and Communications)
http://www.soumu.go.jp/menu_news/s-news/01kiban04_01000006.html (Japanese)

Internet Full Routes Trend

- Internet full routes growing



The number of IPv4 prefix : **over 330,000** in June 2011



The number of IPv6 prefix : **over 6,000** in June 2011

source: BGPmon http://bgpmon.net/stat.php

Overview

- Internet traffic in Japan / full routes have been growing consistently

- Traffic will keep rising in the future
  - ISPs have to …
    - design a robust backbone network to deal with the situation

- The backbone we have been making

- The bandwidth we have

Agenda

1. Current situation of Internet traffic in Japan

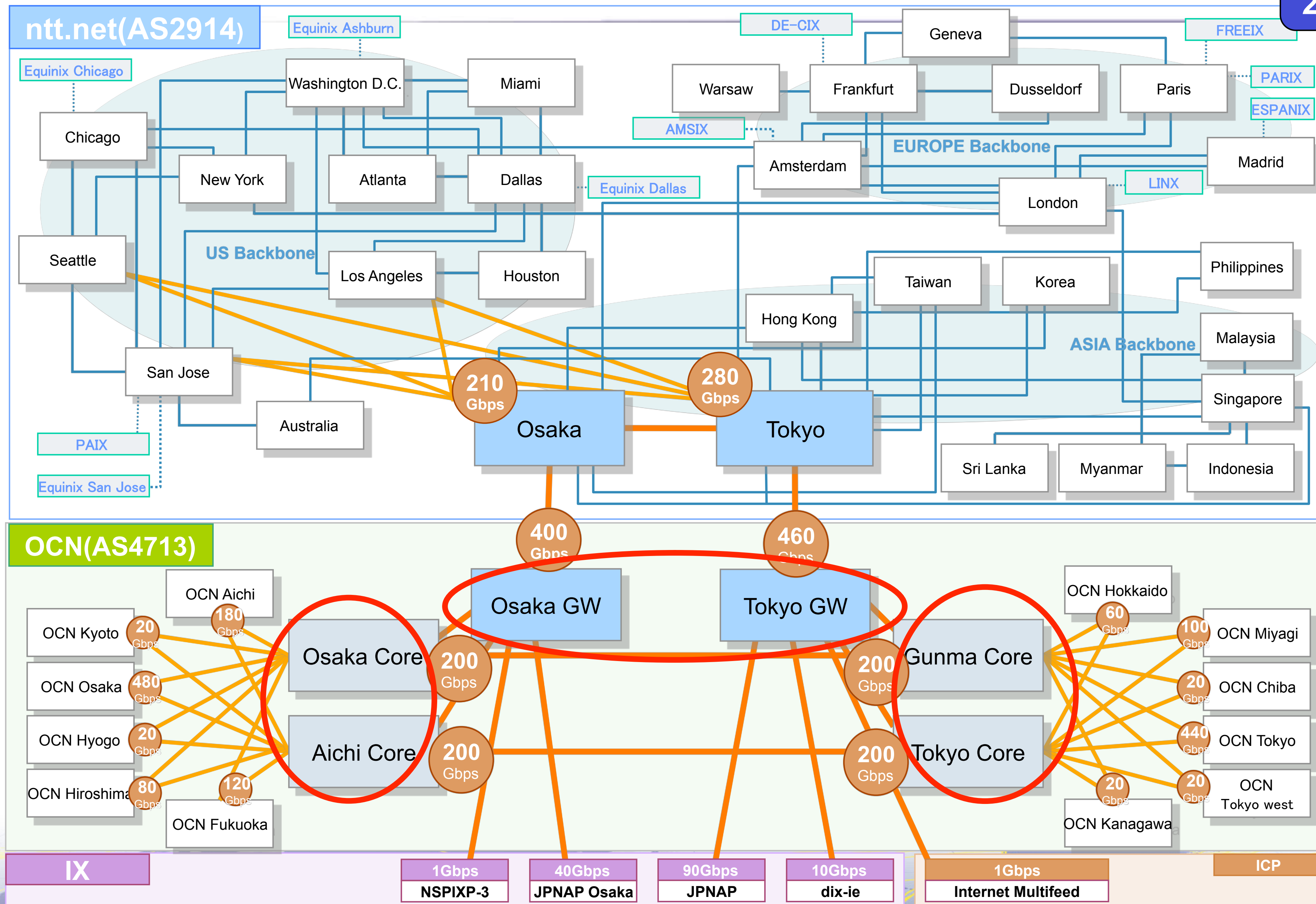
2. **What is OCN?**

3. Current issues we are facing
   - Router Forwarding Table
   - Link Aggregation

4. Future Plan

NTT Communications' IP Backbone Network (ntt.net & OCN)

