" OCN Experience to Handle the Traffic Growth and the Future "

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

NTT Communications, OCN

23th February 2011

Background

- Internet traffic is growing more and more
- One of the most important missions of ISPs
  - to carry the traffic with stability & without any congestion
- Making the backbone robust
- We are talking 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?

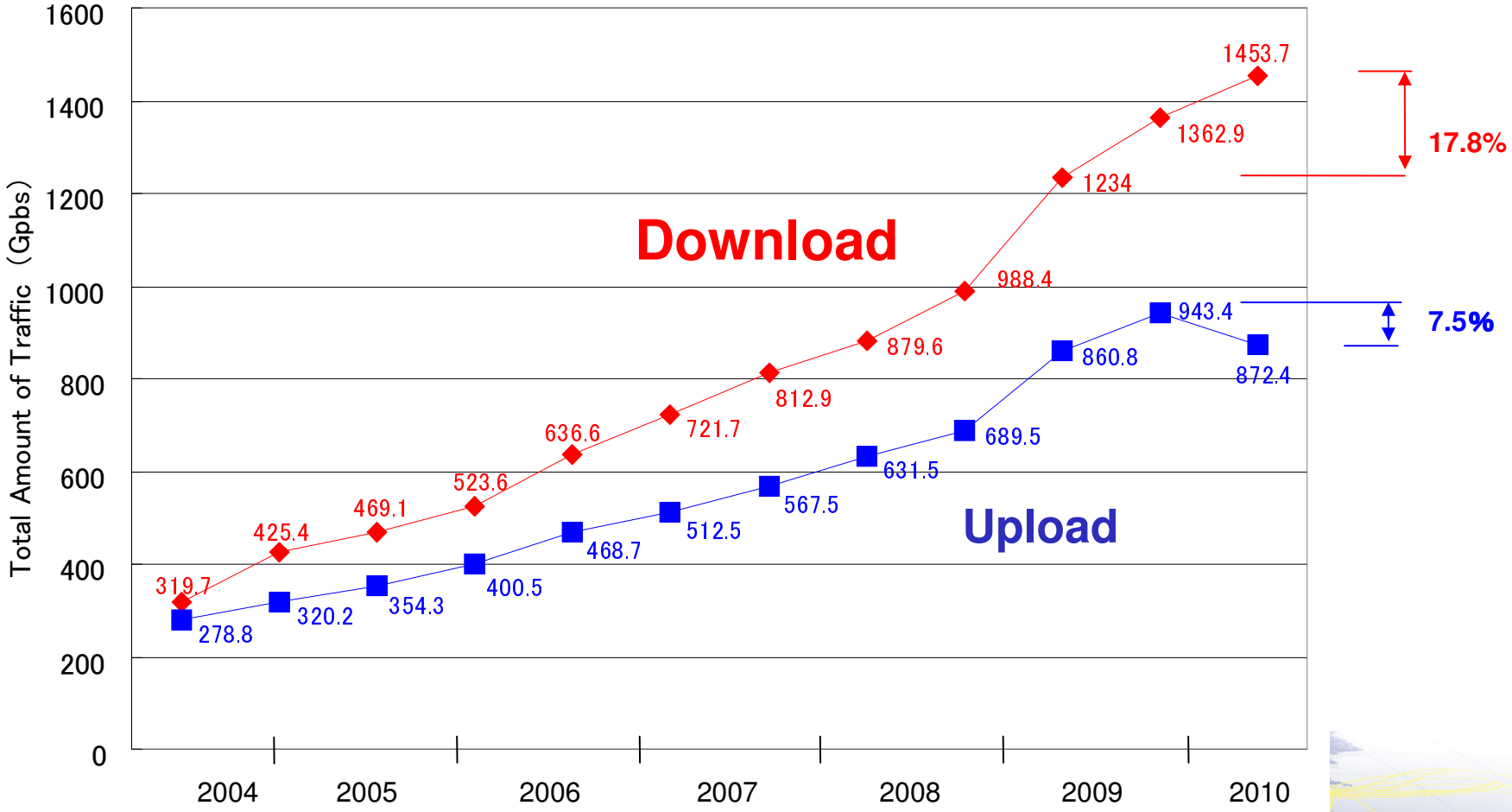
3. Current issues we are facing

4. Future visions

5. Wrap up

Internet Traffic Trend in Japan

- Total amount of Broadband Traffic is 1.46Tbps (Download)
  - 17.8% growth compared to last year
- Upload traffic decreased over the last half year (872Gbps)

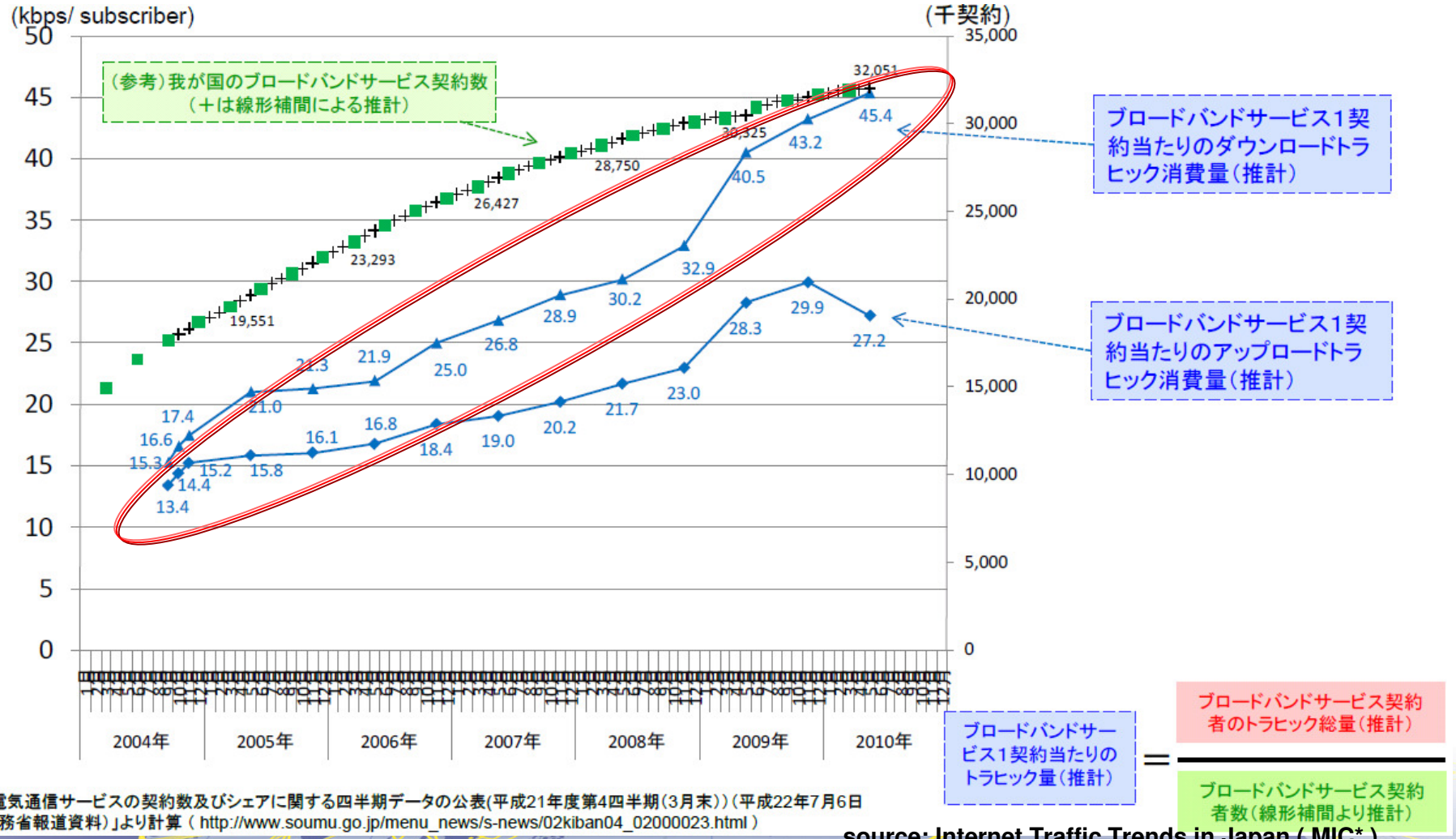


* **Average traffic**

source: Internet Traffic Trends in Japan (MIC*)
MIC: Ministry of Internal Affairs and Communications

Internet Traffic Trend in Japan (cont.)

