IPv6 High Availability Strategies

Ivan Pepelnjak (ip@ipSpace.net)
ipSpace.net

Who is Ivan Pepelnjak (@ioshints)

- Networking engineer since 1985
- Technical director, later Chief Technology Advisor @ NIL Data Communications
- Consultant, blogger (blog.ipspace.net), book and webinar author @ ipSpace.net
- Teaching “Scalable Web Application Design” at University of Ljubljana

Focus:

- Large-scale data centers and network virtualization
- Networking solutions for cloud computing
- Scalable application design
- Core IP routing/MPLS, IPv6, VPN

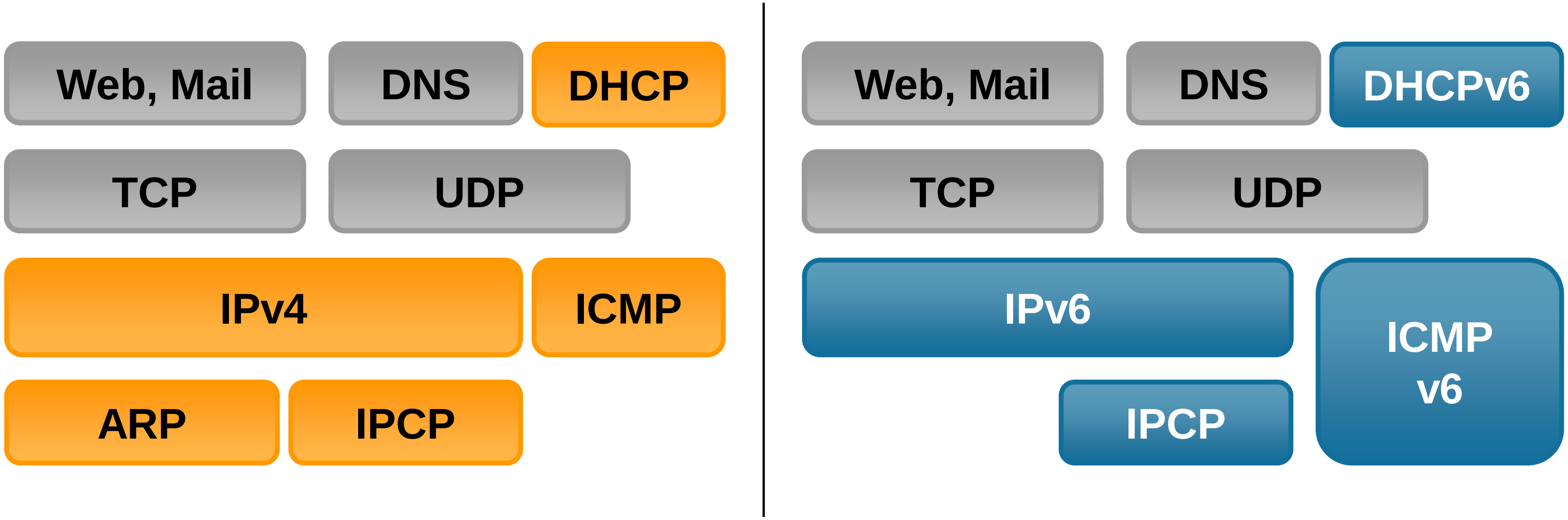
IPv6 Myths and Reality

IPv6 Myths

IPv6 will ...

- enable location/ID separation ✘
- solve IP multihoming issues ✘
- enable more reliable Internet ✘
- improve end-to-end QoS ✘
- give you better security due to embedded IPsec ✘
- be a prerequisite for IP mobility ✘
- be less secure than IPv4 due to lack of NAT ✘
- not require any change to your applications ✘

What is IPv6?



- IPv6 is a network-layer replacement for IPv4
- Longer addresses (128 bits)
- New routing protocols
- Some other changes in L2/L3 protocols
- Upper layers and applications should not change

No Changes To Applications? Keep Dreaming

Java

```
conn = new Socket("example.com",80)
```

Socket API

```
memset(&hints, 0, sizeof(hints));
hints.ai_family = PF_UNSPEC;
hints.ai_socktype = SOCK_STREAM;
error = getaddrinfo("example.com", "http", &hints, &res0);
if (error) { errx(1, "%s", gai_strerror(error)); }

s = -1;
for (res = res0; res; res = res->ai_next) {
        s = socket(res->ai_family, res->ai_socktype, res->ai_protocol);
        if (s < 0) { cause = "socket"; continue; }

        if (connect(s, res->ai_addr, res->ai_addrlen) < 0) {
                cause = "connect";
                close(s);
                s = -1;
                continue;
        }

        break;  /* okay we got one */
}
if (s < 0) { err(1, "%s", cause); }
```

High Availability Components

High Availability 101

A service is available = users can performs the transactions they want

Service availability includes

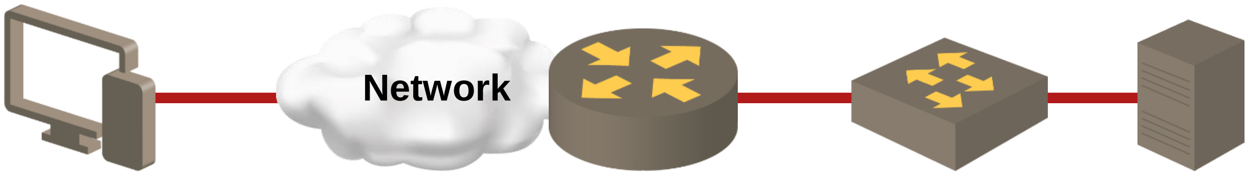
- Application availability
- Server and storage availability
- End-to-end network availability

Network availability includes

- Network services availability (DNS …)
- Network connectivity

Graceful degradation / failure resilience might be better than brute-force HA

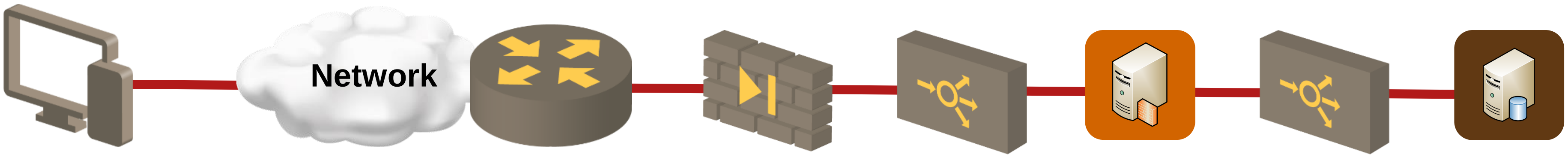
IPv6 Single-Server Applications



Network-level high availability

- Services (DNS – unchanged)
- Layer-2 (unchanged)
- First-hop router (new)
- Core network (new routing protocols, but similar)
- Multihoming (mostly unchanged, more options)