May 2011



Network Design Policy of OCN

Full redundant network

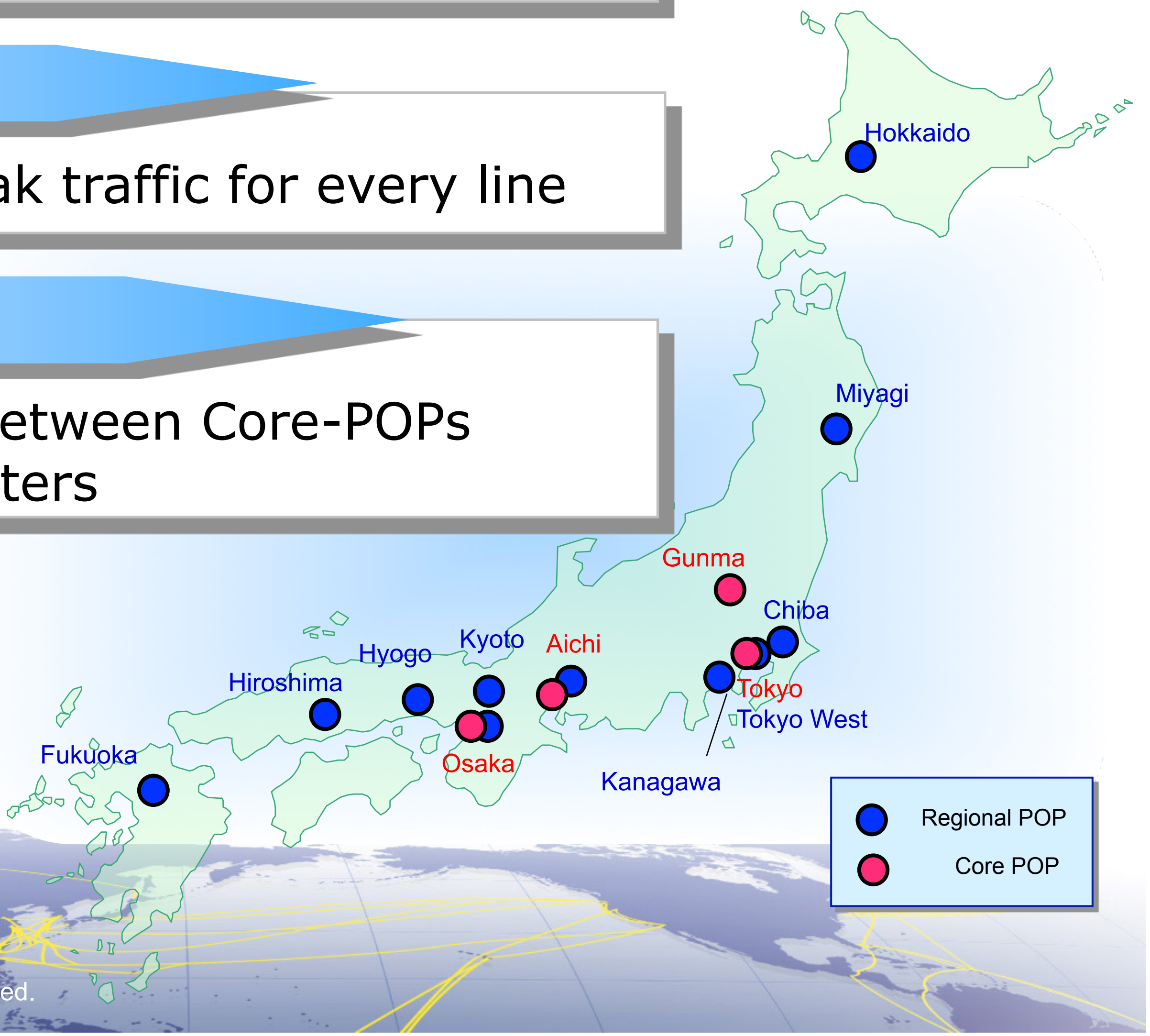
No single point of failure

100% Traffic Relief

Double bandwidth of the peak traffic for every line

Disaster Tolerance

More than 100km distance between Core-POPs minimize impact of the disasters



Agenda

1. Current situation of Internet traffic in Japan

2. What is OCN?

3. **Current issues we are facing**
   - Router Forwarding Table
   - Link Aggregation

4. Future Plan

The issues we are facing

1. Routes Growth
   Scalability of Router Forwarding Tables

2. Traffic Growth
   Link Aggregation

FIB table of OCN

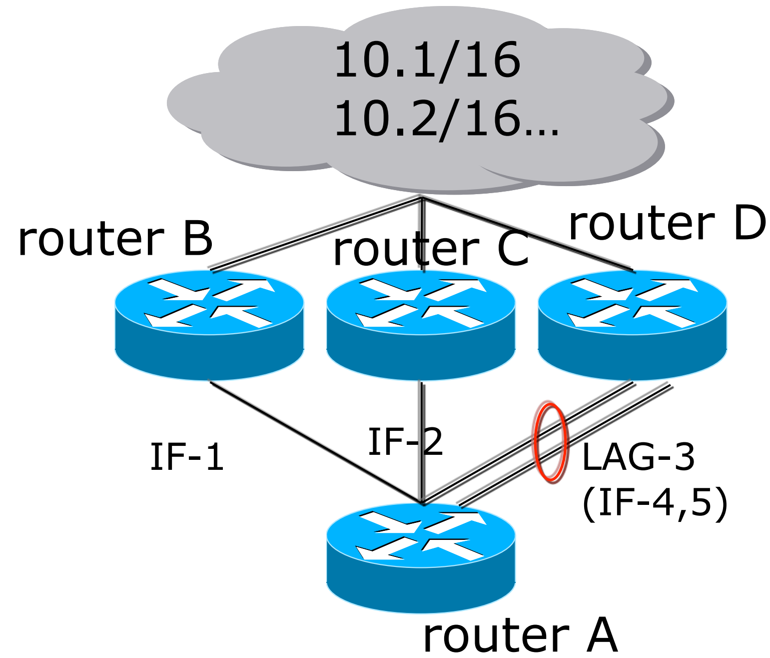
- FIB(Forwarding Information Base) table has been growing
  - 410,000 routes in OCN (June 2011)

- Causes of growing FIB
  1. BGP full routes
  2. Prefixes with no-export (several tens of thousands in OCN)
  3. ECMP, {i, e} bgp-multipath

Scalability of Router Forwarding Tables

- When a rerouting event occurs, potentially thousands of routes must be updated

FIB of router-A	
prefix	output interface(s)
10.1.0.0/16	IF-1
	IF-2
	LAG-3(IF-4, 5)
10.2.0.0/16	IF-1
	IF-2
	LAG-3(IF-4, 5)

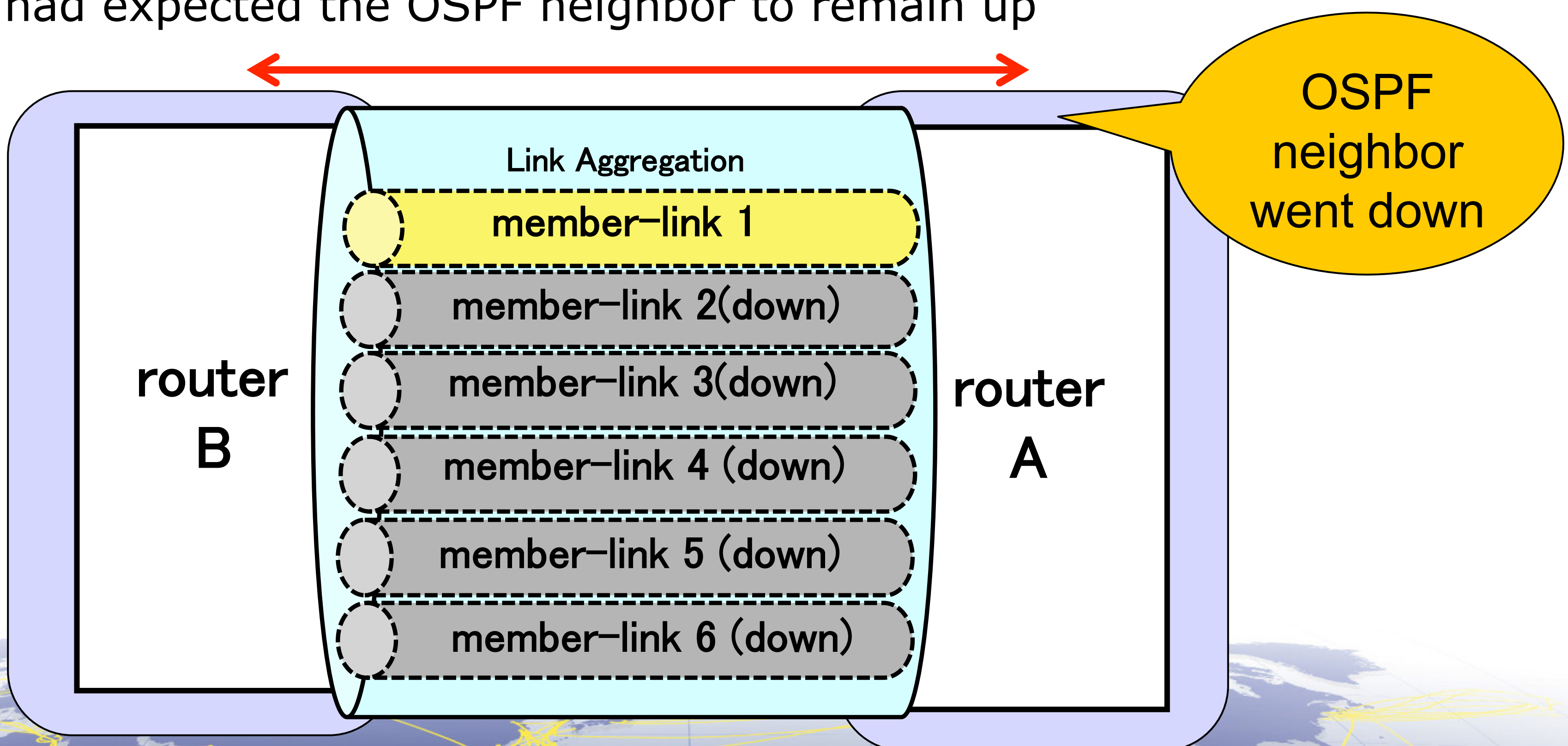


- It took a lot of time to converge the routes
  - When some member-links of a link aggregation were taken down