- Traffic volume per subscriber growing
  - 16kbps (in 2004) -> 45kbps (in 2010)

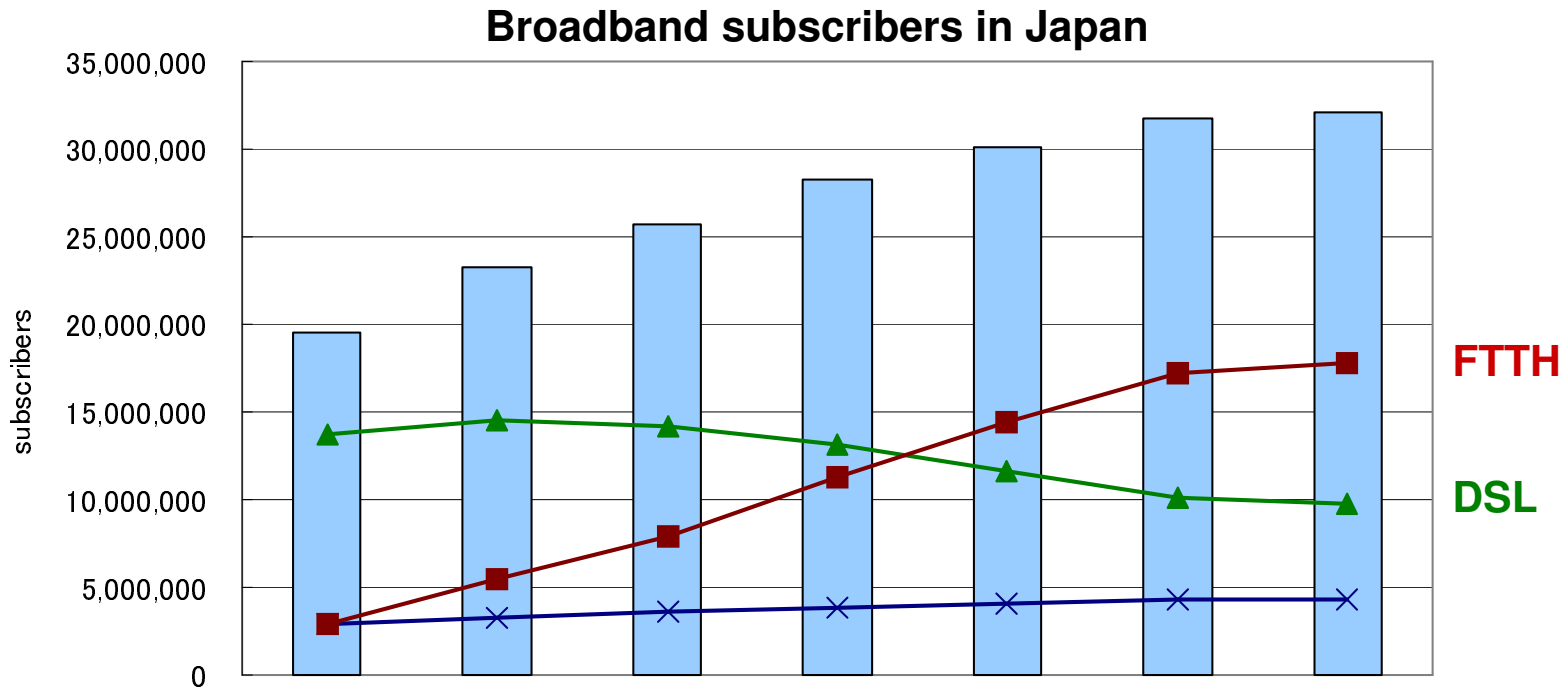


「電気通信サービスの契約数及びシェアに関する四半期データの公表(平成21年度第4四半期(3月末))(平成22年7月6日総務省報道資料)」より計算 (http://www.soumu.go.jp/menu_news/s-news/02kiban04_02000023.html)

source: Internet Traffic Trends in Japan (MIC*)
MIC: Ministry of Internal Affairs and Communications

Internet Traffic Trend in Japan (cont.)

- Growing Broadband subscribers
- Shifting from DSL (metal) to FTTH (optical fiber)



	2004	2005	2006	2007	2008	2009	2010
Total	19,557,100	23,301,100	25,744,769	28,303,003	30,115,989	31,709,116	32,040,792
DSL	13,675,800	14,517,800	14,235,925	13,133,113	11,601,734	10,134,491	9,735,140
CATV	2,959,710	3,309,480	3,565,427	3,827,417	4,083,072	4,300,594	4,352,878
FTTH	2,896,930	5,457,690	7,931,837	11,329,886	14,418,215	17,195,696	17,788,535

source: Ministry of information and Communications Statistic Database

Overview

- Internet traffic in Japan has been growing consistently

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

- How backbone we have been making?
- How bandwidth we have?

Agenda

1. Current situation of Internet traffic in Japan

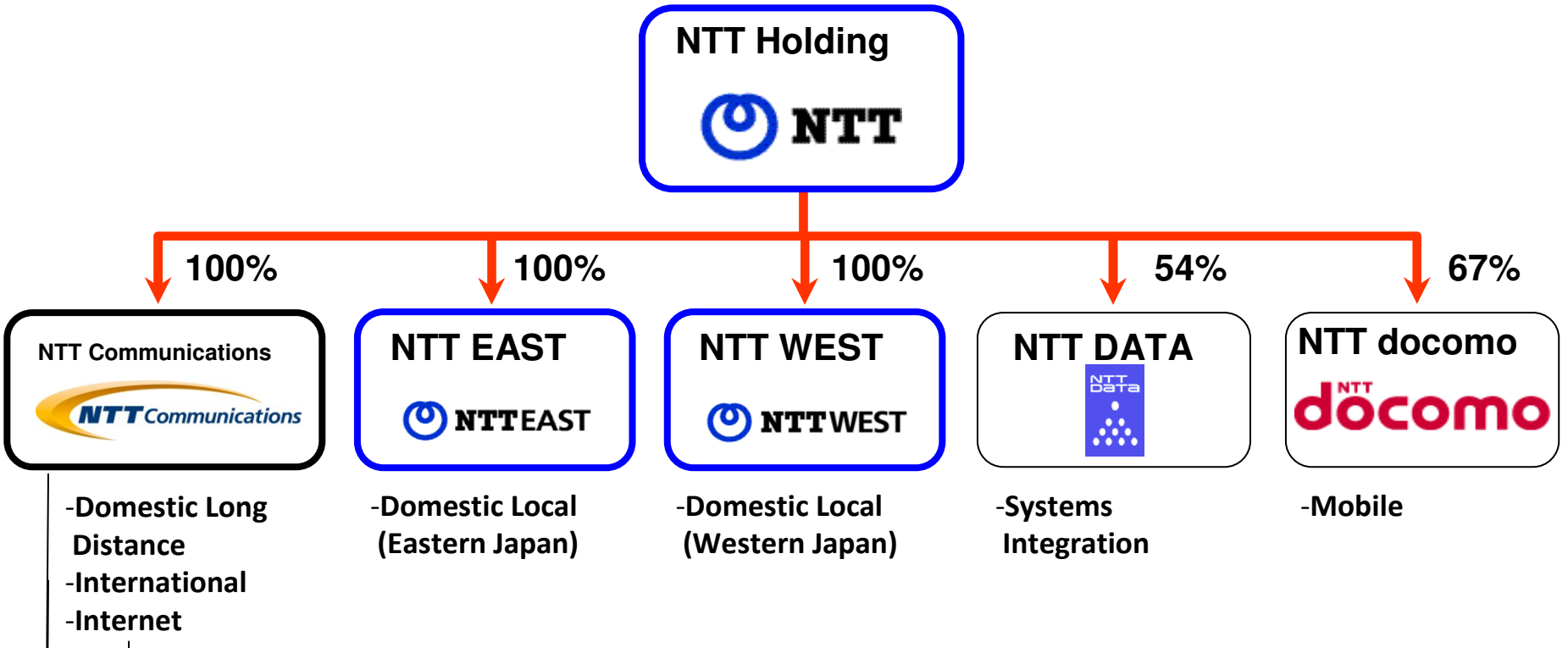
2. What is OCN?