Complex IPv6 Application Stacks



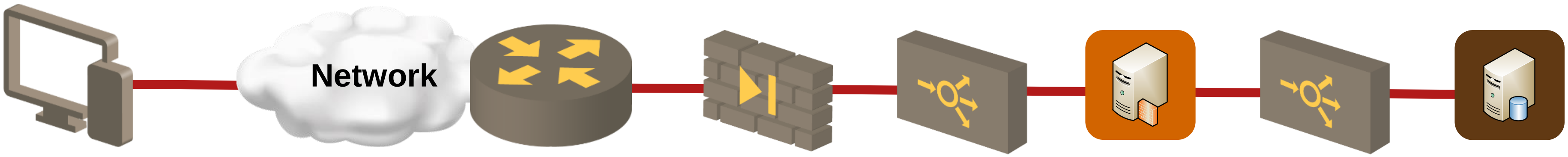
Additional application-level requirements

- Server-to-server communication
- Dependencies between application layers

Additional network-level high availability requirements

- Services: DNS, firewalls, load balancers

Beyond Networking



High availability components

- Connectivity
- Security
- Failure resilience
- Failover mechanisms
- Scale-out architectures

Review of IPv6 First-Hop Mechanisms

Review: Configuring Host IPv6 Parameters

Minimum set of parameters:

- Host IPv6 address
- Routing information (minimum: first-hop router’s IPv6 address)
- DNS server IPv6 address (could use IPv4 DNS server in dual-stack environments)

Configuration mechanisms:

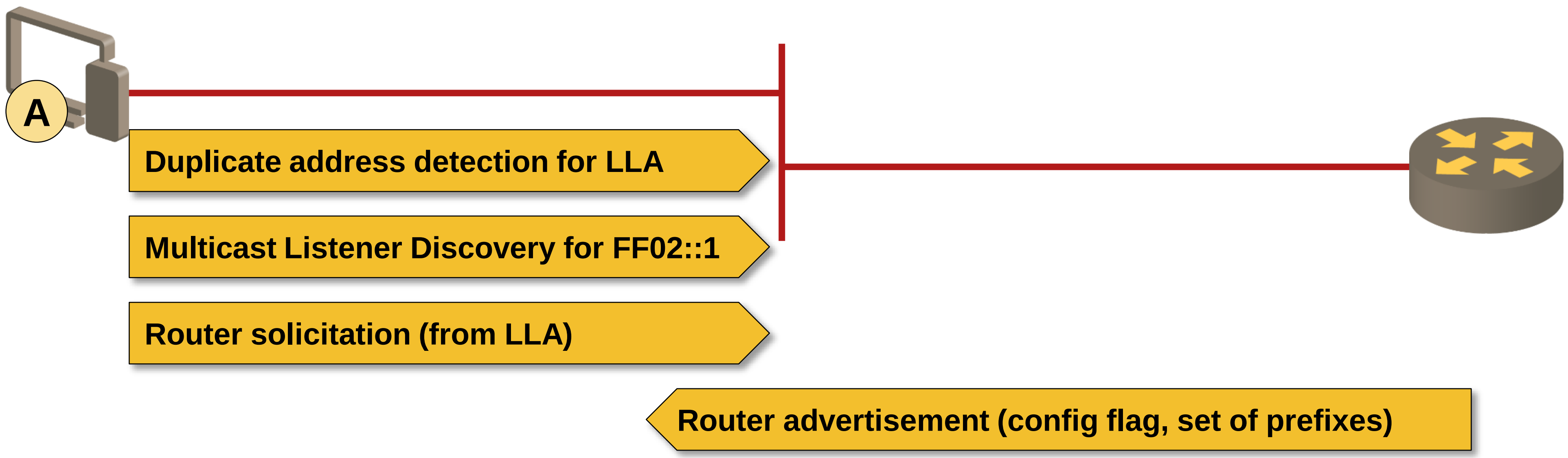
- Static configuration (servers, routers)
- Stateless Autoconfiguration (SLAAC) using Router Advertisements
- DHCPv6-based configuration

Review: Dynamic Host Configuration Options

Parameter	ICMPv6 (ND/RA)	DHCPv6
Host IPv6 address	Yes (SLAAC)	Yes
First hop router’s IPv6 address	Yes (RA)	No
DNS server’s IPv6 address	Yes (RFC 6106)	Yes

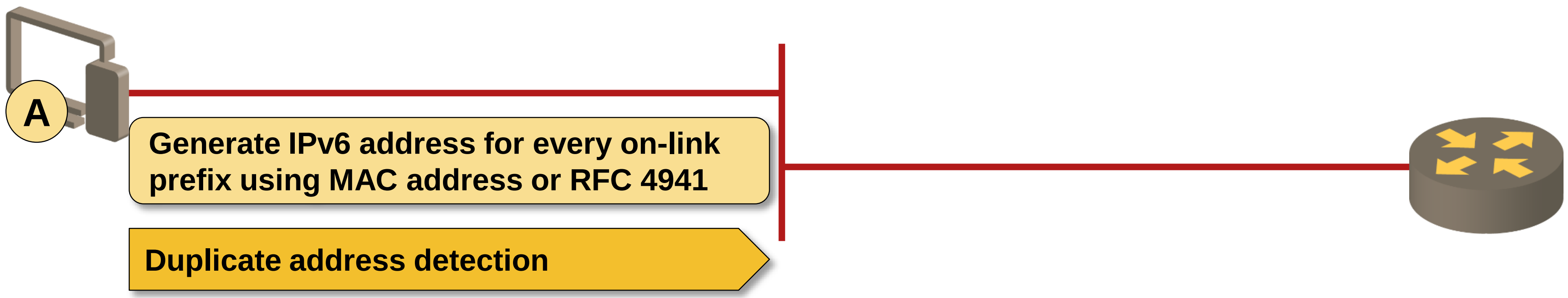
- RFC 6106 is not widely supported yet
- In most cases you need both RA and DHCPv6
- SLAAC with dynamic DNS registration is preferred to DHCPv6-based address allocation on client segments