	FIB table(IPv4)	Convergence time (flattened FIB)
a certain router	360,000	more than 130sec
	500,000	more than 210sec

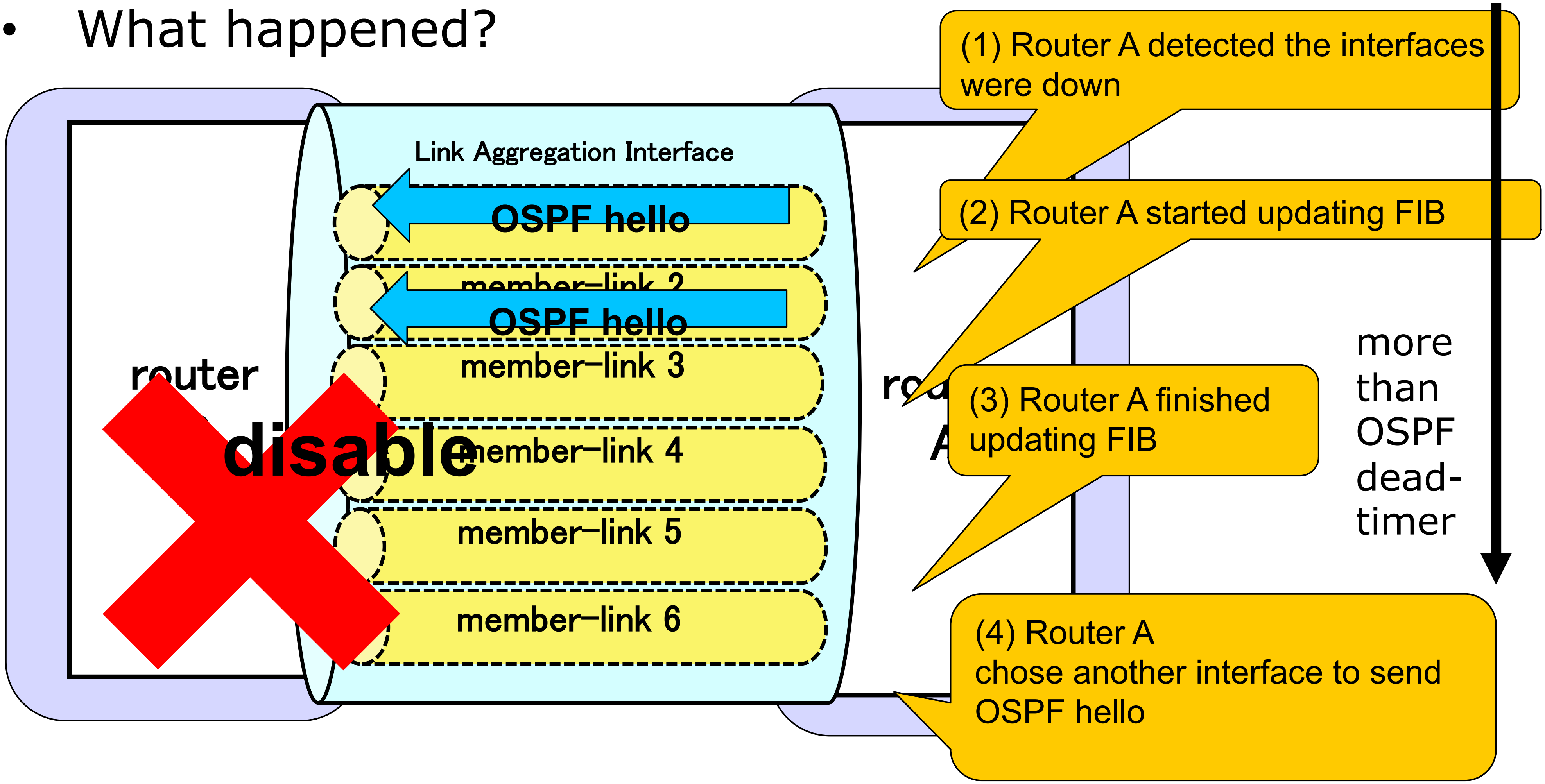
Scalability of Router Forwarding Tables

- We were facing a problem:
  - OSPF neighbor went down due to FIB table convergence
- Between router A and B
  - Link Aggregation (LAG) had been enabled (minimum-links = 1)
  - OSPF neighbor had been connected through the LAG interface
- When all member-links but one had been disabled
  - We had expected the OSPF neighbor to remain up



Scalability of Router Forwarding Tables

- What happened?



Router-A could not send any OSPF hello packets during (1) – (3), then the neighbor went down

Scalability of Router Forwarding Tables

- Hierarchical FIB
  - Cisco: BGP Prefix Independent Convergence(PIC)
  - Juniper: indirect-nexthop

  For more information: BGP Convergence in much less than a second
  http://www.nanog.org/meetings/nanog40/presentations/ClarenceFilsfils-BGP.pdf

- Fewer routes to be updated

- Improving the route convergence time

a certain router	FIB table(IPv4)	Convergence time (flattened FIB)	Convergence time (hierarchical FIB)
	500,000	more than 210sec	around 25sec

Agenda

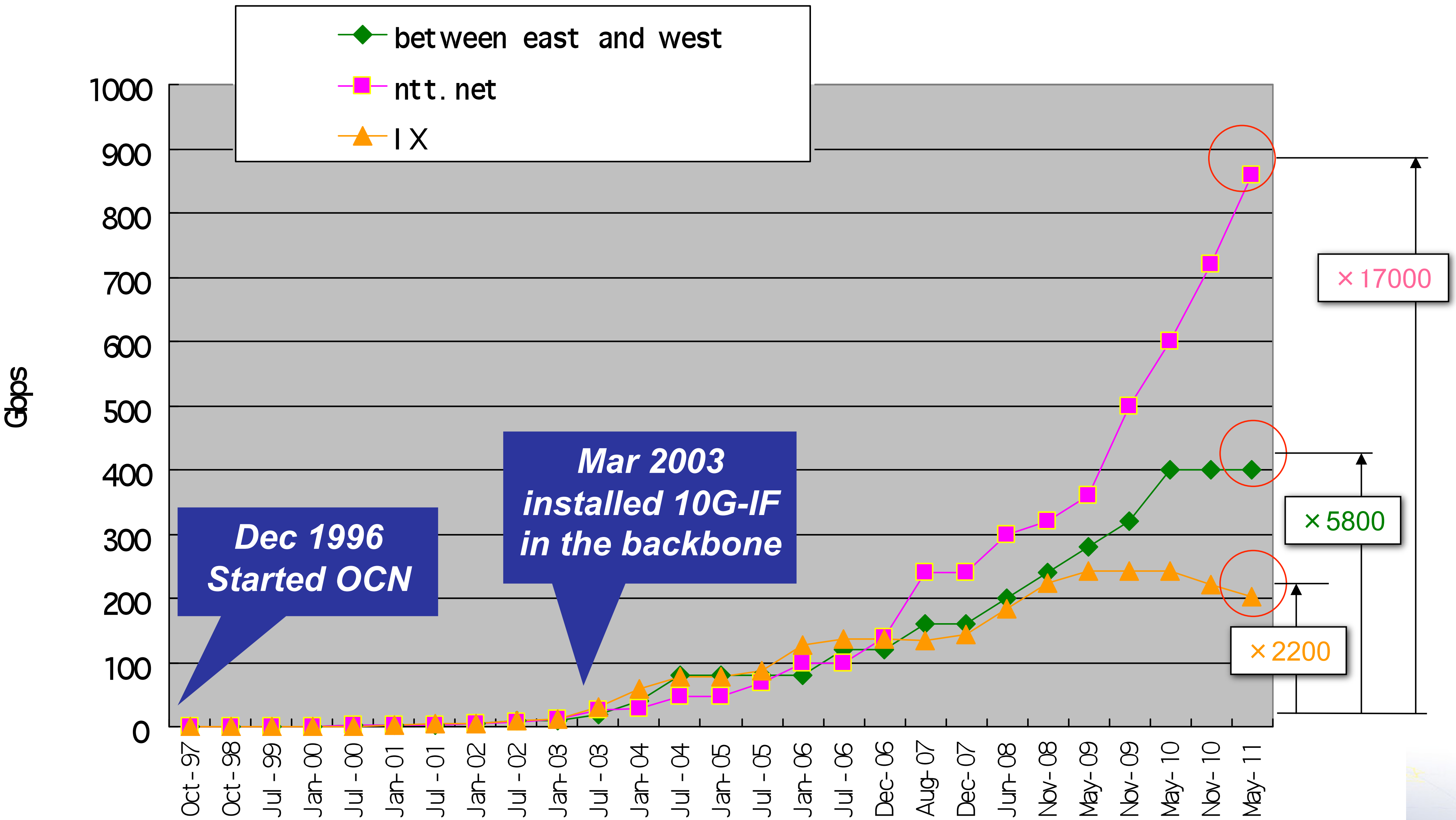
1. Current situation of Internet traffic in Japan