3. Current issues we are facing

4. Future visions

5. Wrap up

Outline of NTT Group



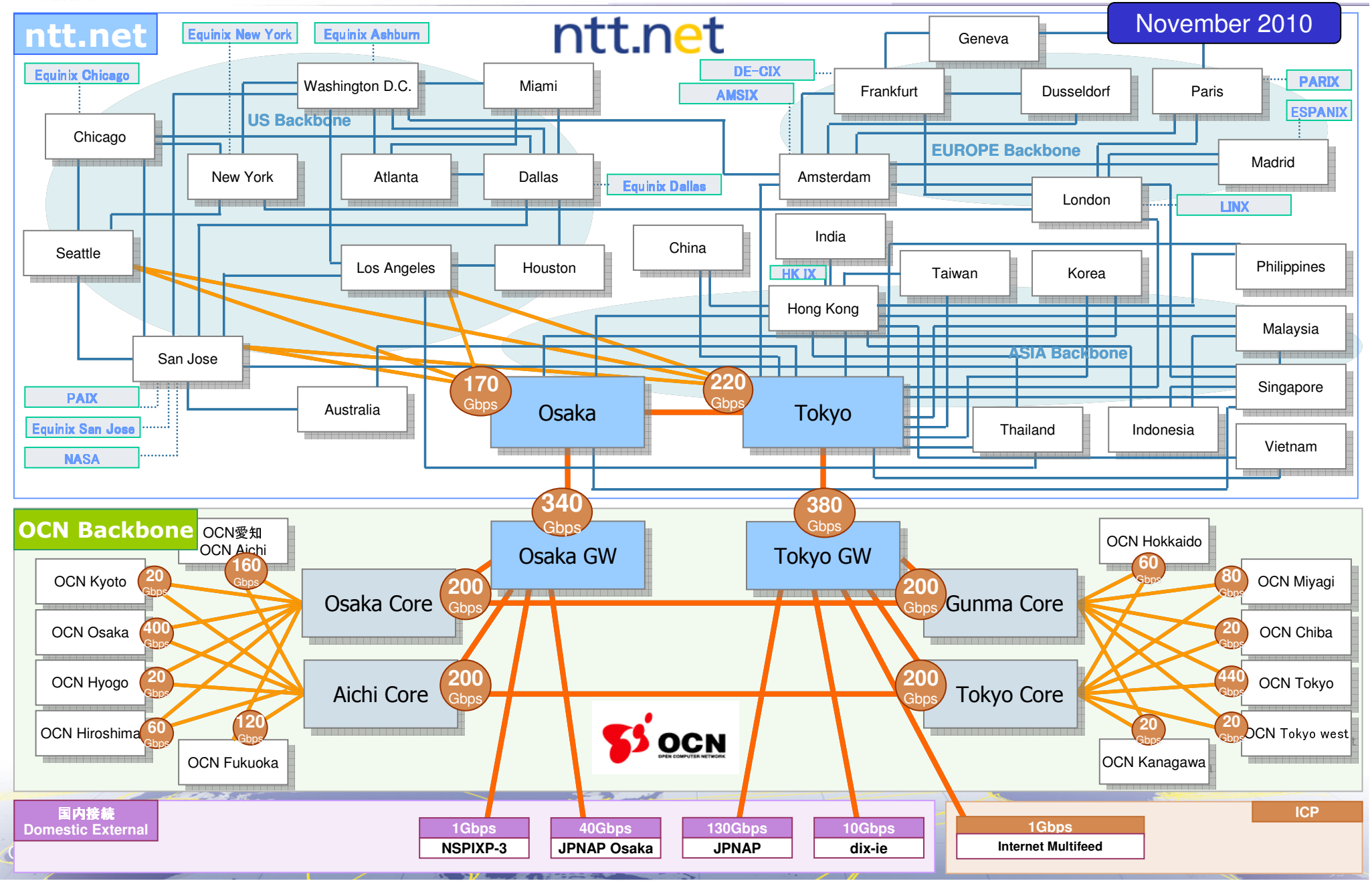
OCN : *Domestic Open IP Service Brand*

ntt.net : *International Open IP Service Brand*

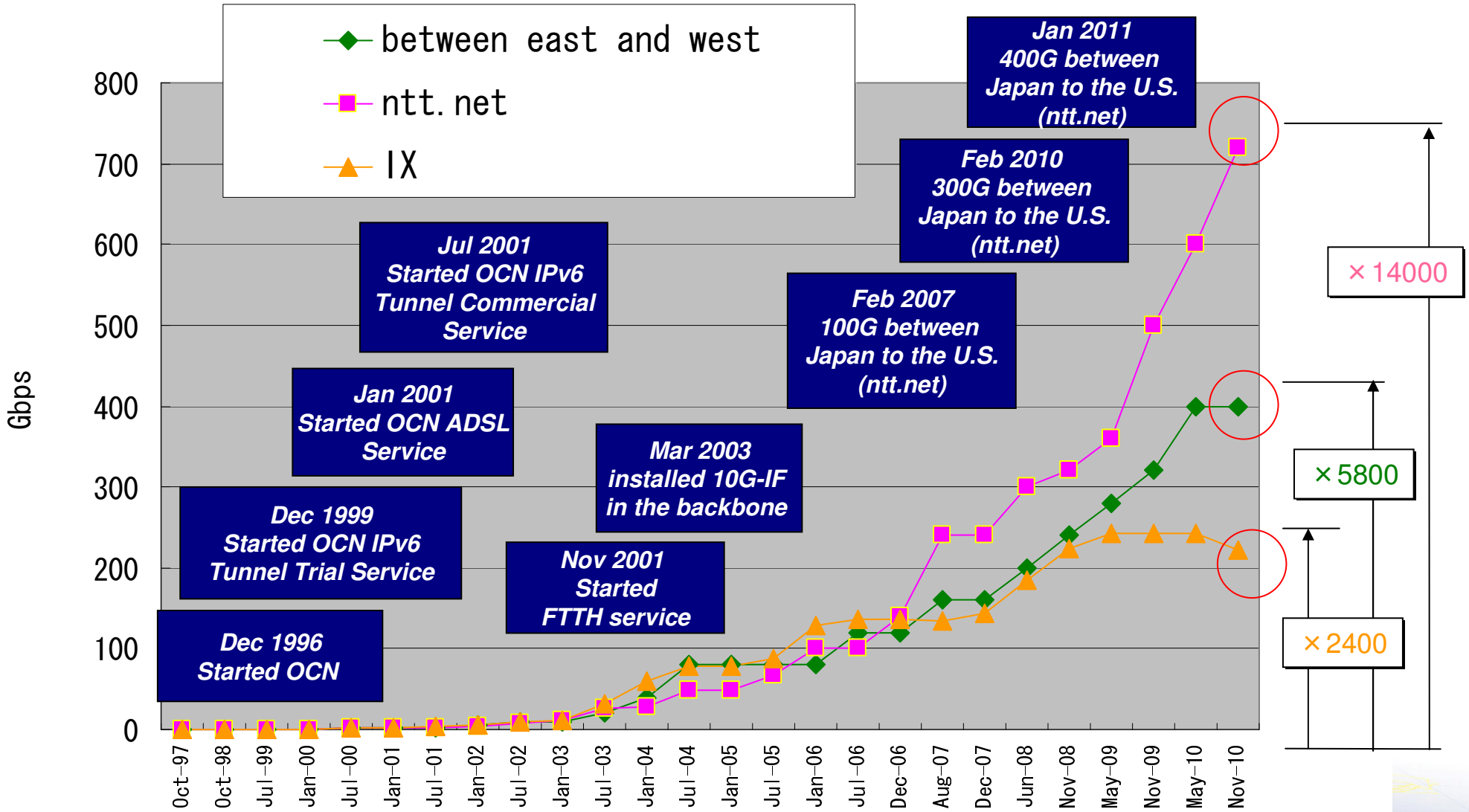
100% NTT Communications | NTT Com Asia — http://www.hk.ntt.com/en/index.html

100% HKNET — http://www.hknet.com/en/home/index.htm

NTT Communications' IP Backbone Network



Bandwidth history of OCN



What we are doing

- Make our network larger and larger as Internet traffic grows
- Issues we have been facing
- Efforts we have been making

Agenda

1. Current situation of Internet traffic in Japan
2. What is OCN?
3. **Current issues we are facing**
4. Future visions
5. Wrap up

