Operational routing experience in NTT/OCN

Routing-SIG @ APNIC19

NTT Communications / OCN

Tomoya Yoshida <yoshida@ocn.ad.jp>

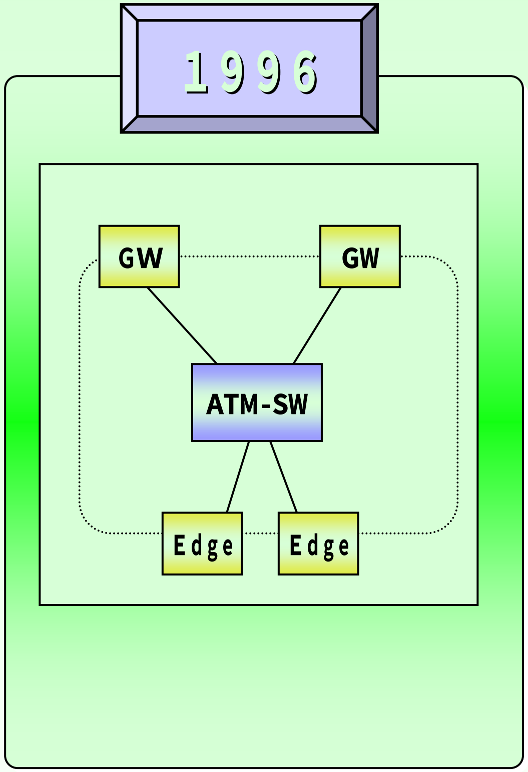
Our History from “OCN Economy”

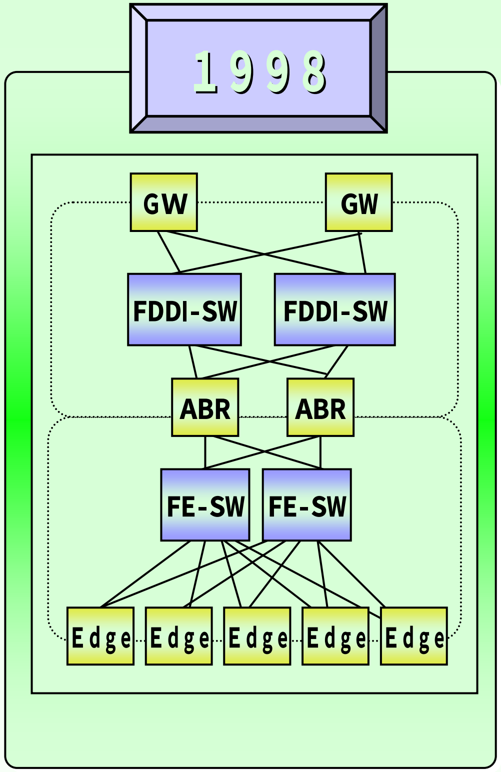
- We started “OCN Economy” Service in 1996
  - This is the epoch-making service
    - » The Price was very cheap at that time : ¥38,000 128Kbps
  - We distribute|assign /28 or /29 to users
- /28 or /29 is redistributed to OSPF by external route
  - Static route information on the edge router is redistributed to OSPF
- Many OSPF external routes is growing

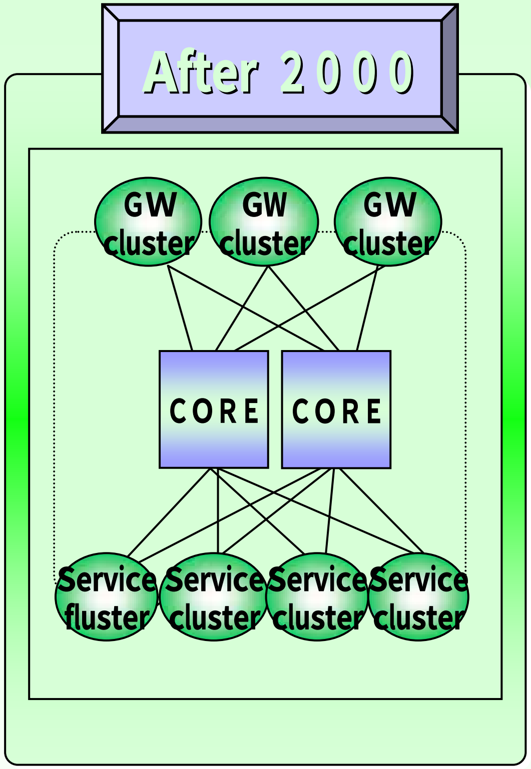
Our History Cont.

- When the OSPF external route reached around 20,000, OSPF convergence time needed more and more
  - We tried to separate OSPF domain
    - Operation would be complicated
    - Extension would be difficult
  - We changed from OSPF to BGP around 2000
  - iBGP route is growing and growing very fast
  - Then we use route reflector hierarchy
- Address problem
  - We could not get enough address to assign at once
  - As the result it was difficult to aggregate the route