2. What is OCN?

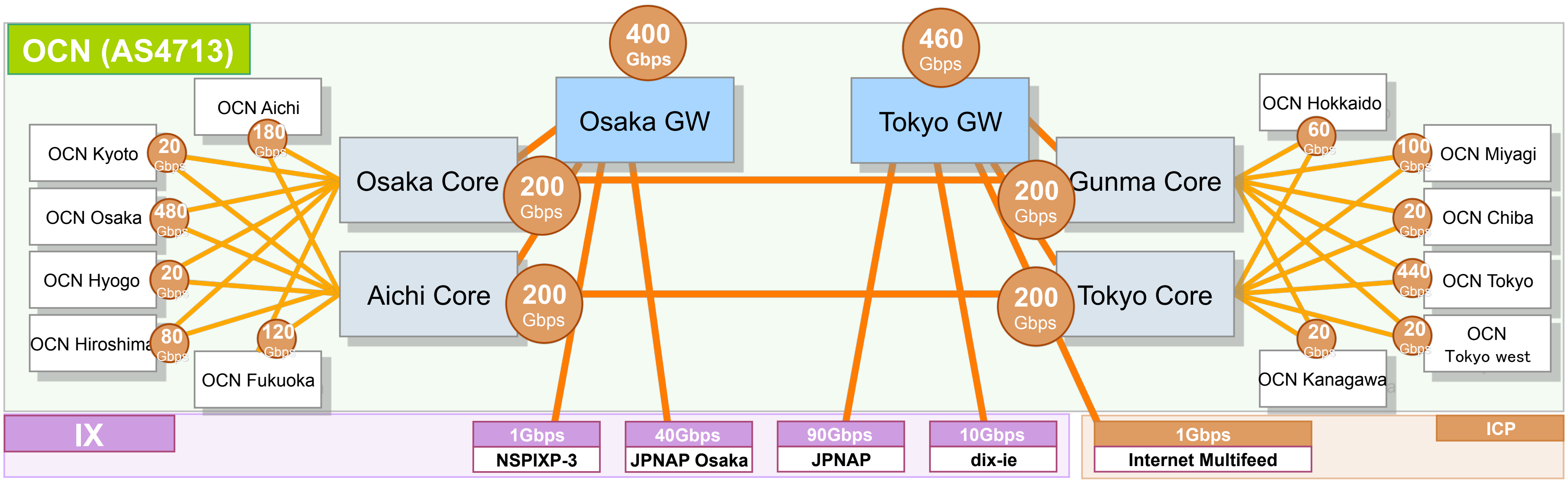
3. **Current issues we are facing**
    - Router Forwarding Table
    - **Link Aggregation**

4. Future Plan

Bandwidth History of OCN



A lot of Link Aggregation in OCN



- A large number of 10GE Interfaces
- A lot of Link Aggregation 10GE Interfaces in the backbone

Link Aggregation Issues

A) Traffic load-balancing issues (Traffic Polarization)
- Background
1. Traffic-unbalance by variation of flow - may skip -
2. Limited number of hash elements
3. Combination of ECMPs and LAGs
   - Case 1: ECMP and LAG at the same Node
   - Case 2: ECMP and LAG at different Node
   - Case 3: ECMP and ECMP at different Node

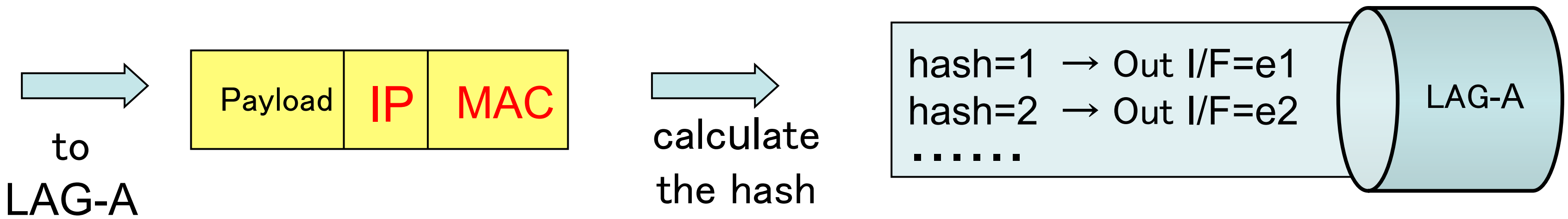
B) Operational Considerations
1. LACP - might skip -
2. minimum-links - may skip -
3. Ping to each physical interface

C) Other issues

A) Traffic load-balancing issues: Background

- Condition of traffic load-balancing method in the LAG
  - ***Can't use per-packet round-robin***
    - Simple round-robin bring about packet reordering in a flow
    - Should use flow-based traffic load-balancing method

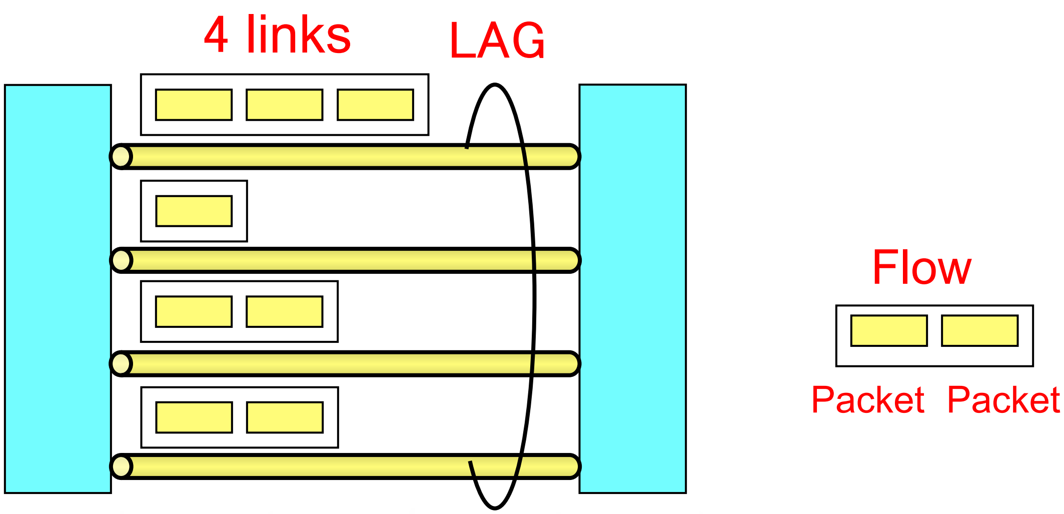
- Hash value is used for flow-based traffic load-balance
- Hashing algorithm: calculate the hash value based on the packet information (IP address, MAC address, and etc.) to decide Output I/F