The issues we are facing

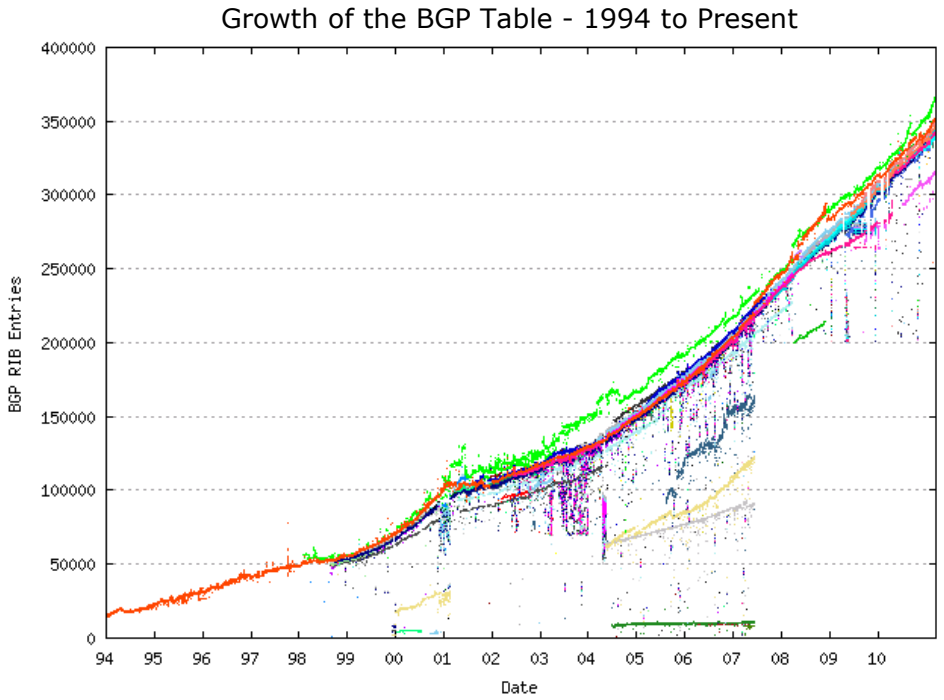
1. Scalability of Router Forwarding Tables

2. Link Aggregation

FIB table of OCN

- FIB(Forwarding Information Base) table has been growing

- Causes of growing FIB
  1. BGP full routes (more than 340,000 in February 2011)
  2. Prefixes with no-export
  3. ECMP, {i, e} bgp-multipath

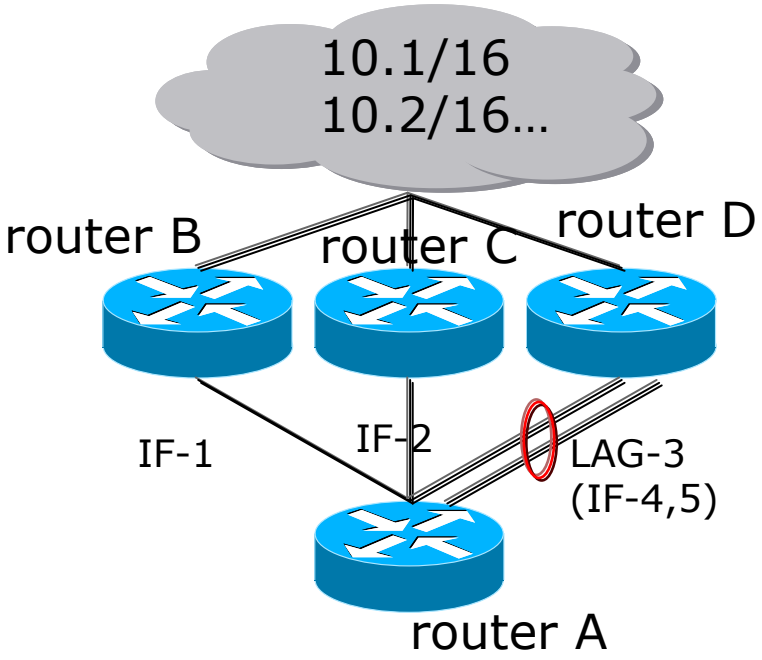


source: http://bgp.potaroo.net/

Scalability of Router Forwarding Tables

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

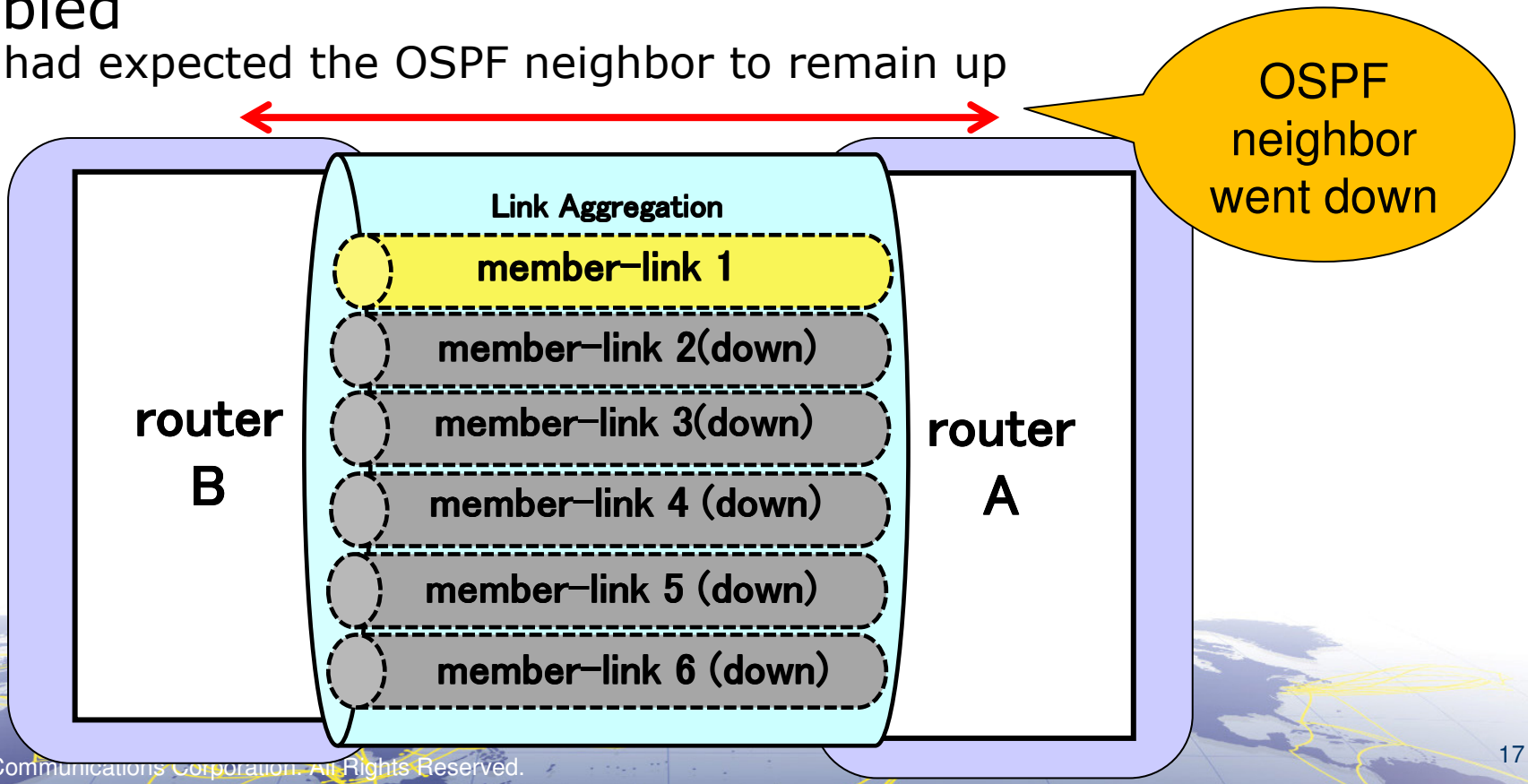
FIB of router-A	
prefix	output interface(s)
10.1.0.0/16	IF-1
	IF-2
	LAG-3(IF-4, 5)
10.2.0.0/16	IF-1
	IF-2
	LAG-3(IF-4, 5)