The changes of OCN Backbone Topology





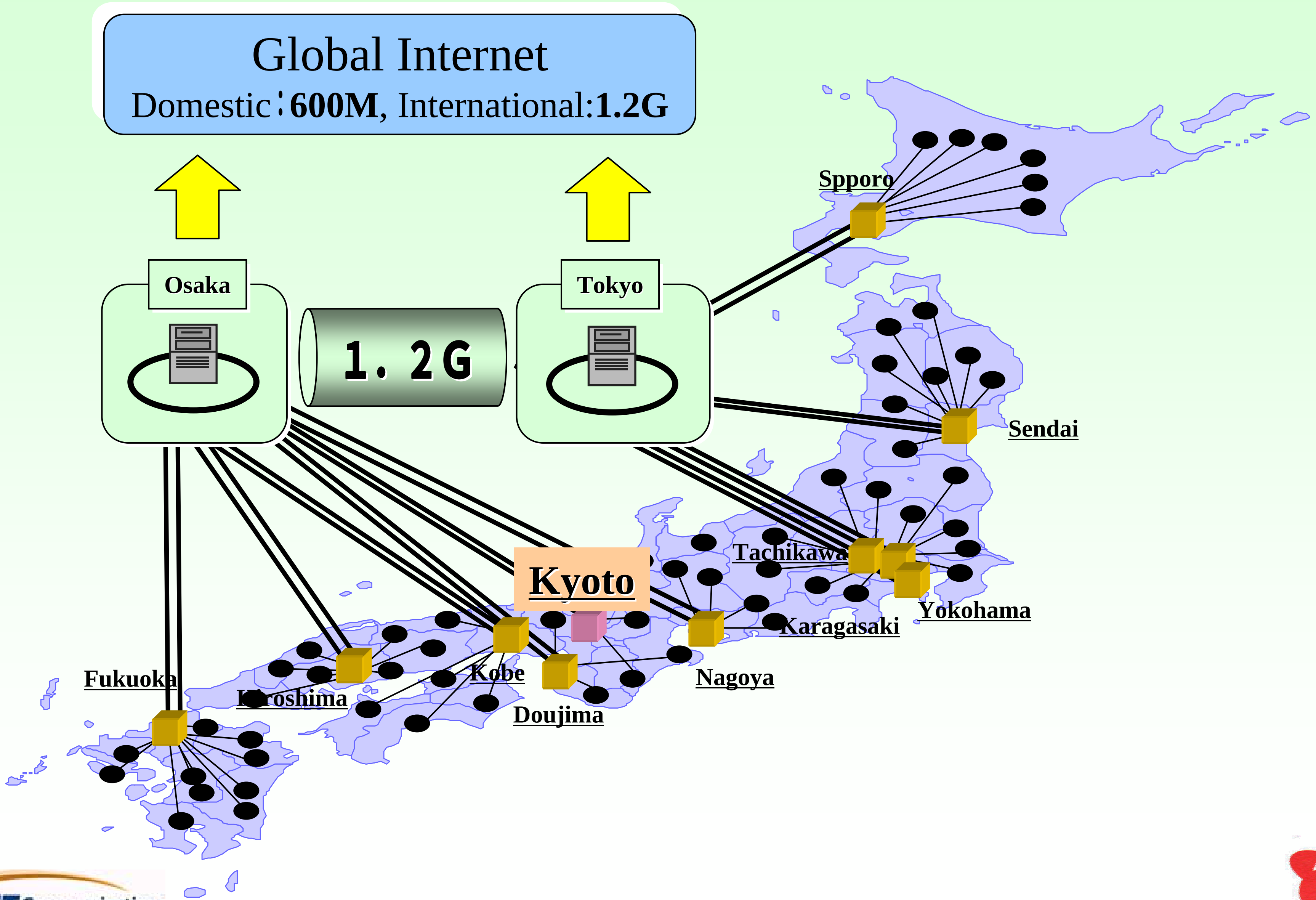


- full-mesh topology
- using ATM-SW etc

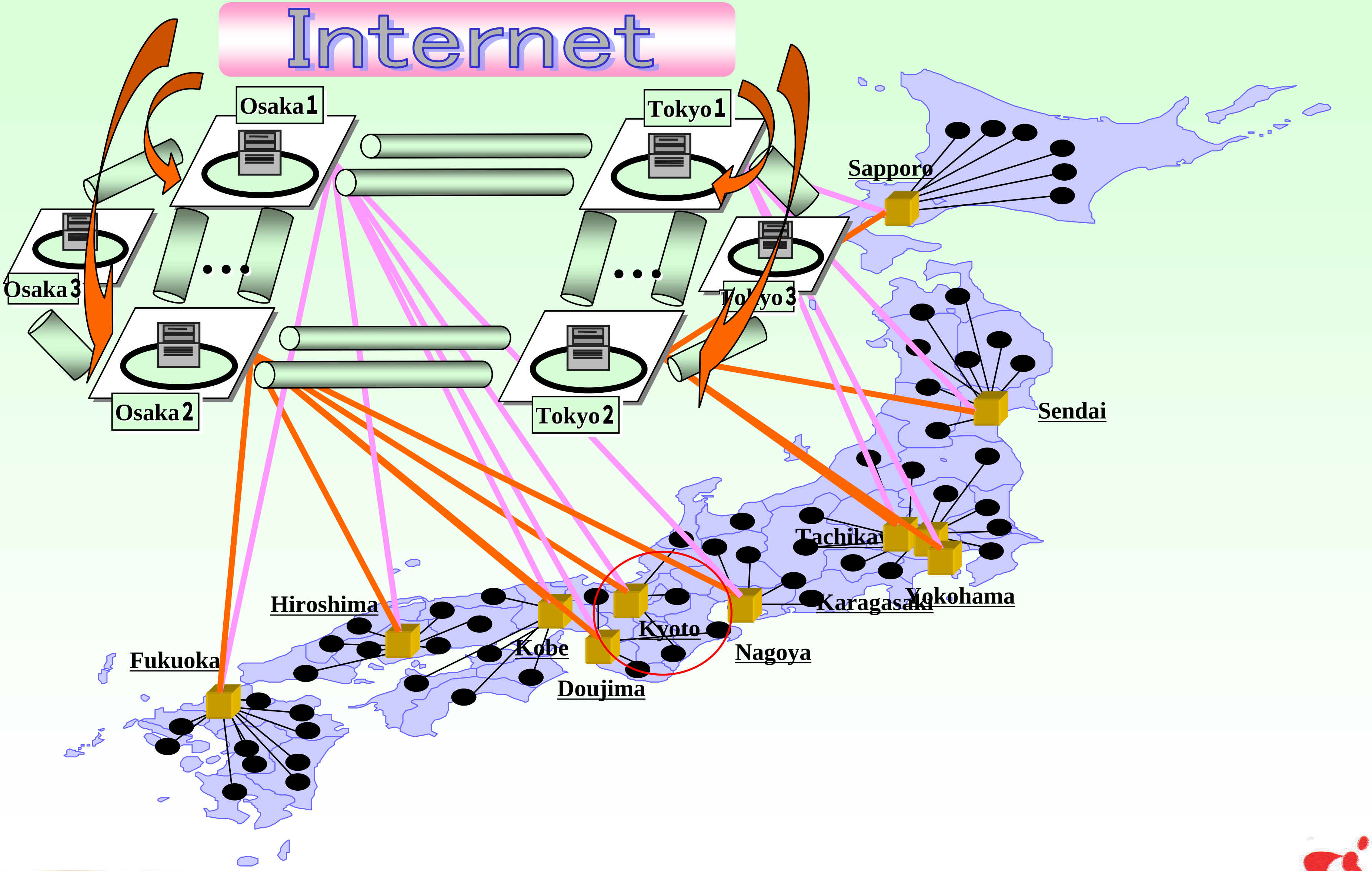
- divided OSPF area
- FDDI、FE-SW
- Reduction of routing