A) Traffic load-balancing issues

- Issue 1: Traffic-unbalance by variation of flow
  - Each flow has each size
  - Small issue
    - Each 10Gbps physical link has a huge number of flows

Skip this slide due to limited time



A) Traffic load-balancing issues

- Issue 2: Limited number of hash elements
  - Due to this, traffic cannot be evenly distributed
    - ➢Less effective use of bandwidth
  - The less # of hash elements, the worse traffic balance

e.g.: Traffic balance in a LAG when # of hash elements is 8

5 link 10GE LAG	4 link 10GE LAG	3 link LAG	
IF#1 H1、H6 IF#2 H2、H7 IF#3 H3、H8 IF#4 H4 IF#5 H5	IF#1 H1、H5 IF#2 H2、H6 IF#3 H3、H7 IF#4 H4、H8	IF#1 H1、H4、H7 IF#2 H2、H5、H8 IF#3 H3、H6	
2:2:2:1:1	2:2:2:2	3:3:2	<- Traffic balance ratio
10+10+10+10*1/2+10 *1/2 = 40	10+10+10+10=40	10+10+10*2/3 = 26.7	<- Effective bandwidth in the LAG

Use only 40G / 50G

Use only 27G / 30G

A) Traffic load-balancing issues

- cf. Difference in traffic load-balance by # of hash elements

e.g.1: Traffic balance in a LAG when # of hash elements is 8

5 links LAG	4 links LAG	3 links LAG
IF#1 H1, H6 IF#2 H2, H7 IF#3 H3, H8 IF#4 H4 IF#5 H5	IF#1 H1, H5 IF#2 H2, H6 IF#3 H3, H7 IF#4 H4, H8	IF#1 H1, H4, H7 IF#2 H2, H5, H8 IF#3 H3, H6
40	40	26.7

The more # of hash elements, the better traffic balance

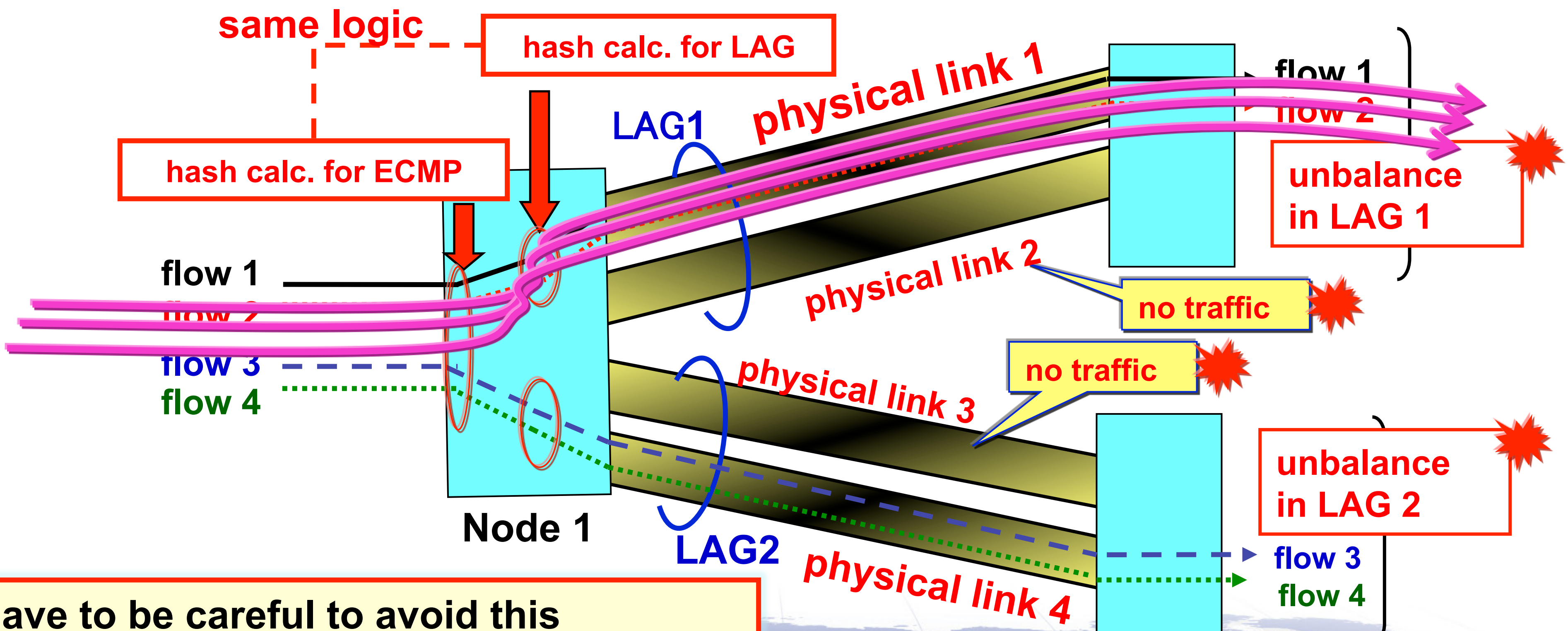
Should avoid odd number of member-links in a LAG

e.g.2: Traffic balance in a LAG when # of hash elements is 32

5 links LAG	4 links LAG	3 links LAG
IF#1 H1, H6, ･･･H26, H31 IF#2 H2, H7, ･･･H27, H32 IF#3 H3, H8, ･･･H28 IF#4 H4, H9, ･･･H29 IF#5 H5, H10, ･･･H30	IF#1 H1, H5, ･･･H29 IF#2 H2, H6, ･･･H30 IF#3 H3, H7, ･･･H31 IF#4 H4, H8, ･･･H32	IF#1 H1, H4, ･･･H28, H31 IF#2 H2, H5, ･･･H29, H32 IF#3 H3, H6, ･･･H30
7:7:6:6:6 10+10+10*6/7+10*6/7＋ 10*6/7 = 45.7	8:8:8:8 10+10+10+10 = 40	11:11:10 10+10+10*10/11 = 29.1