- It took a lot of time to converge the routes

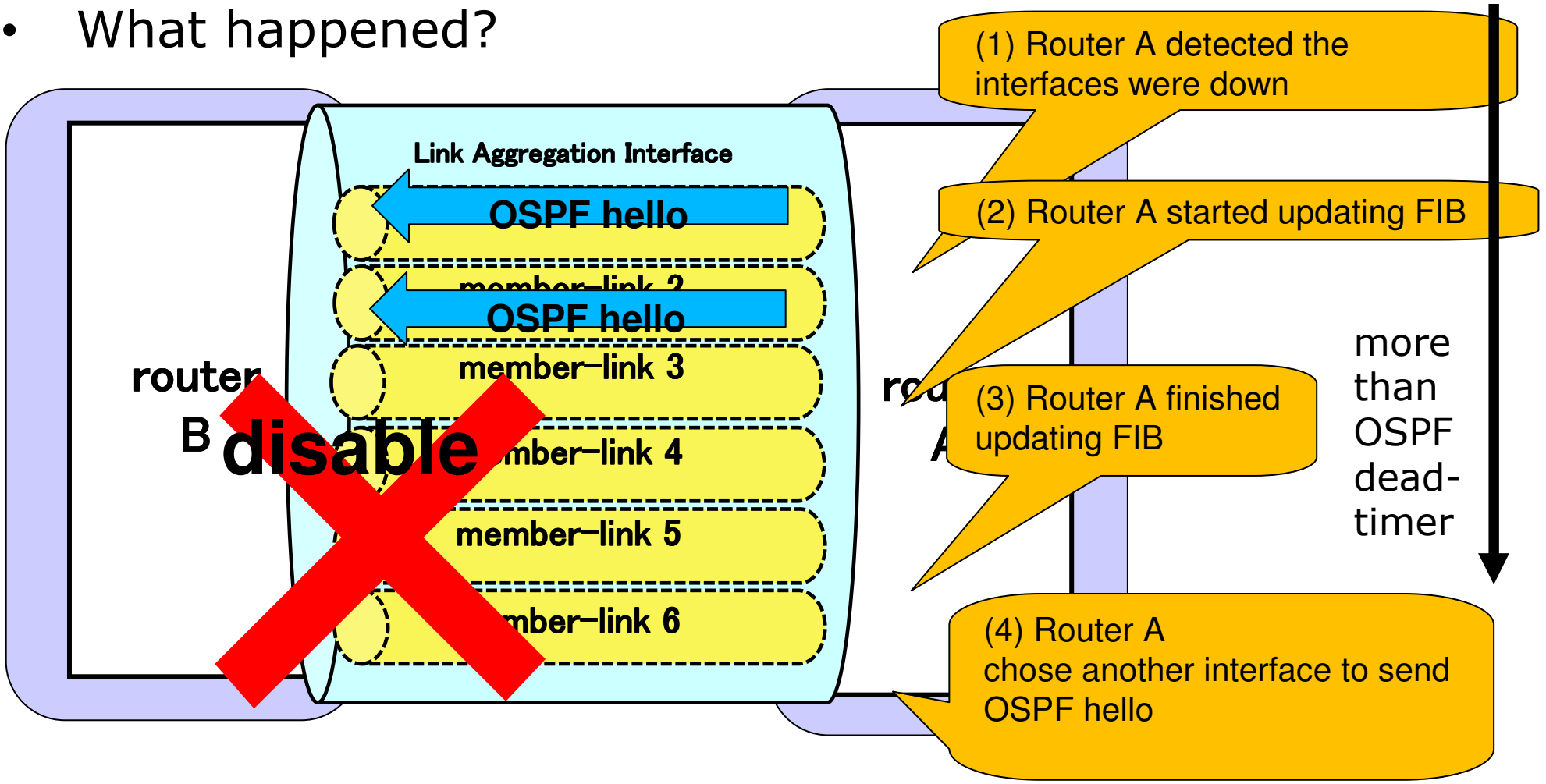
Scalability of Router Forwarding Tables

- 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 to make 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

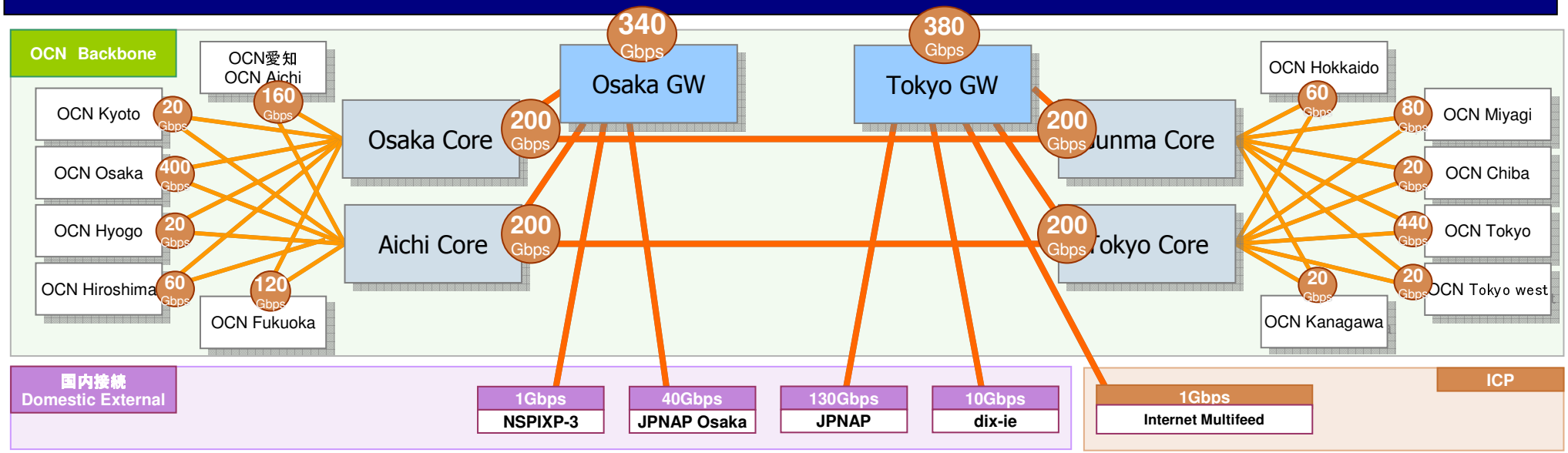
  For more information: BGP Convergence in much less than a second
  http://www.nanog.org/meetings/nanog40/presentations/ClarenceFilsfils-BGP.pdf

- Fewer routes to be updated

- Improving the route convergence time

Link Aggregation Issues

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

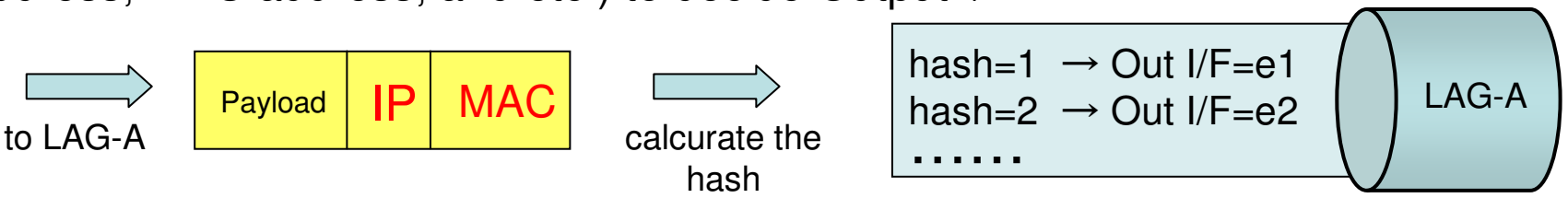


- Traffic balance issues (Traffic Polarization)
- Operation issues
- Other issues

LAG Issues

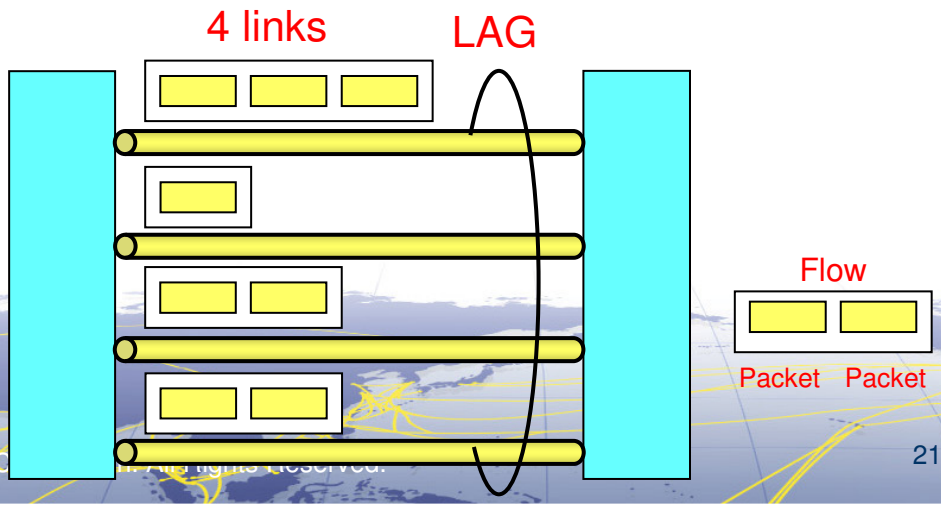
~traffic balance issues (1/3)~

- Traffic balance in the LAG（1）
  - ***Can't use per-packet round-robin***
    - Simple round-robin bring about packet reordering in a flow
  - Hashing algorithm: calculate the hash value based on the packet information (IP address, MAC address, and etc.) to decide Output I/F