Review: Host Configuration, Part 1



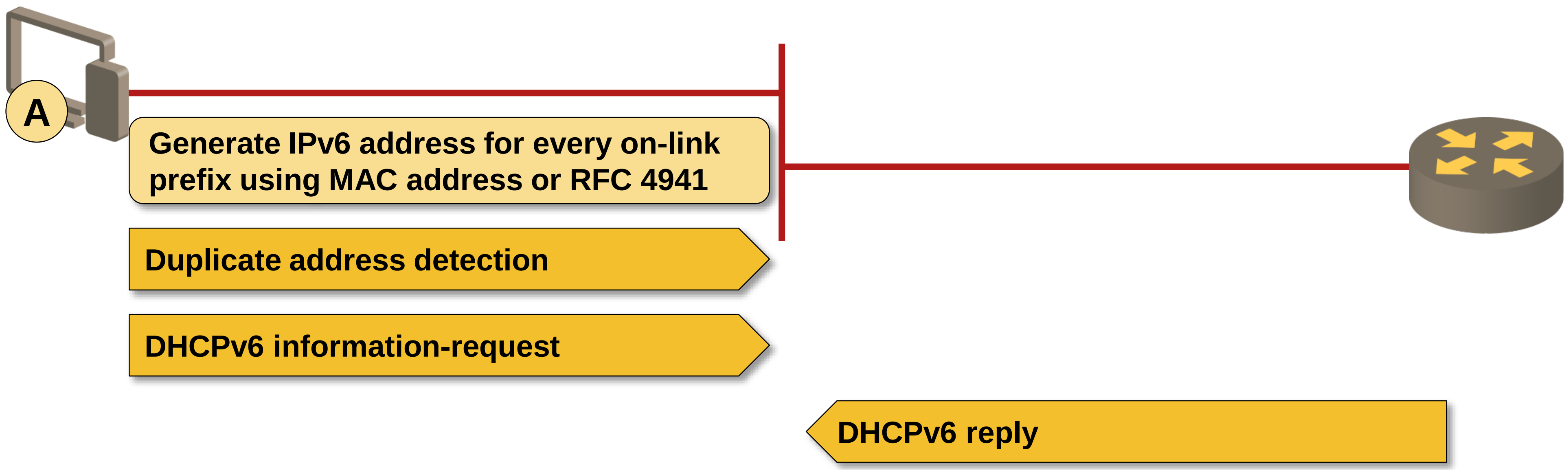
- Newly started host must first get a LLA (using its MAC address)
- Duplicate address detection is used to check LLA uniqueness
- Host joins the all-nodes multicast group (MLD needed for L2 switches)
- Host tries to find an adjacent router to get configuration mechanisms and on-link prefixes
- DHCPv6 may be used if no routers are present on the link

Review: Host Configuration – SLAAC



- Host generates an IPv6 address for every prefix with auto-configuration (A) flag advertised by the routers
- Host uses duplicate address detection to check generated address uniqueness
- Non-RFC4941 SLAAC fails if the host encounters a duplicate IPv6 address (indicating duplicate Interface ID – MAC address)
- IPv4 DHCP can be used in dual-stack environments to specify DNS server IPv4 address

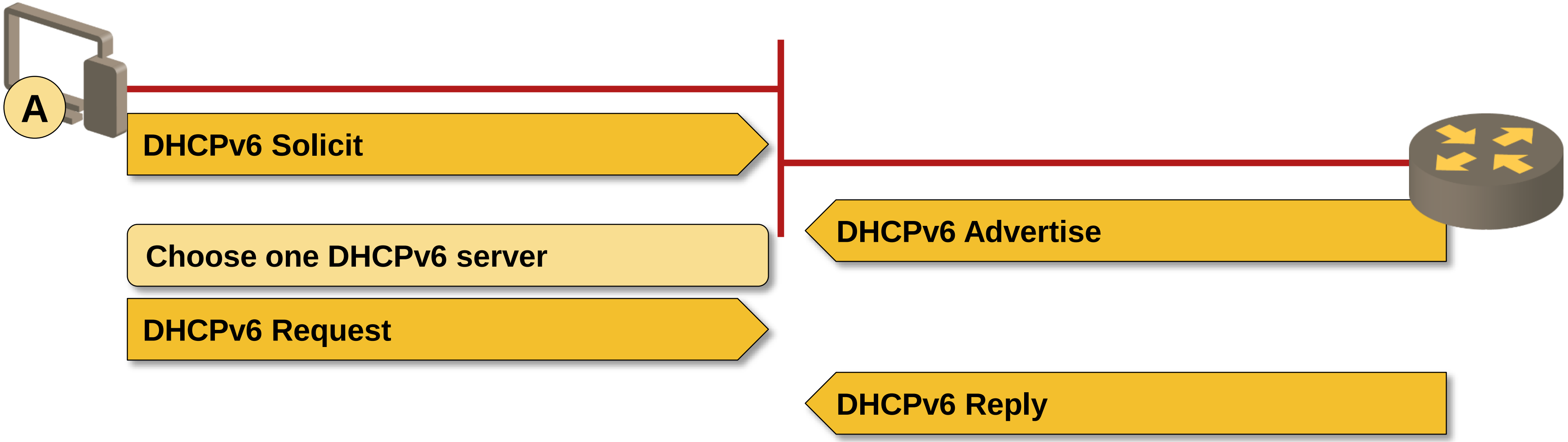
Review: Host Configuration – SLAAC + DHCPv6



SLAAC+DHCPv6 used when O (Other configuration) flag is set in RA

- SLAAC is used to generate IPv6 addresses for all prefixes advertised with A flag
- DHCPv6 request is sent to retrieve non-address parts of the configuration (DNS server IPv6 address)
- Router can reply to the DHCPv6 request or relay it to a central DHCPv6 server