A) Traffic load-balancing issues

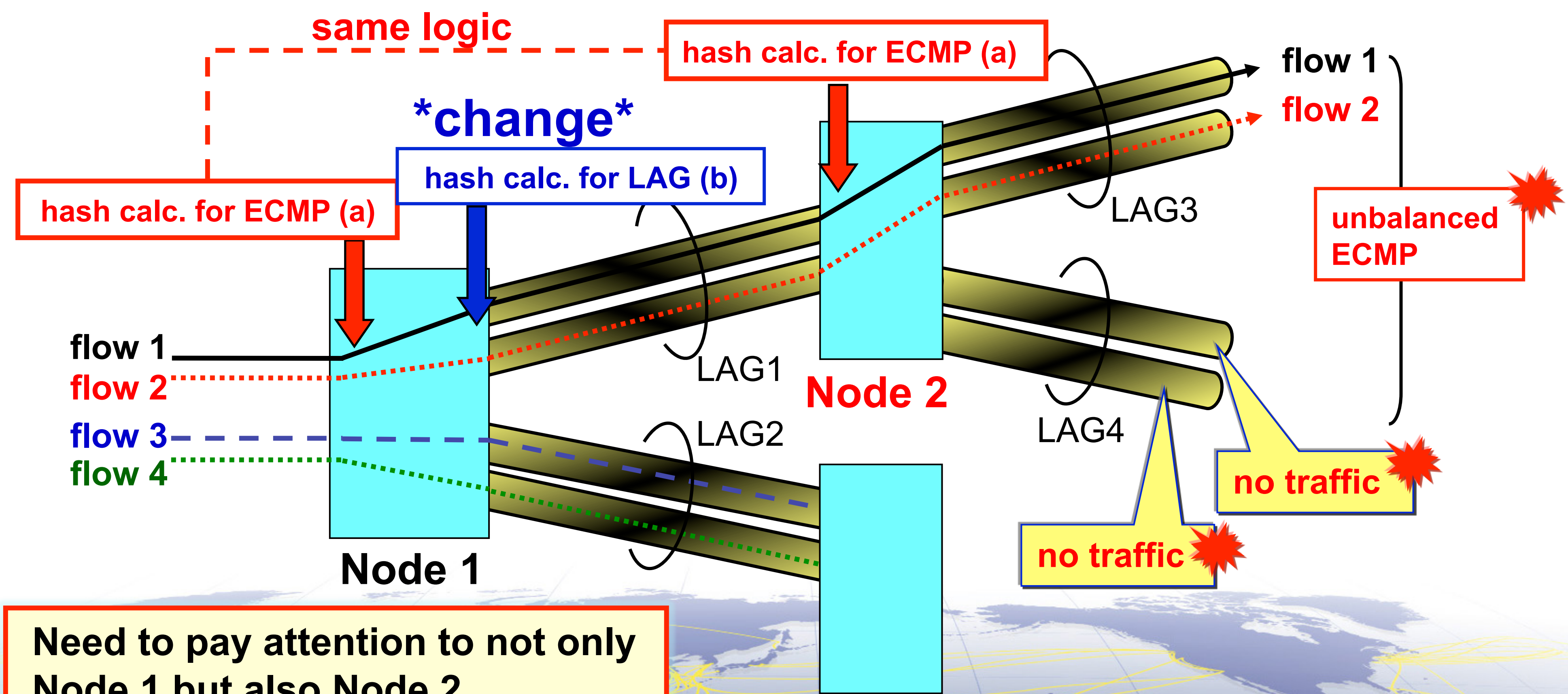
- Issue 3: Combination of ECMPs (Equal Cost Multi Path) and LAGs
- Case 1:
  - When hash calculation logic of LAG is the same as ECMP's, it will bring about unbalanced traffic in physical links



- Have to be careful to avoid this
- Some routers have the same calculation logics for ECMP and LAG as a default

A) Traffic load-balancing issues

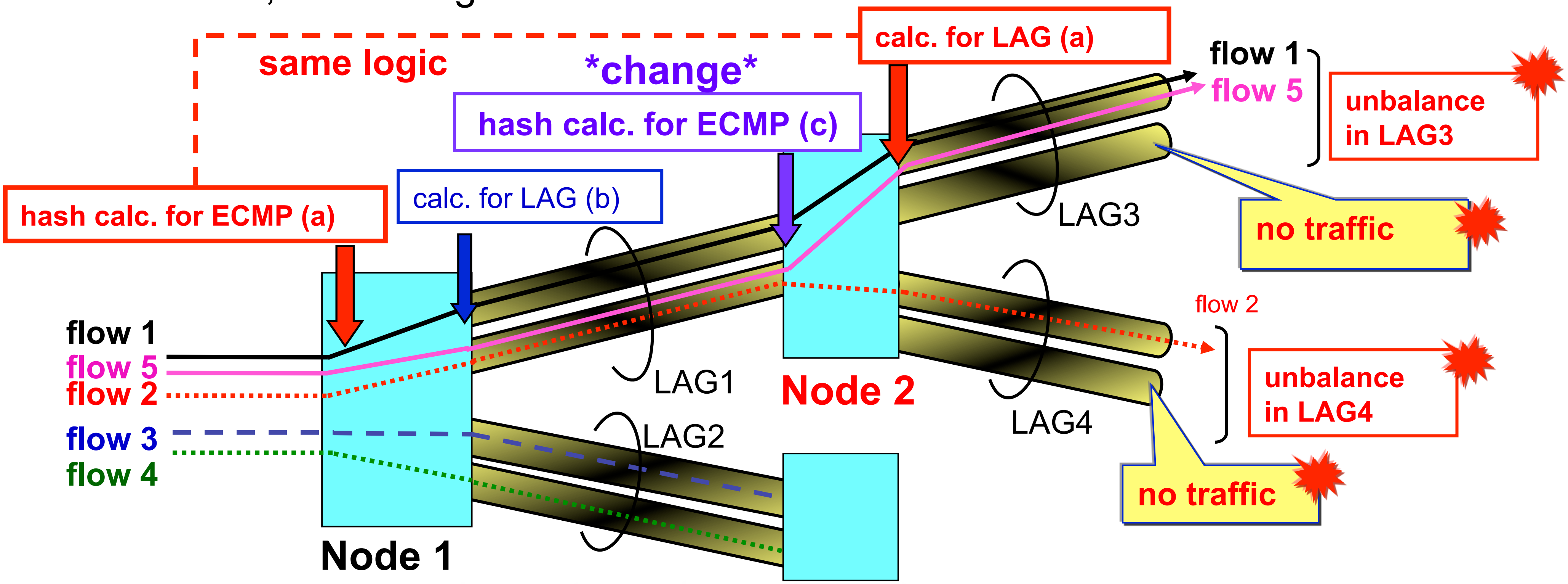
- Issue 3: Combination of ECMPs and LAGs
- Case 2:
  - When calculation logic of ECMP is the same as that of next node, it will bring about unbalanced traffic



Need to pay attention to not only Node 1 but also Node 2

A) Traffic load-balancing issues

- Issue 3: Combination of ECMPs and LAGs
- Case 3:
  - When calculation logic of ECMP is the same as that of LAG at the next node, it will bring about unbalanced traffic



* Some latest routers can include a router-ID in the seed of hash to avoid case 2,3

Need to consider balance logics, network topology, configurations

B) Operational Considerations

- Consideration 1:
  - In the case of silent-failure, traffic through the fault link will drop

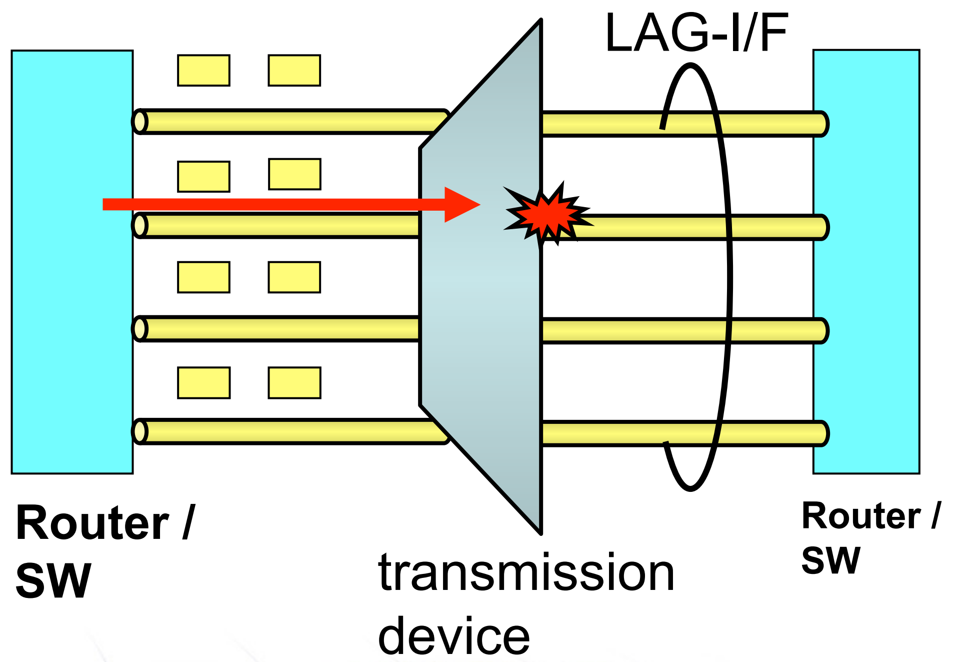
Might skip this slide due to limited time