  - Traffic are distributed per flow using the hash values

➢ Issue 1: traffic-unbalance by variation of flow



LAG Issues

~traffic balance issues (2/3)~

- Traffic balance in the LAG（2）
  - Issue 2: The less # of hash elements, the worse traffic-balanced
    - ➢as a result, less effective use of bandwidth

e.g.: 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
2:2:2:1:1 10+10+10+10*1/2+10*1/2=40	2:2:2:2 10+10+10+10=40	3:3:2 10+10+10*2/3=26.7

Traffic cannot be evenly distributed due to the hash mechanism

←Traffic balance ratio
←Effective bandwidth in the LAG

Only use 40G / 50G

Only use 27G / 30G

LAG Issues

~traffic balance issues (2/3)~

cf. Difference in traffic 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**

A large number of hash elements is better

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

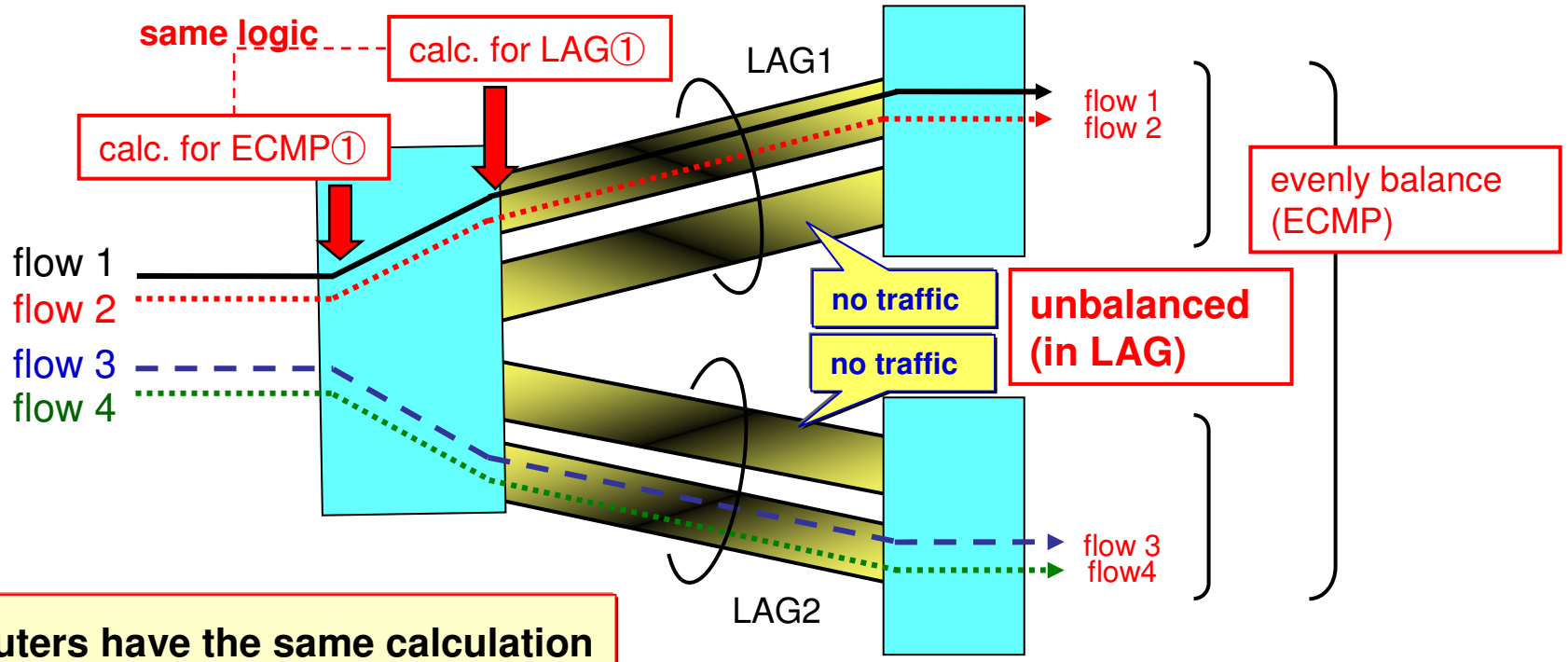
5 links LAG	4 links LAG	3 links LAG
IF#1 H1, H6, ･･･H26, H31 IF#2 H2, H7, ･･･H27, H32 IF#3 H3, H8, ･･･H28 IF#4 H4, H9, ･･･H29 IF#5 H5, H10, ･･･H30	IF#1 H1, H5, ･･･H29 IF#2 H2, H6, ･･･H30 IF#3 H3, H7, ･･･H31 IF#4 H4, H8, ･･･H32	IF#1 H1, H4, ･･･H28, H31 IF#2 H2, H5, ･･･H29, H32 IF#3 H3, H6, ･･･H30
7:7:6:6:6 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**

←Traffic balance ratio
←Effective bandwidth in the LAG

LAG Issues

～traffic balance issues (3/3 - 1)～

- Traffic balance by ECMP (Equal Cost Multi Path) and LAG: Case1
  - If calculation logic of LAG is the same as ECMP's, it will bring about unbalanced traffic in physical links