- Clustering topology according to the service or routing

Backbone Topology in 1999

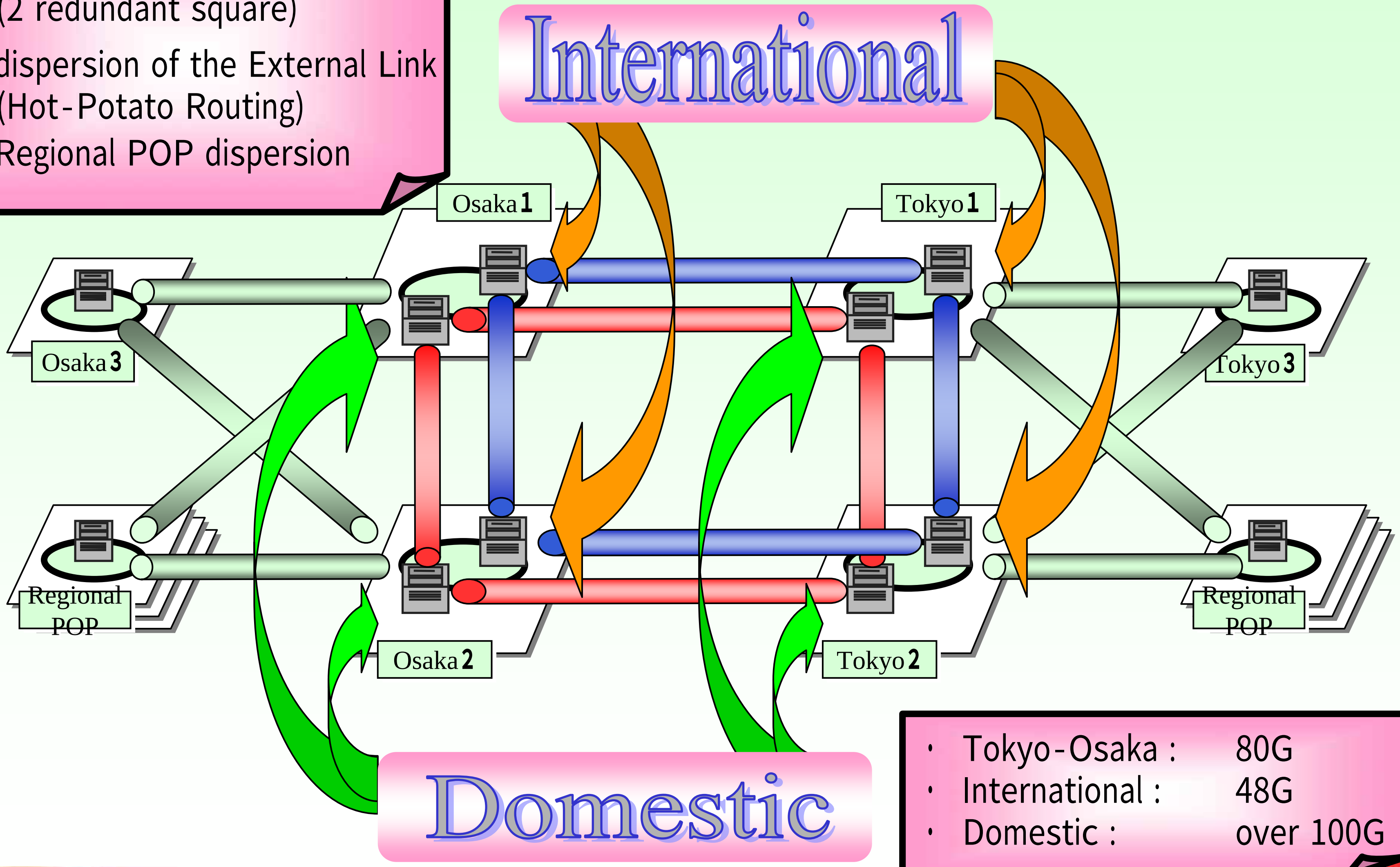


Current OCN Backbone Topology



Square Backbone

- Square Topology (2 redundant square)
- dispersion of the External Link (Hot-Potato Routing)
- Regional POP dispersion



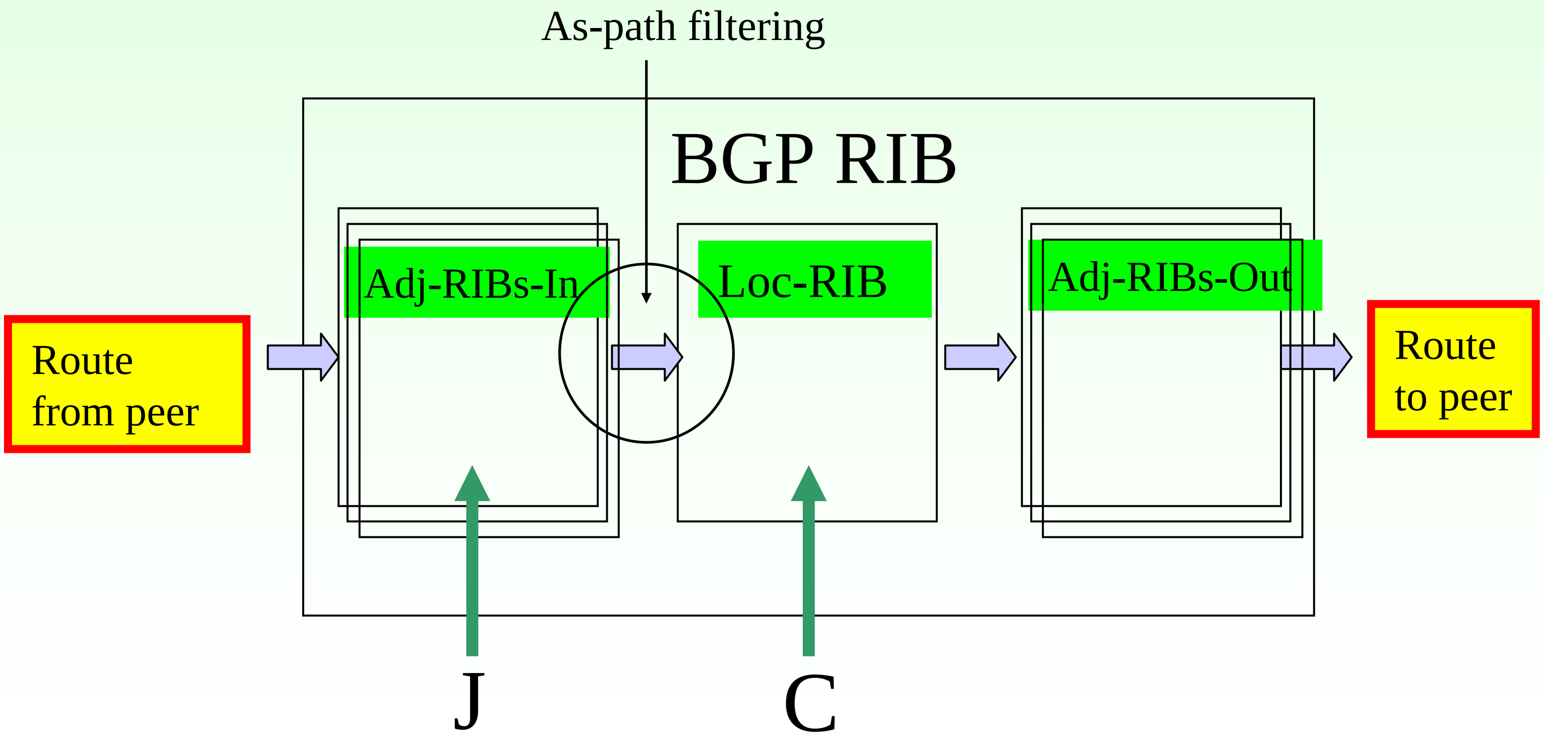
Tokyo-Osaka :	80G
International :	48G
Domestic :	over 100G

Routing (OSPF/BGP)

- OSPF:IGP
  - Backbone area and many other areas : normal design
  - Cost design is basically equal cost load balancing
  - Distribution the function of DR/BDR in the same router for more than two segment
  - Restriction of the number of router in the same area
- BGP:EGP
  - Route reflector hierarchy topology
  - Distribute for needed cluster

BGP prefix limitation experience

- Both Cisco and Juniper have a limitation mechanism of the BGP route from peer
- But those implementation are different