LACP (Link Aggregation Control Protocol)

・Send and receive control frames in physical links

・Attention to detail Interoperability

- Basically good
- Different default mode (fast / slow)
- Different reaction to null ID (bug) LACP (keep down / once down then go up)

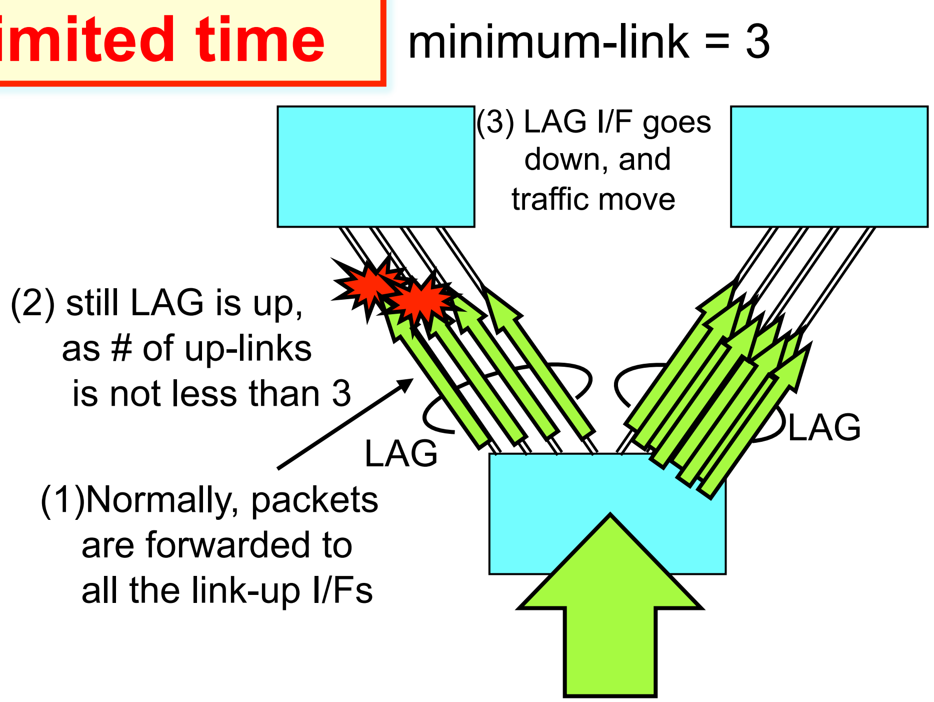


BFD Per Member Link
(Bidirectional Forwarding Detection)

B) Operational Considerations

May skip this slide due to limited time

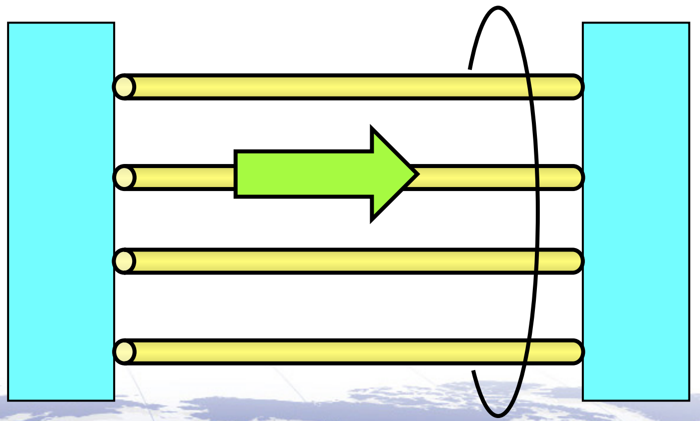
- Consideration 2:
  - Switching policy of LAG-I/F
    - minimum-link (trunk-threshold)
    - threshold whether LAG-I/F is up or down
    - This switching policy is important for effective use of LAG
    - **should consider the entire network topology to use minimum-links**



# of links in LAG	3	4	5	6	7	8	9	10
minimum-link	3	3	4	5	5	6	7	7

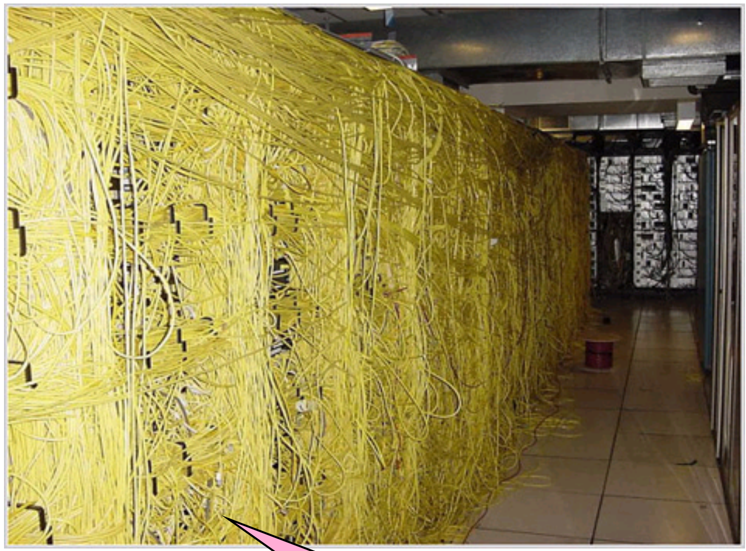
B) Operational Considerations

- Consideration 3:
  - Ping for test
    - Packet goes through only one physical interface
    - Need to test each interface with letting the rest go down
    - Some recent routers and switches support Ethernet OAM to avoid this troublesome job



C) Other Issues

- Limitations on # of links in a LAG
- Issues of physical wiring
  - Increased # of physical links
    -> Complicated maintenance
- Need a well-thought-out plan for LAG
  - How to assign physical links to Line Cards
    - Redundant policy
    - MTBF for each part
    - Cost, etc.
  - e.g. Policy 1: keep LAG-I/F up as much as possible
    - assign each physical link to each LC, minimum-link = 1
  - e.g. Policy 2: Switching traffic to the other LAG immediately
    - assign all physical links to one LC, minimum-link = # of links
  - e.g. Policy 3: Between policy 1 and policy 2
- **LAG is troublesome**
  - **many LAGs, many member-links**