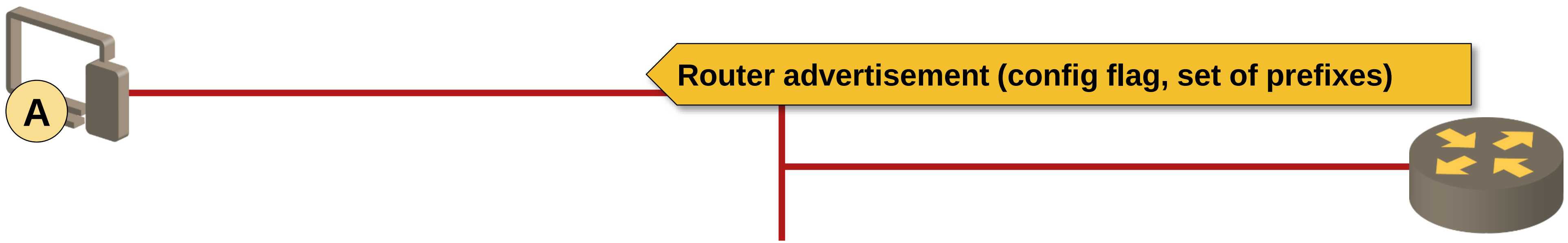
Review: Host Configuration – DHCPv6 Only



Used when router advertisements contain M (Managed addresses) flag:

- DHCPv6 is used to assign IPv6 addresses (and other parameters) to the hosts
- Two-step process like IPv4 DHCP
- Router can run a DHCPv6 server or relay DHCPv6 requests
- Rapid commit (one step process – Solicit message answered with Reply message) can be used if supported by the client and the DHCPv6 server

Review: On-Net/Off-Net Determination



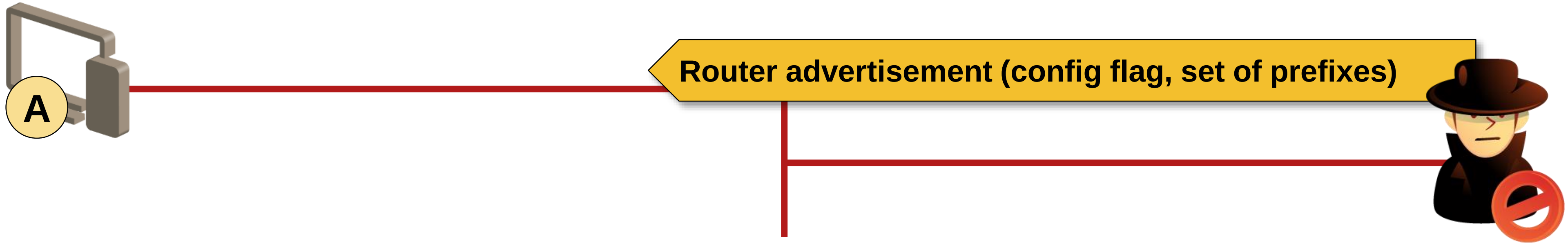
Routers advertise locally-significant IPv6 prefixes in router advertisements

- Prefixes with A flag set are used for SLAAC
- Prefixes with L flag set are on-net prefixes
- First-hop router is the source IPv6 address of the RA

Default host IPv6 packet forwarding procedures

- Destination IPv6 address in a prefix with L flag ➔ send directly
- All other IPv6 destinations ➔ send to first-hop router
- Behavior in multi-router environment is unspecified (and varies by OS)
- Static configuration usually overrides RA-derived information

Why Is This Relevant?



An intruder might start sending IPv6 RA messages

- IPv6 is enabled by default on most operating systems
- Servers will auto-configure themselves
- Intruder can advertise itself as IPv6 default router and IPv6 DNS
- IPv6 DNS might take precedence over IPv4 DNS
- IPv6 transport **will** take precedence over IPv4 transport
- With proper RA messages (prefixes without on-net flag) all traffic goes through the intruder's node

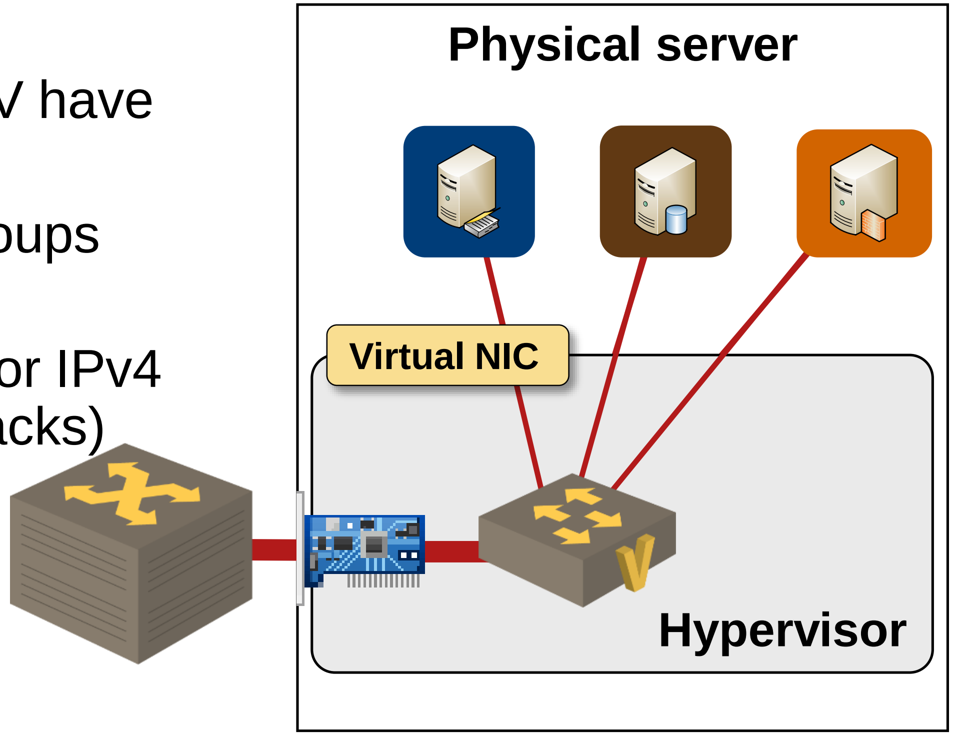
First-hop IPv6 security mechanisms are a MUST