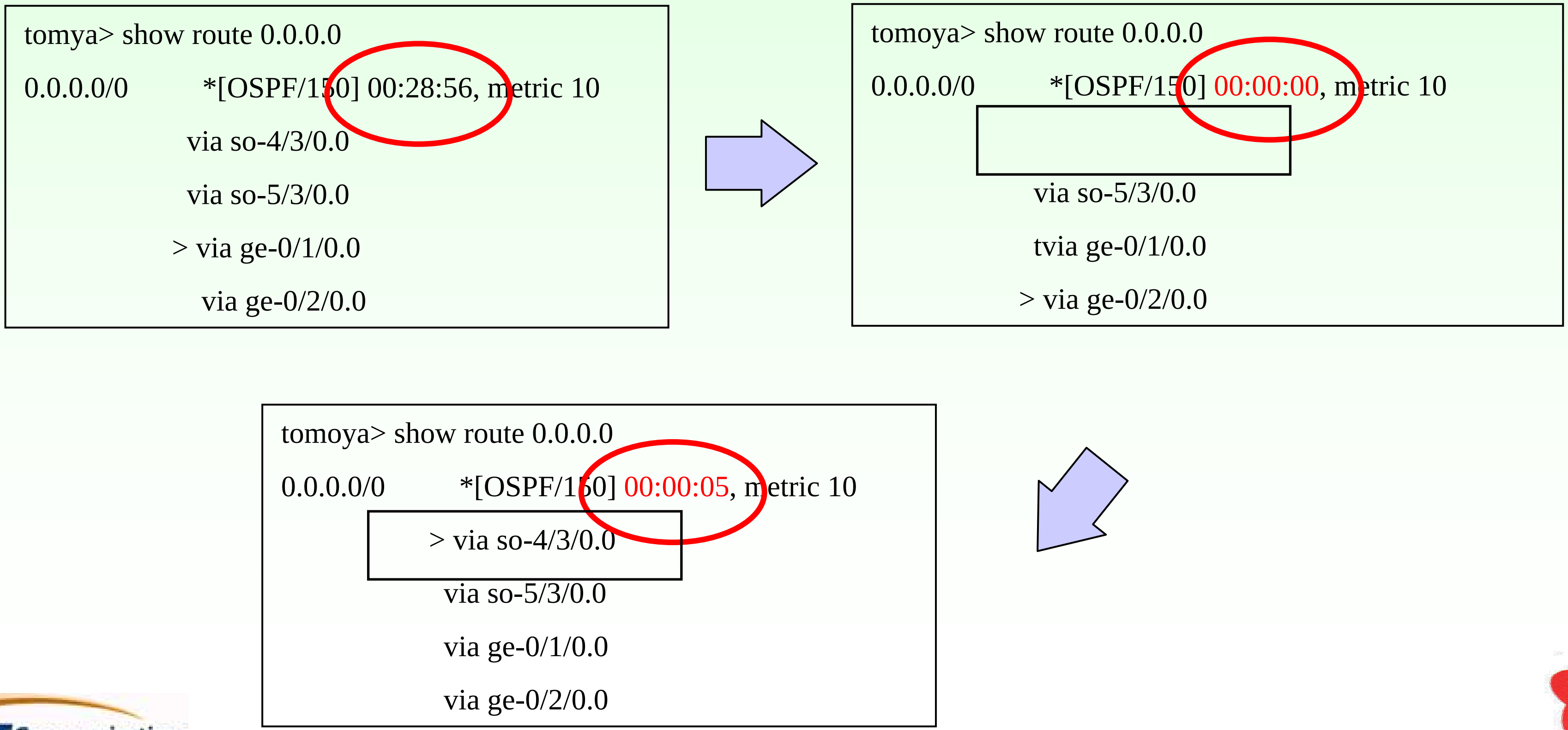
Next-hop self / redistribution

- If you forget next-hop-self at the eXchange border route and not redistributed to your backbone the IX segment around /24

- In Japan, 3 major IXs is announcing around /20 the part of the IX’s segment IP like /24, so when some ISP forget the next-hop-self and not redistribute those segment to IGP, traffic will go to the IX’s AS (dix-ie, JPIX, JPNAP’s AS)

LSA refresh experience

- Some LSA was flapping
  - Default refresh timer is different
    - » Cisco is 30 minutes, Juniper is 50 minutes



```
tomya> show route 0.0.0.0
0.0.0.0/0          *[OSPF/150] 00:28:56, metric 10
                     via so-4/3/0.0
                     via so-5/3/0.0
                   > via ge-0/1/0.0
                     via ge-0/2/0.0
```