NOTE: this is NOT NTT Communications' equipment

Agenda

1. Current situation of Internet traffic in Japan

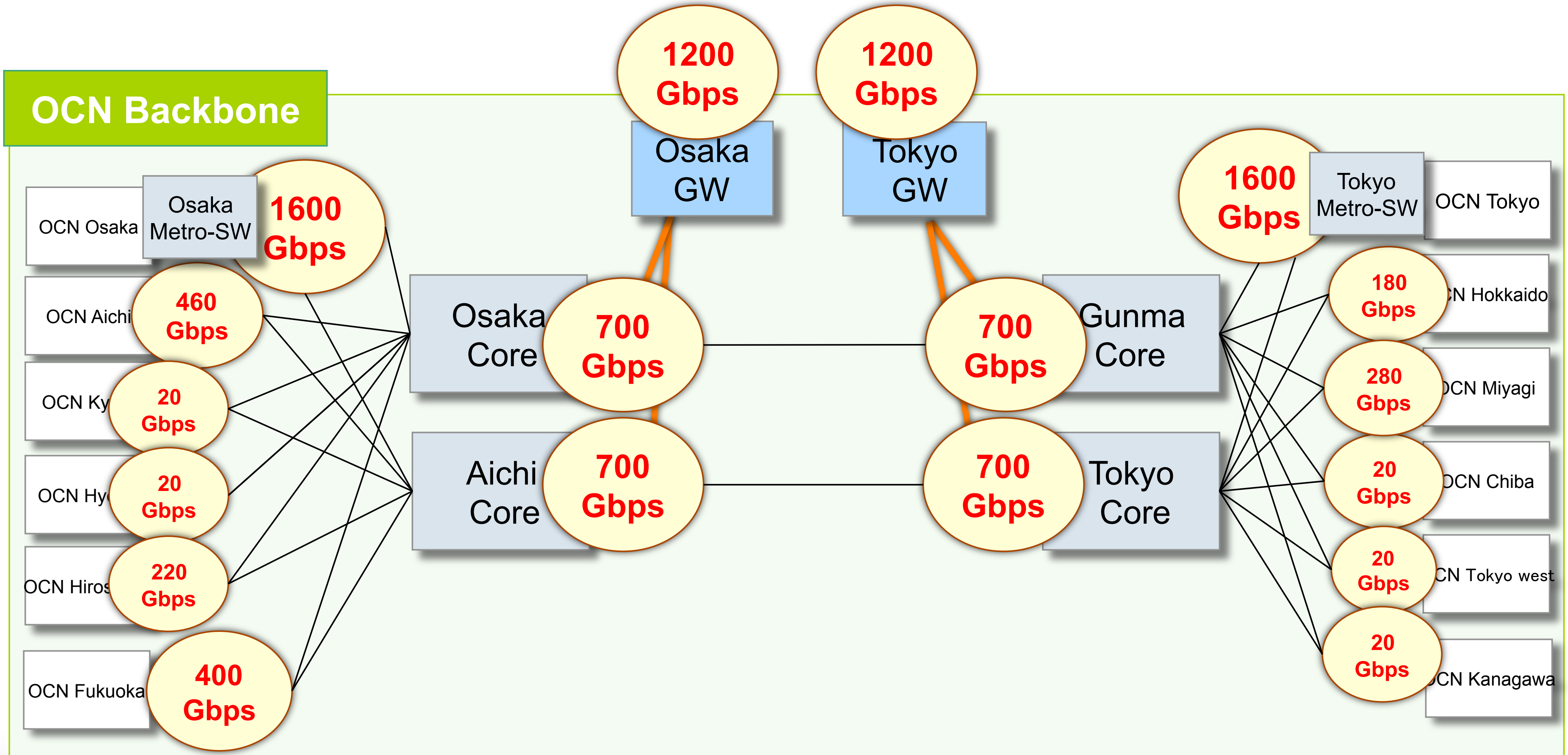
2. What is OCN?

3. Current issues we are facing

4. **Future Plan**

OCN future plan

- **More bandwidth**

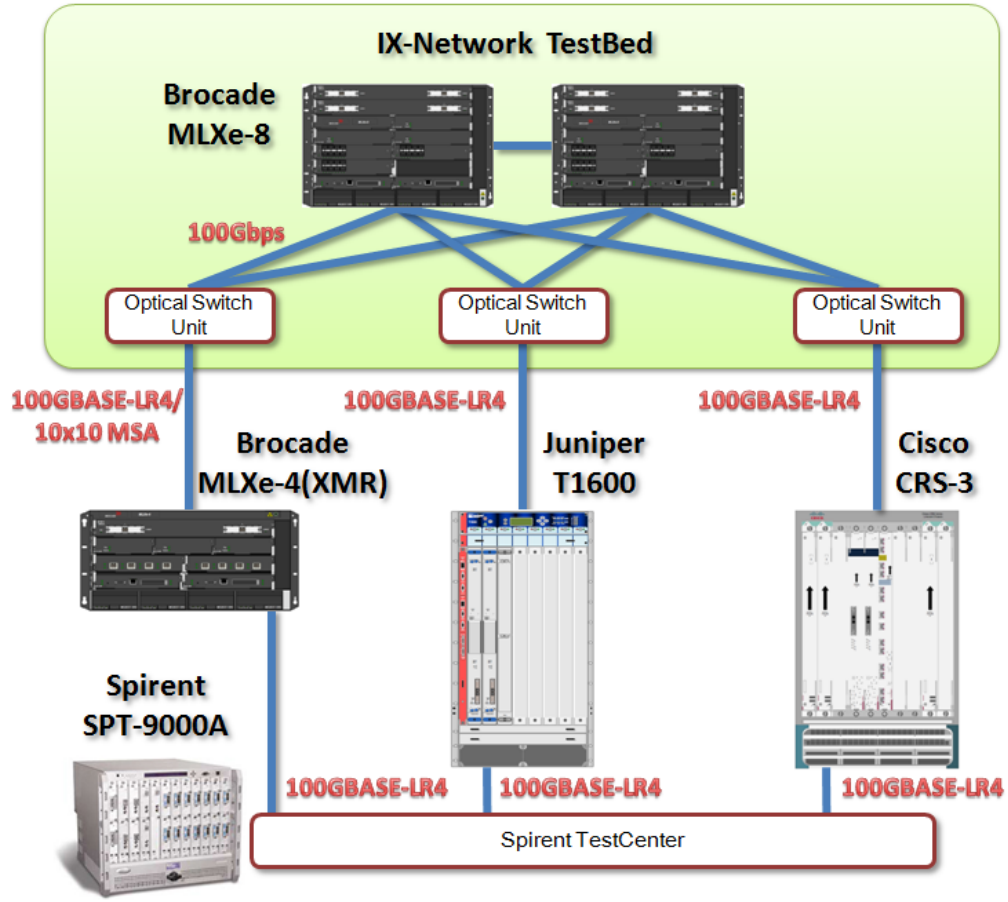


Expectation for 100GE

- Need 100GE I/Fs
  - Bandwidth over 1Tbps
  - LAG is troublesome
- Request
  - Lower price
    - CFP is expensive
    - 10 x10 MSA (LR10)
  - Long-distance transmission (ER4)
  - Higher Capacity
    - Capacity per chassis will be decreased when migrating from 10GEs to 100GEs in some current routers
  - LAG of 10GE and 100GE simultaneously
  - good Interoperability, easy-operation 100GE LAG, convenient Ether OAM
  - Next step: 400GE, 1T Ether

100GE Joint Interoperability Test at JPNAP

- Brocade, Cisco, Juniper and Toyo Corporation (Spirent)
- JPNAP, IIJ, and NTT Communications
- Success of 100 Gigabit Ethernet joint interoperability test at IX
- Confirmed the interoperability between different vendors' products especially at an IX environment
  - Good interoperability
  - Some small issues with each product
    - feedback to vendors with some requests



- Further information is available at: http://www.mfeed.co.jp/english/press/2011/20110601-e.html

Summary

- The traffic in Japan and BGP table has been consistently growing
- We need to consider growth of both routes and traffic to keep our backbone stable
- LAG is troublesome
- We need 100GE to deal with the traffic growth

Thank you!