The Virtual Fiasco

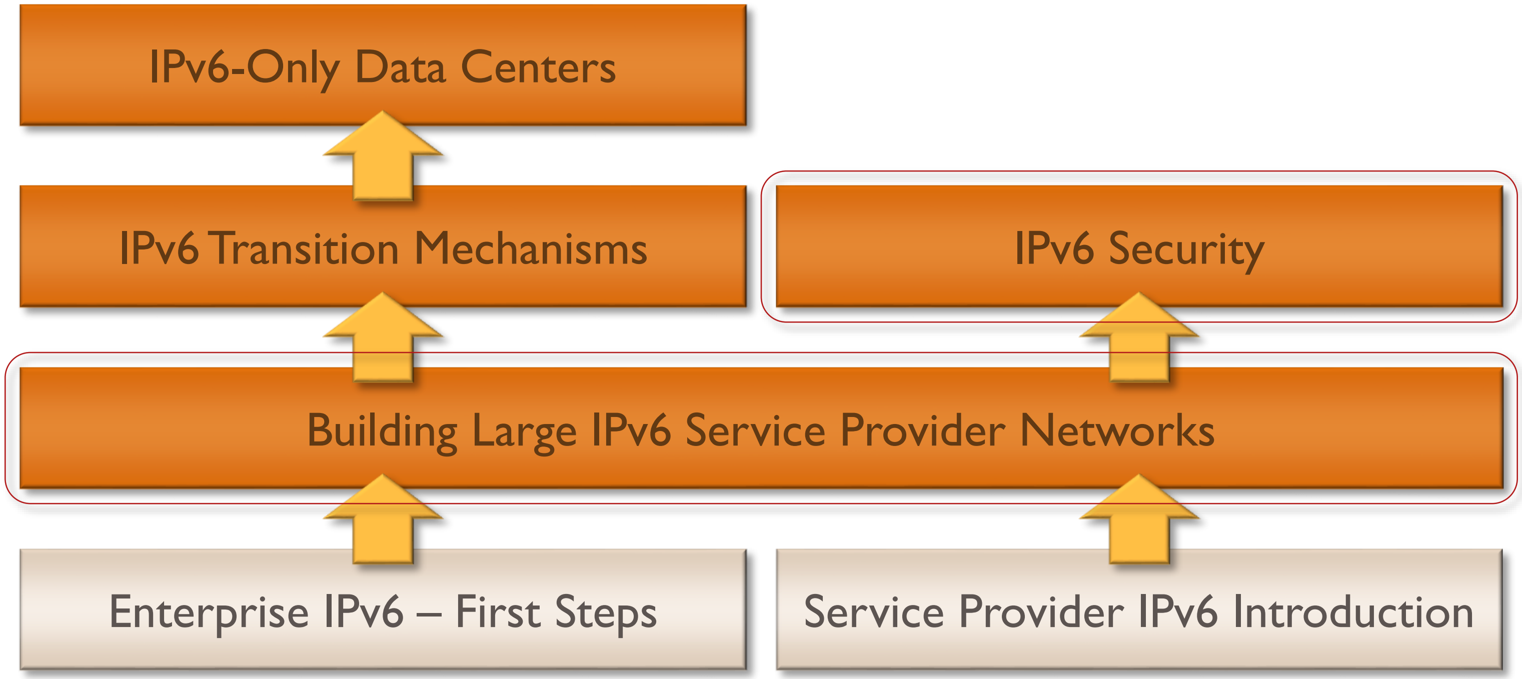
- First-hop security MUST be implemented on the first layer-2 switch
- In virtual environments the first switch is the virtual switch
- Virtual switch MUST implement IPv6 first-hop security features: RA guard, DHCPv6 guard, Source/Destination guard, Binding Integrity guard

State-of-the-art:

- vSphere 5.5, vCNS 5.5 and Nexus 1000V have no IPv6 security features
- OpenStack Havana has IPv6 security groups (and little else)
- Hyper-V implements layer-3 forwarding for IPv4 and IPv6 (and thus blocks most IPv6 attacks)
- Amazon VPC does not support IPv6 (but does not propagate it either)