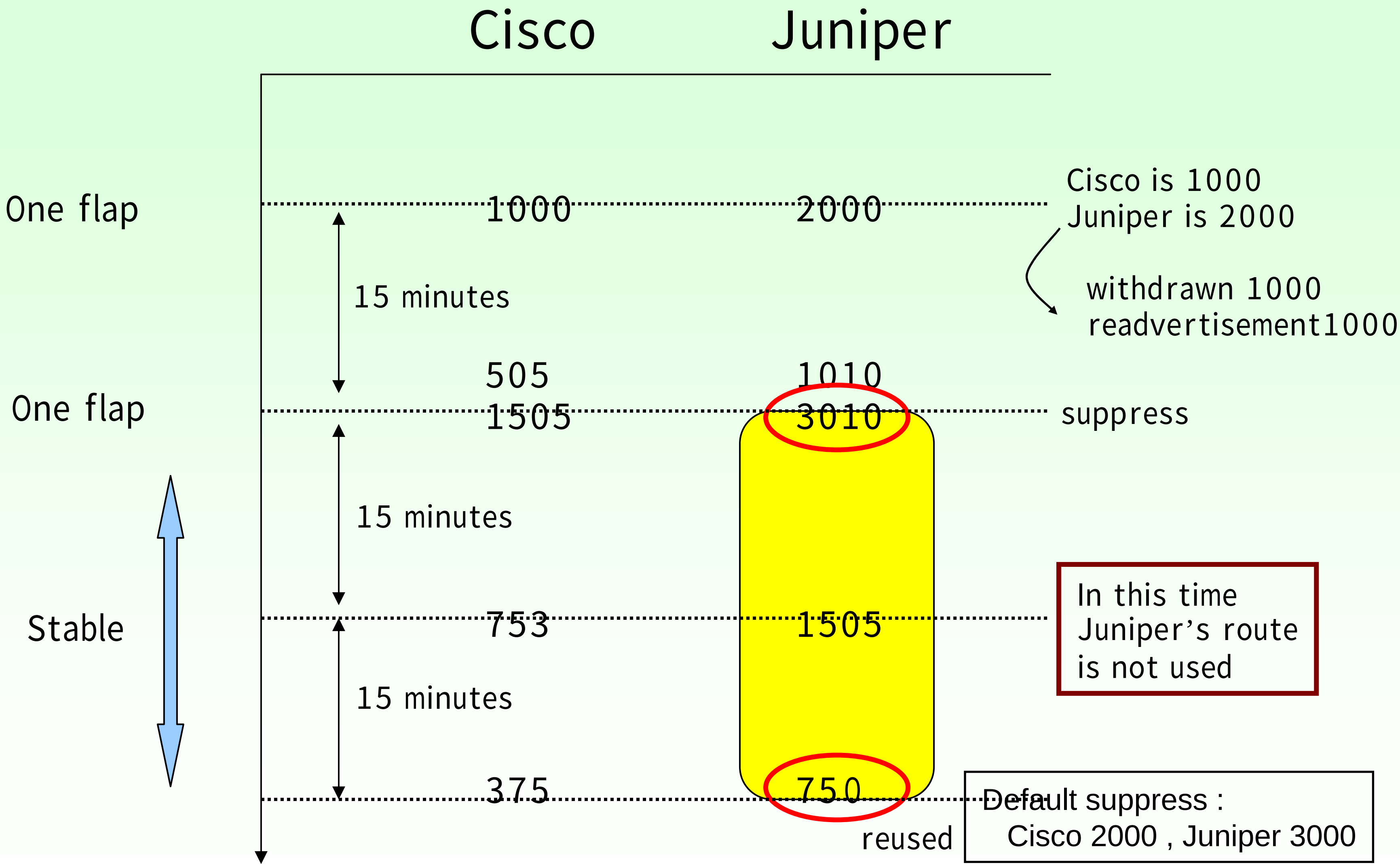
```
tomoya> show route 0.0.0.0
0.0.0.0/0          *[OSPF/150] 00:00:00, metric 10

                     via so-5/3/0.0
                     tvia ge-0/1/0.0
                   > via ge-0/2/0.0
```

```
tomoya> show route 0.0.0.0
0.0.0.0/0          *[OSPF/150] 00:00:05, metric 10
                   > via so-4/3/0.0
                     via so-5/3/0.0
                     via ge-0/1/0.0
                     via ge-0/2/0.0
```

Route cache is very useful

- Currently almost vendor is implemented “route refresh capability”
- But soft-reconfiguration inbound (for crs-1 need always keyword) is very useful
- When you set a new peer, you set low priority to this new peer, but more specific is strong!
  - Firstly check the route not receiving any route, only monitor the route from peer by using cache then receive