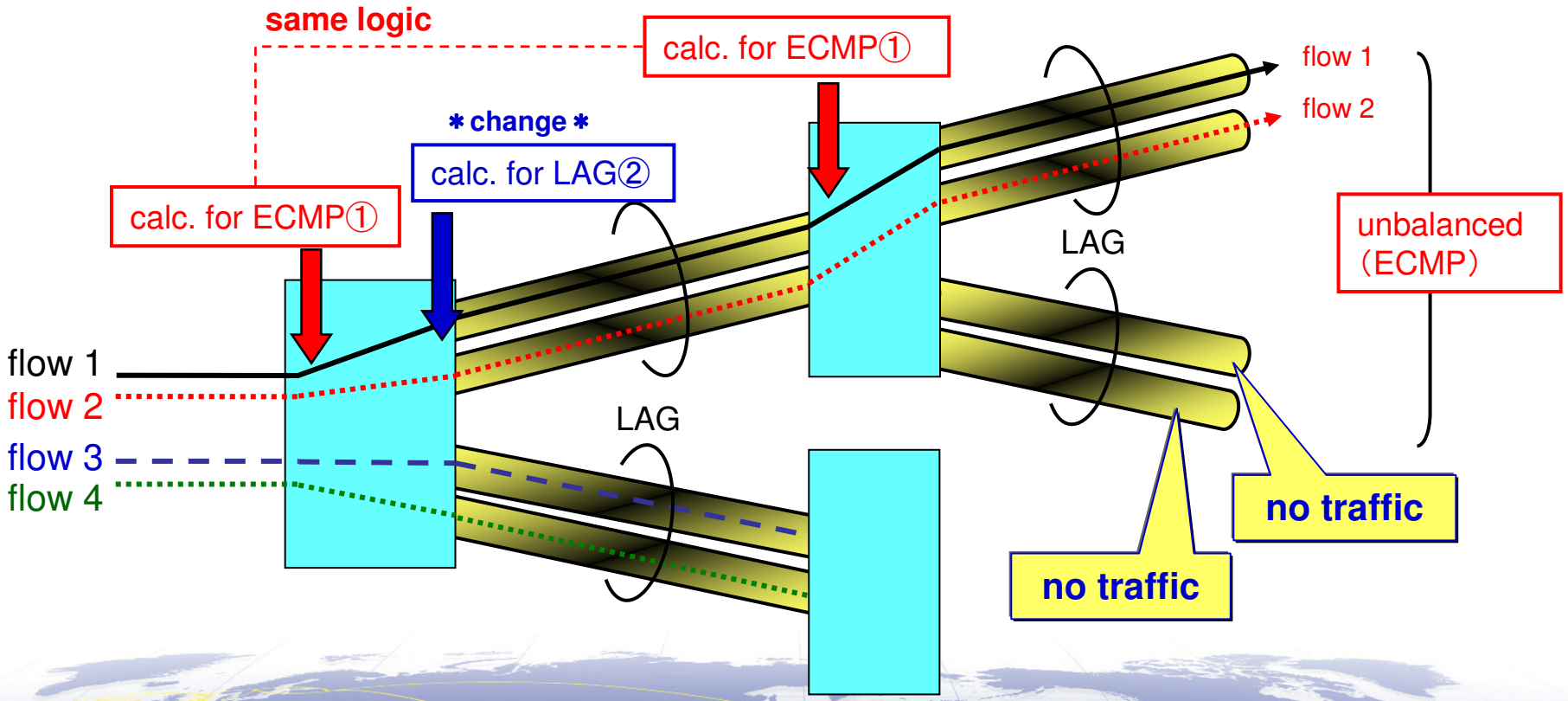


Some routers have the same calculation logics for ECMP and LAG as a default

LAG Issues

～traffic balance issues (3/3 - 2)～

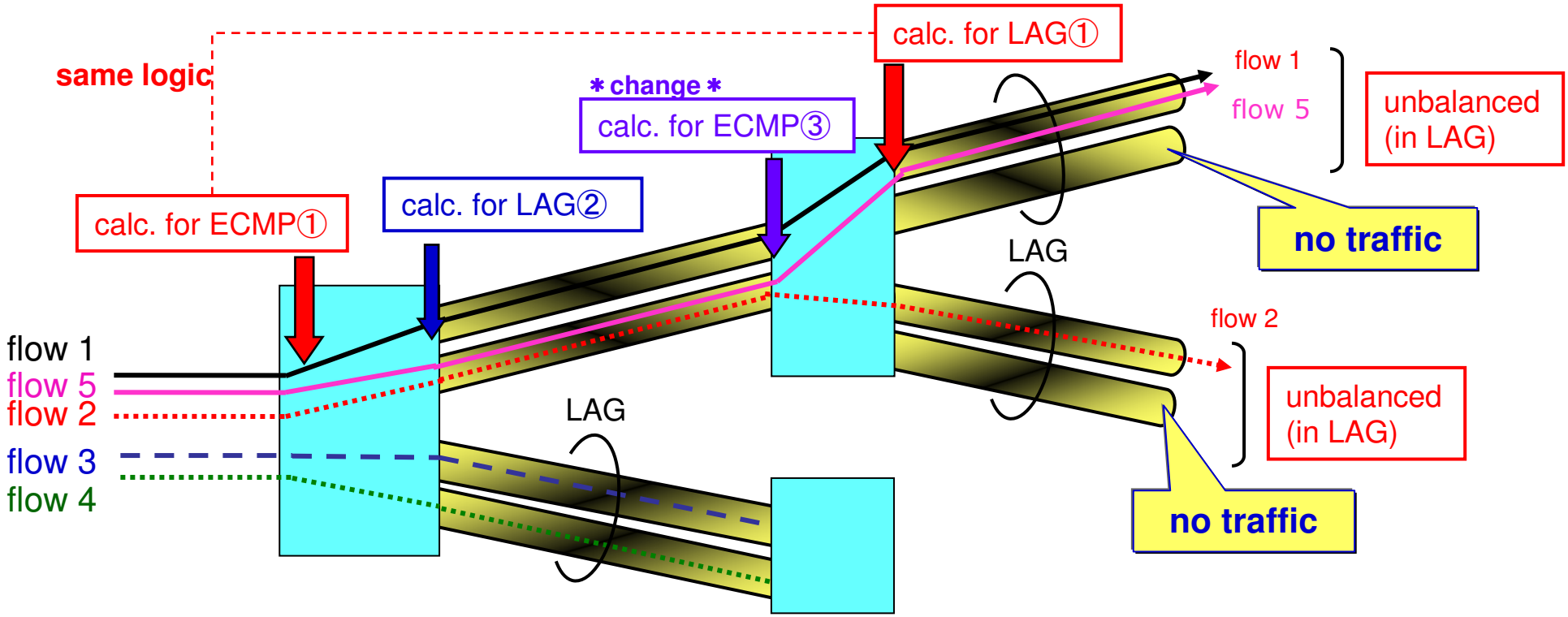
- Traffic balance by ECMP and LAG : Case2
  - If calculation logic of ECMP is the same as that of previous ECMP, it will bring about unbalanced traffic



LAG Issues

～traffic balance issues (3/3 - 3)～

- Traffic balance by ECMP and LAG : Case3
  - If calculation logic of LAG is the same as that of ECMP at the previous 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

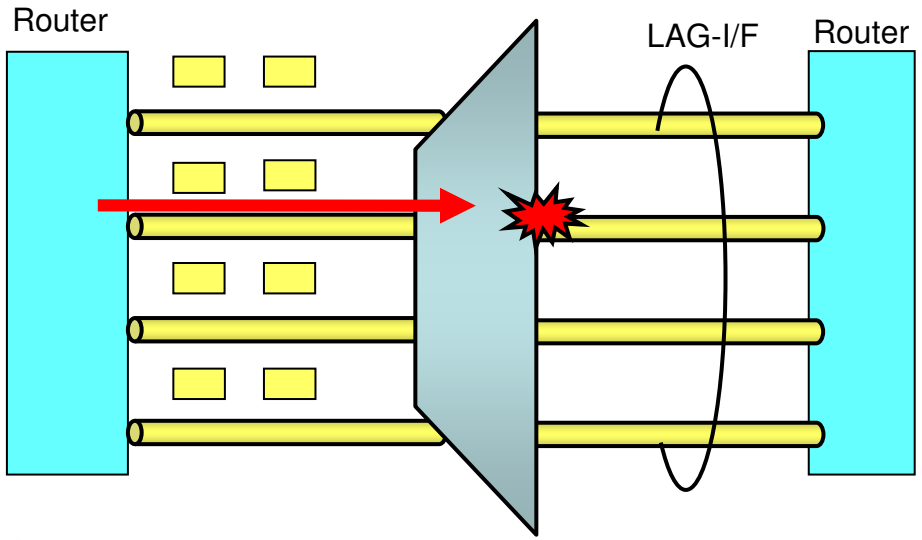
Need to consider balance logics, network topology, configurations

LAG Issues

～operation issues (1/3)～

- LAG operation (1)
  - In the case of silent-failure, traffic through the fault link will drop

  - LACP (Link Aggregation Control Protocol)
    - Sending and receiving control frames in physical links
    - Attention to Interoperability
  - BFD Per Member Link
    (Bidirectional Forwarding Detection)

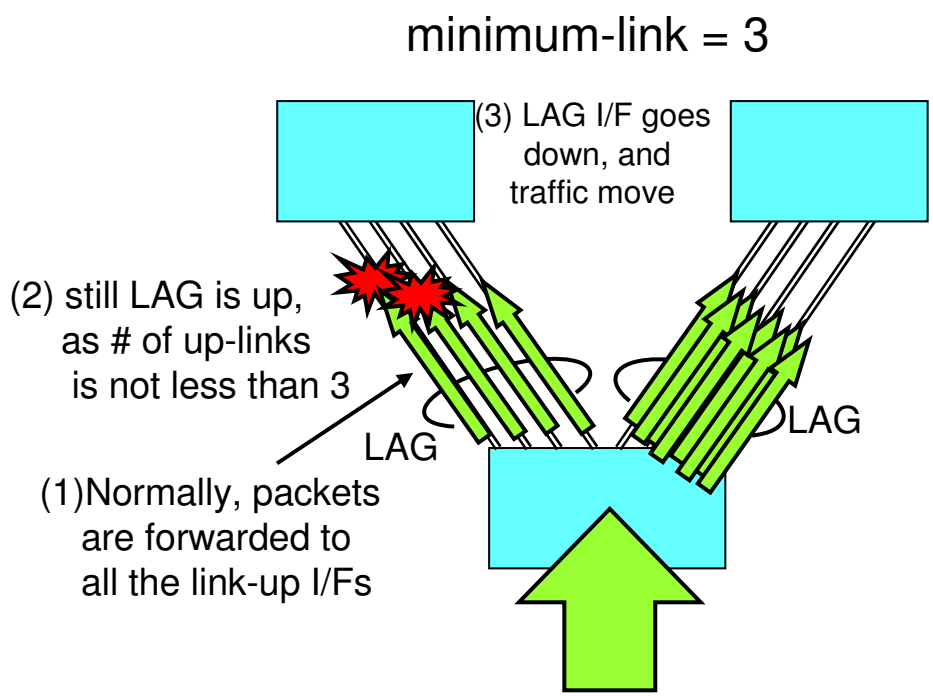


LAG Issues

～operation issues (2/3)～

- LAG operation (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
    - consider the entire network topology