IPv6 Webinars on ipSpace.net



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

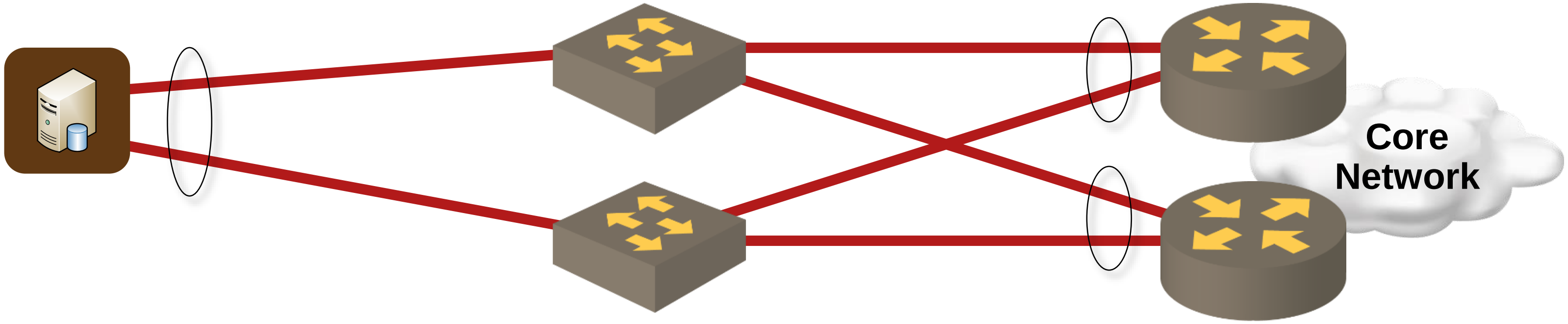
Other options

- Customized webinars
- ExpertExpress
- On-site workshops

More information @ http://www.ipSpace.net/IPv6

IPv6 First-Hop High Availability

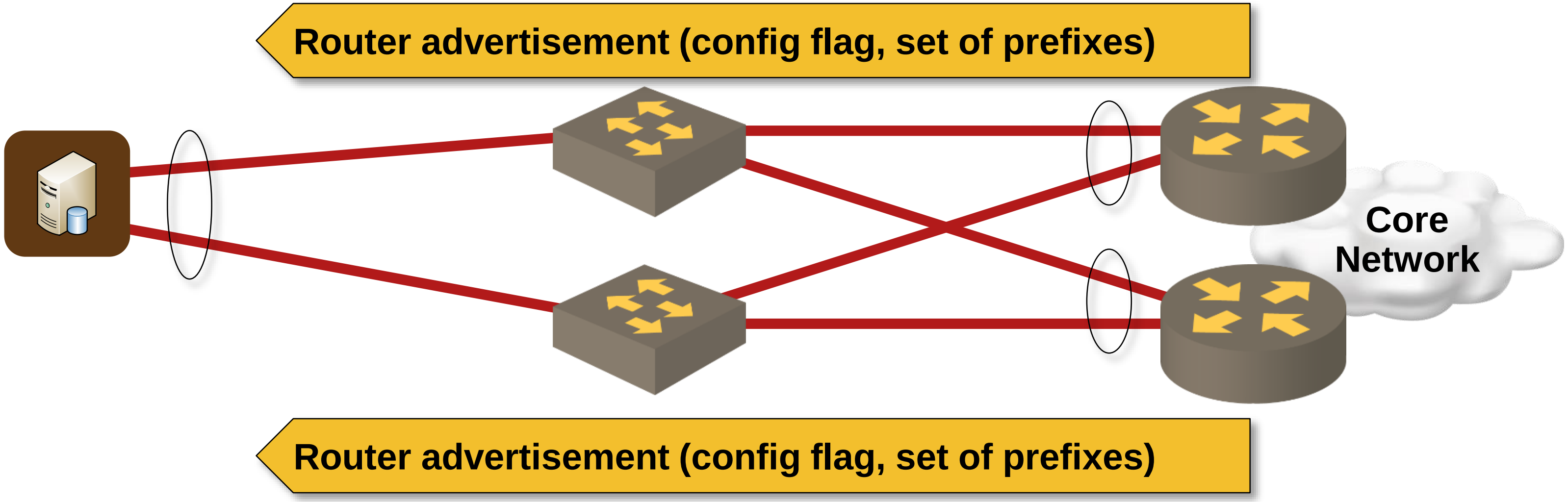
Typical High-Availability Setup



IPv6-specific modifications:

- No changes on servers (all NIC teaming modes work as expected)
- No changes on L2 switches (might need MLD snooping)
- First-hop L3 switches must be configured for high-availability environment

Router Advertisements in Dual-Router Environment



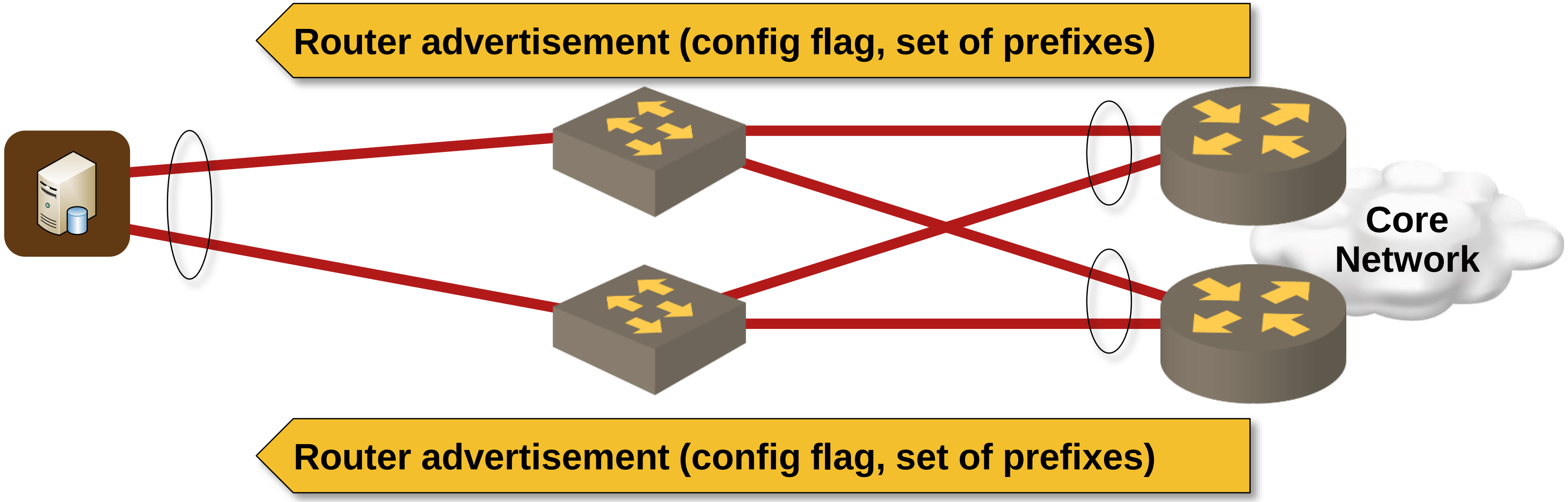
All routers advertise their presence with RA messages

- Router's LLA and physical MAC address

Host behavior varies between operating systems (and OS versions)

- Use the first RA received as long as it's valid
- Load-balance between all first-hop routers
- Use the last RA received (flip-flopping between routers)

Are Router Advertisements Good Enough?