Route flapping experience



Cisco

Juniper

One flap

1000

2000

Cisco is 1000
Juniper is 2000

withdrawn 1000
readvertisement1000

15 minutes

505

1010

One flap

1505

3010

suppress

15 minutes

Stable

753

1505

In this time
Juniper’s route
is not used

15 minutes

375

750

reused

Default suppress :
Cisco 2000 , Juniper 3000

Routing Hijack

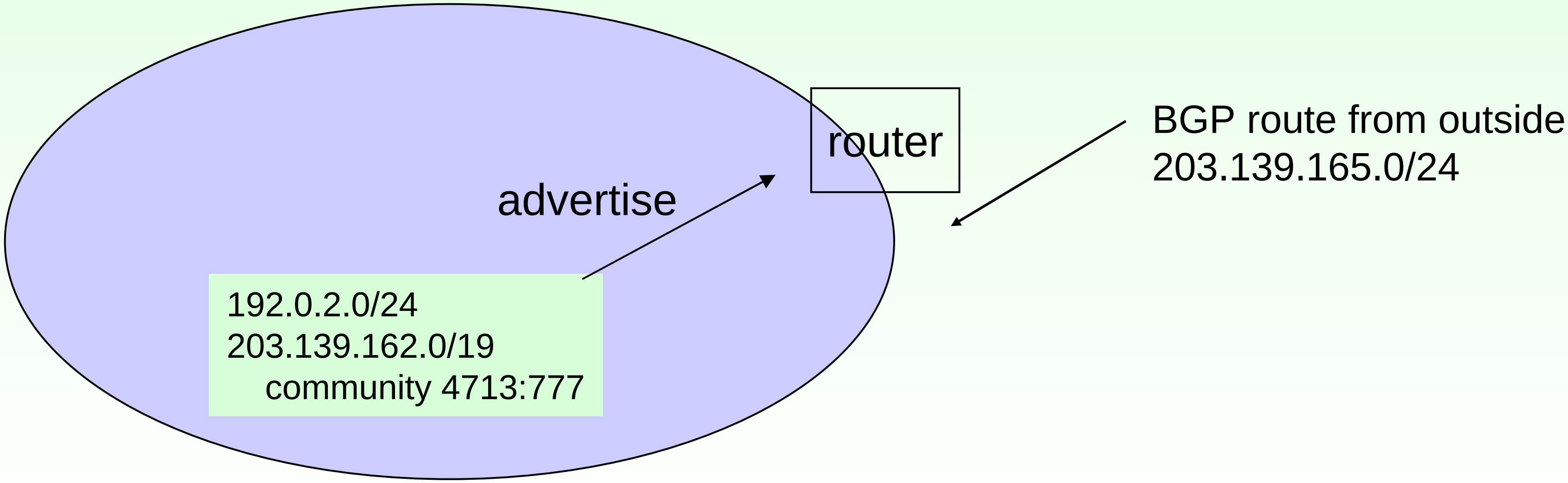
- We have around /10 IP blocks
- Sometimes our prefix hijacked
- When we hijacked our route, we announce more specific prefix to the internet
  - But When someone hijack /24, it is very difficult
    - We announce two /25s but almost ISPs cannot receive
    - Also we announce /24 in addition to /16 our PA
- We need BGP origin validation security mechanism
  - sBGP/So-BGP or IRR etc.

We need…

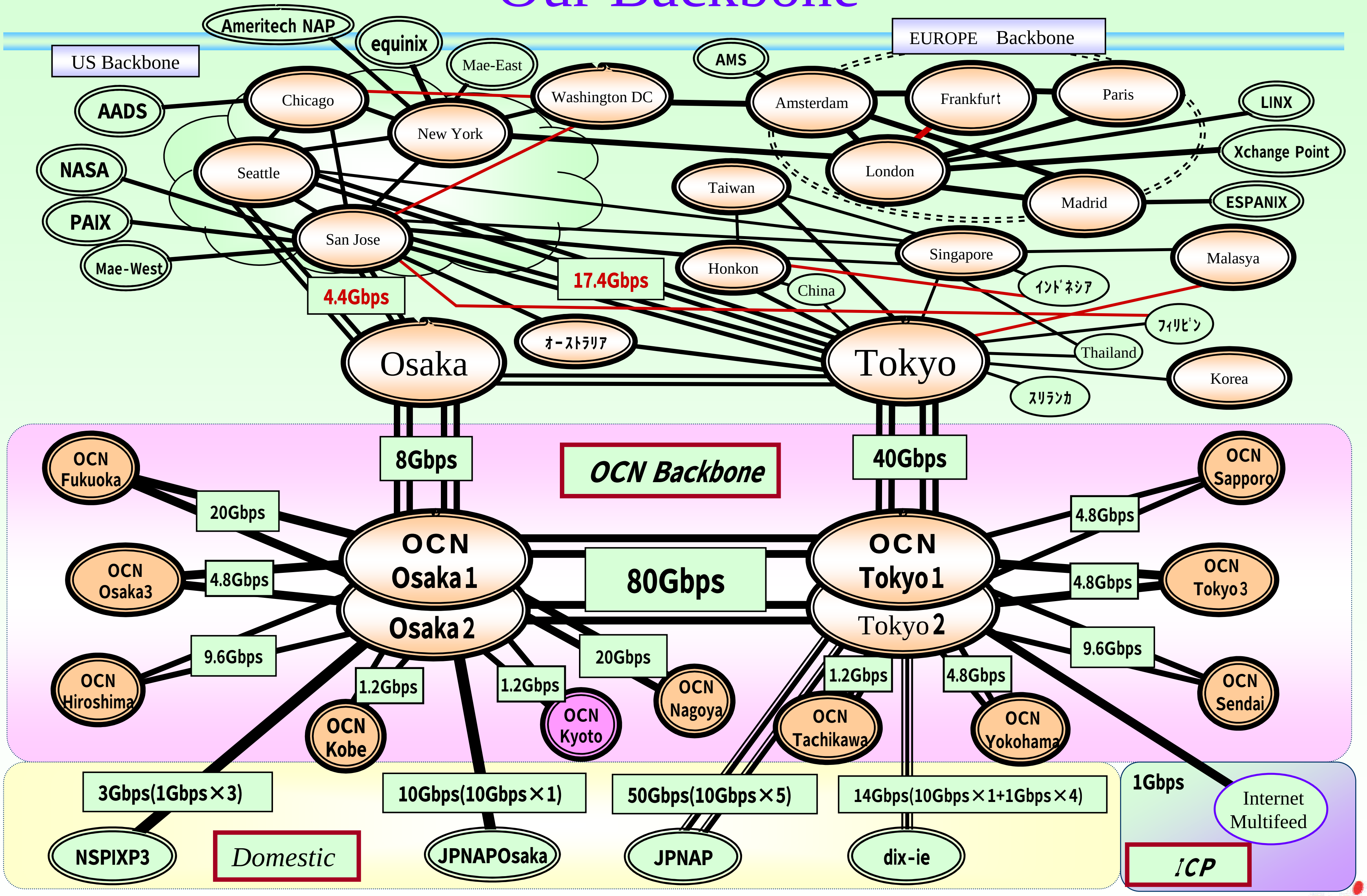
- TTL hack security mechanism for many vendor
- Prefix limitation by using LOC-RIB for Juniper
- Mac accounting for 10G
- Feasible path reverse path forwarding for uRPF
  - Strict mode is dangerous
  - Loose mode is just loose…
- BGP Inactive reason for Cisco is coming
  - Cisco implemented for CRS-1
  - Operational additional information is very important
- Dynmic filtering by using bgp community, just my idea