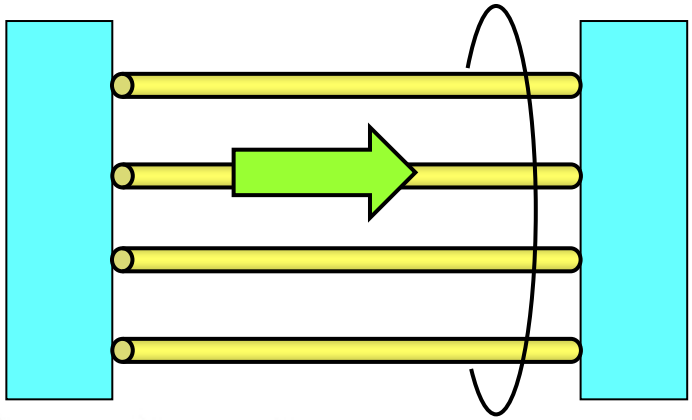
e.g.: minimum-link when the policy is 70% in LAG

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

LAG Issues

～operation issues (3/3)～

- LAG operation (3)
  - Ping for test
    - Packet goes through only one physical interface
    - Need to test each interface with letting the rest go down
    - expect Ethernet OAM

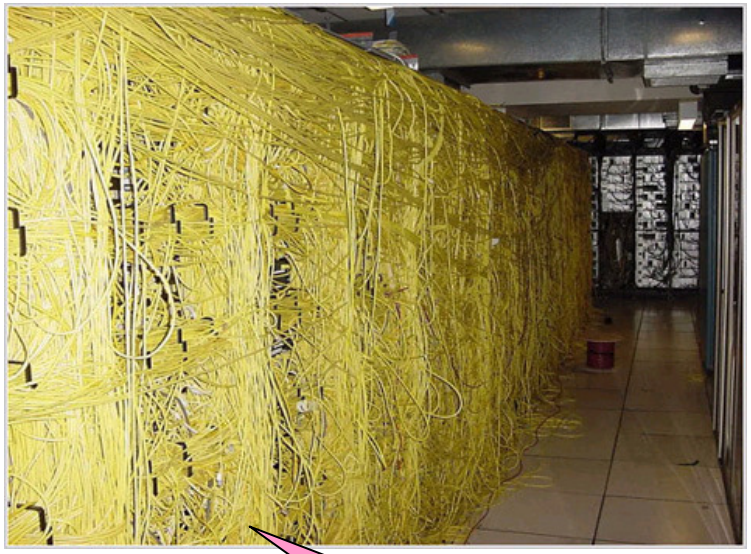


LAG Issues

～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
    - based on redundant policy
    - MTBF for each part
    - Cost
  - 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



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 visions**
5. Wrap up

How to handle traffic growth

(1) Change Network	• new routers, new switches • new router-interface (100GE)
(2) Control Traffic	• Cache Servers • CDN

Expectation for 100GE

- Need 100GE I/F
  - Bandwidth over 1Tbps
  - LAG is troublesome
- Request
  - Lower price
    - CFP is expensive
    - LR10
  - Support 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
  - Interoperability, 100GE LAG, Ether OAM
  - Next step: 400GE, 1T Ether

Our preparation for 100GE

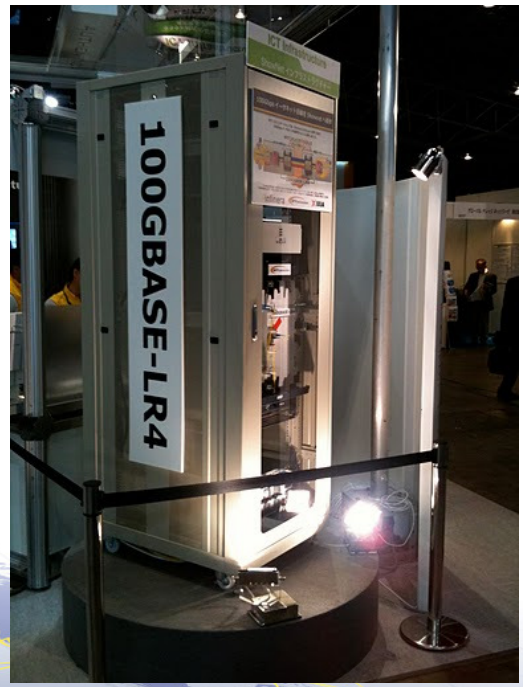
The Best award of Interop Tokyo for 2 years in a row

2009 100GE-SR10 demonstrated with transmission equipment and traffic generator

2010 100GE transmission network (100GE-LR4) was provided for practical operation

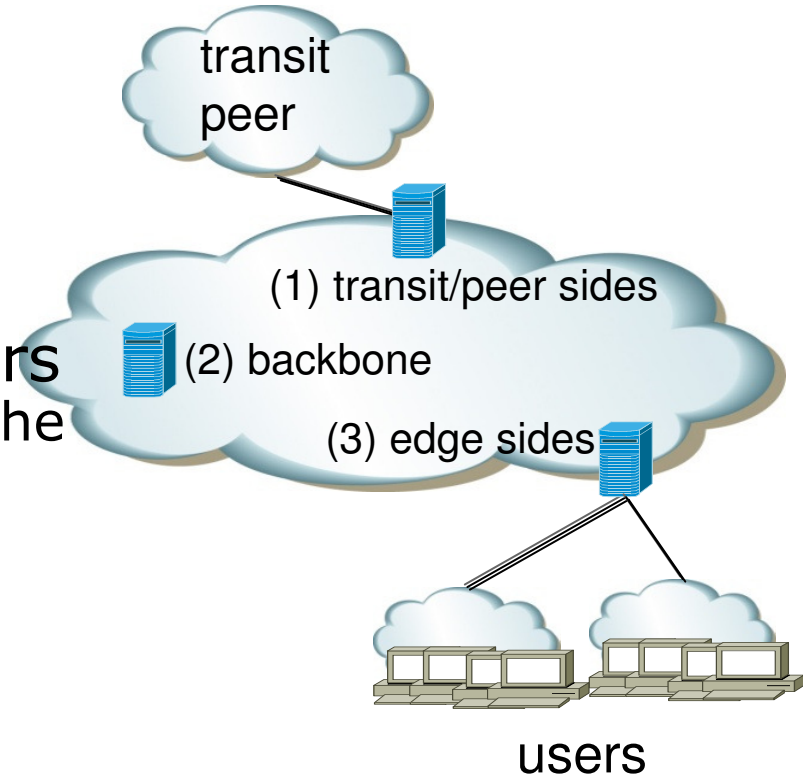
2010/6/8 News Release

NTT Com, Infinera and Ixia to Provide World's First Practical 100 Gbps Ethernet Interconnection at Interop Tokyo 2010



Cache servers

- Legal changes in January 2010 in Japan
  - became legal to install cache servers without permission(s)
- Demerits not to install cache servers
  - Streaming delays because of carrying the traffic from peer/transit network
  - more bandwidth
  - costs for transit network
- Merits to install cache servers
  - Transit cost saving,
  - Bandwidth saving
  - Fixing delay
- Issues
  1. Equipment performance (cache hit ratio, lack of bandwidth…)
  2. Where to place
  3. When equipment failure



Agenda

1. Current situation of Internet traffic in Japan
2. What is OCN?
3. Current issues we are facing
4. Future visions
5. **Wrap up**

Wrap up

- The total traffic in Japan has been consistently increasing.
  - The traffic will keep growing in the future.
- We are continuing to design a strong backbone network.
  - But we have some designing/operational issues
- We are going to need 100GE in the near future to deal with the situation.
- How is your network? Do you have any ideas or suggestions to cope with the expected growth of traffic in the future?

OCN
OPEN COMPUTER NETWORK

Thank you!

NTT Communications