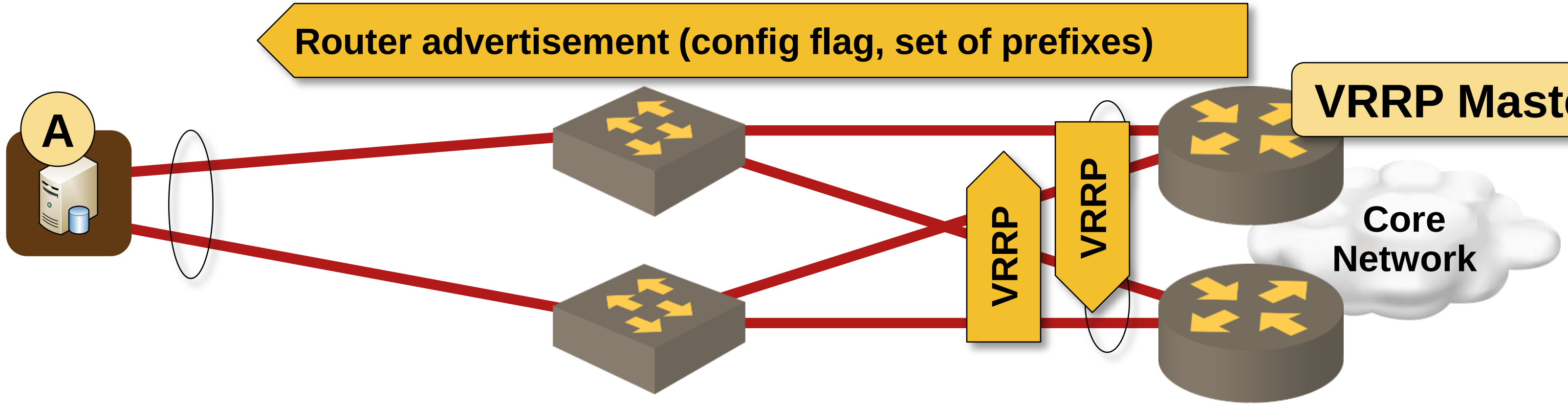
RA timers can be adjusted on most routers and L3 switches

- Minimum RA interval = 30 msec (Cisco IOS)
- Minimum RA lifetime = 1 sec
- Hosts will stop using a failed router after RA expiration

RA-based failover

- Uses CPU cycles on every attached host
- Might be good enough in some environments

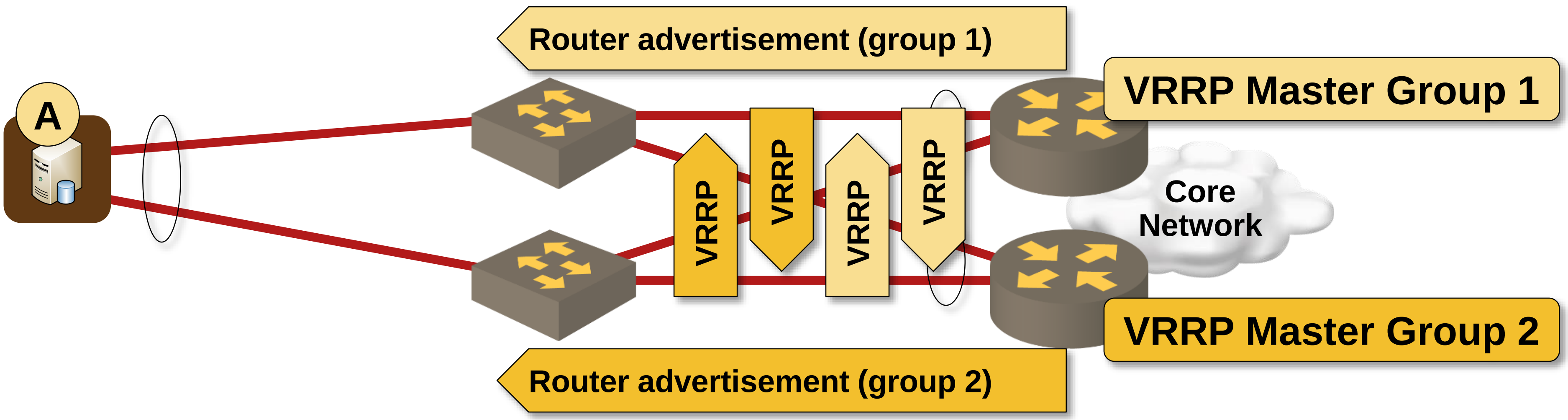
VRRP v3 = FHRP for IPv6



- VRRP configured on server-facing subnets
- Routers elect VRRP master
- VRRP master sends RA messages with VRRP IPv6 and VRRP MAC address
- VRRP backup router takes over VRRP MAC address after VRRP primary router failure

Sub-second convergence is possible (based on VRRP implementation)

Load Balancing with VRRP v3

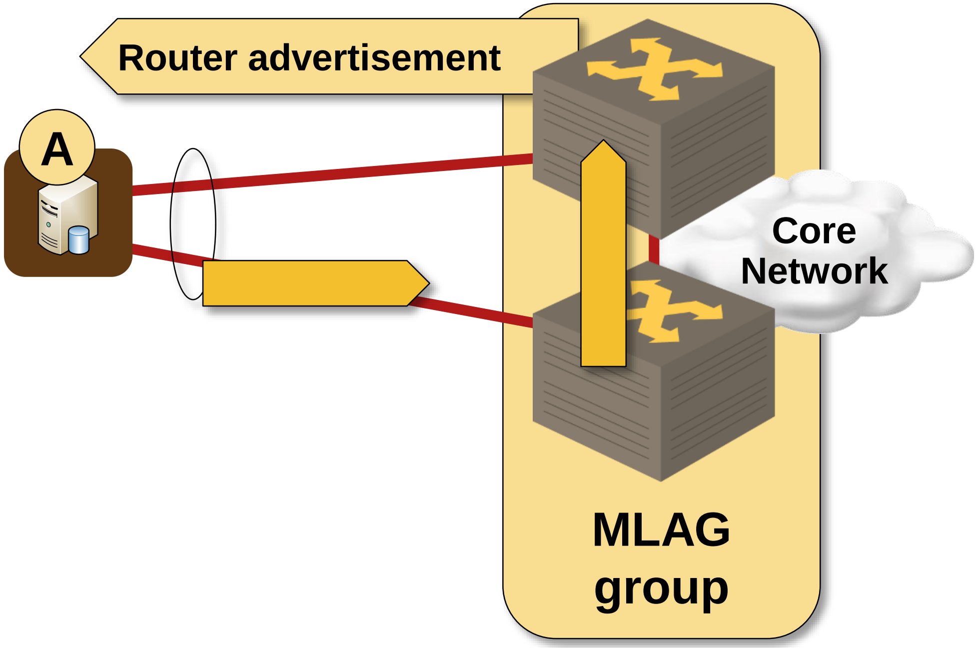


- Multiple VRRP groups configured on the same interface
- Multiple VRRP masters (one per group)
- Each VRRP master sends RA messages with its group's IPv6 and virtual MAC address
- Hosts **might** load-balance across multiple VRRP routers