Dynamic Filtering : just idea

- If you receive the BGP route with this community (4713:777 attribute), the route which in scope of this community will be rejected automatically
  - Useful for filtering for your PA



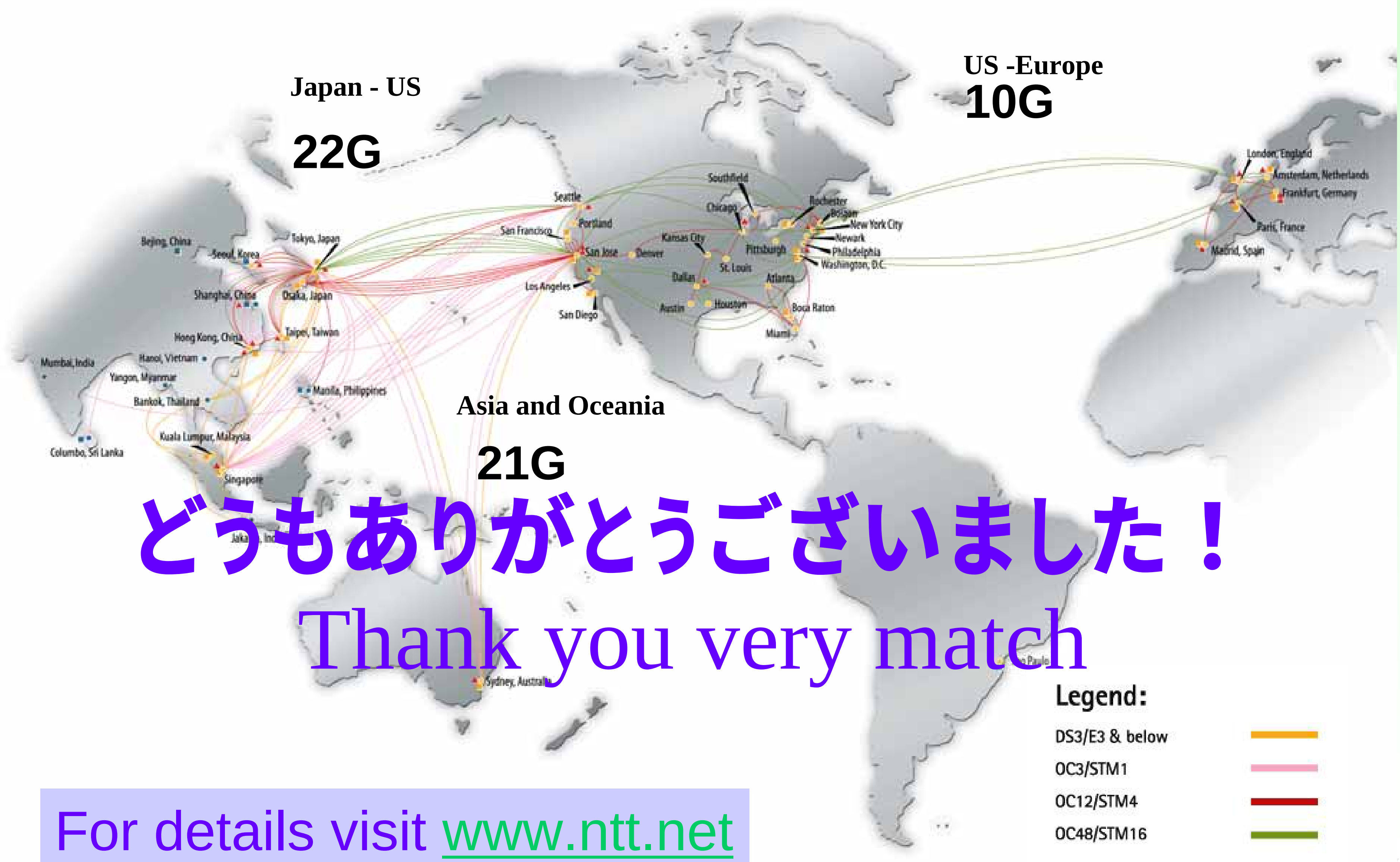
Our Backbone



NTT Communications

OCN

NTT Communications Global IP Network



Japan - US

22G

US -Europe

10G

Asia and Oceania

21G

どうもありがとうございました！

Thank you very match

For details visit www.ntt.net