Might require static server configuration (no first-hop router in DHCPv6)

First-Hop Redundancy on Layer-3 Switches

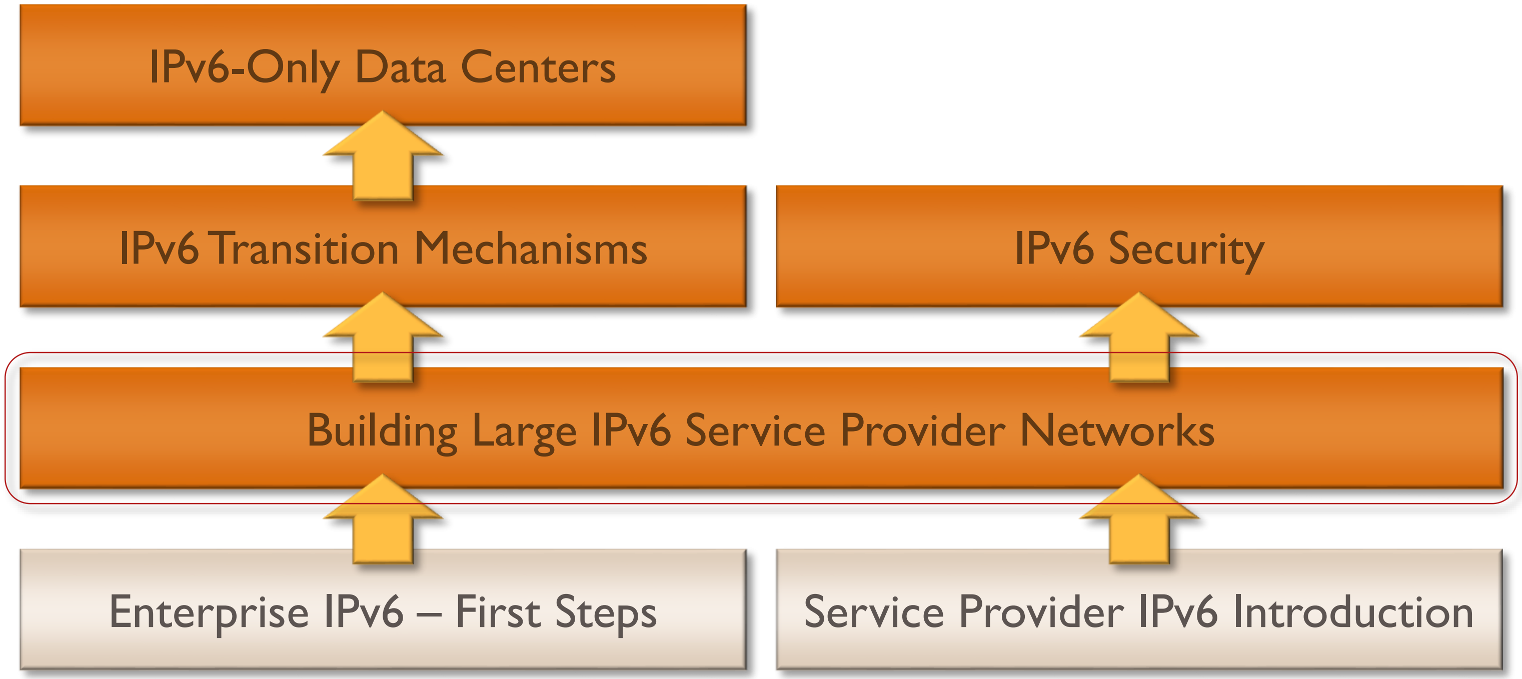
- Each L3 switch advertises its own physical MAC address
- Packet forwarding may become suboptimal
- Loop prevention logic might prevent proper packet forwarding



Correct design:

- Use VRRP v3 (or HSRP for IPv6)
- Both switches forward traffic sent to virtual MAC address

IPv6 Webinars on ipSpace.net



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

Service Endpoint High Availability

IPv6 Solutions Almost Identical to IPv4 Solutions

Local high availability

- Clusters with shared IP address
- Load balancers

Redundant Internet connectivity

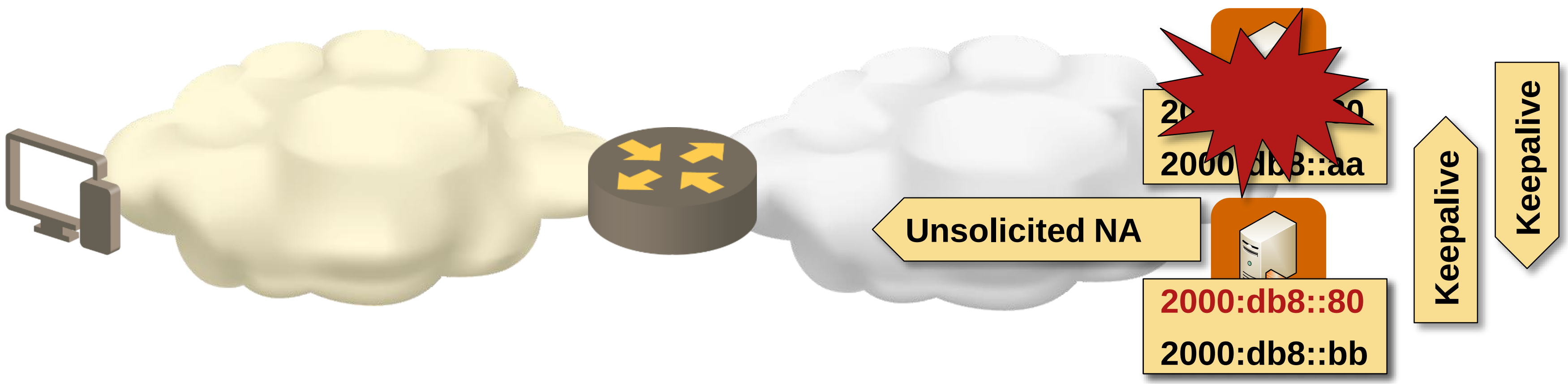
- BGP multihoming
- NAT/NPT with multiple uplinks (clients only)
- Mobile IP (clients only – better integrated in IPv6)
- LISP (new)

Global high-availability

- DNS-based solutions (including geolocation)
- Anycast

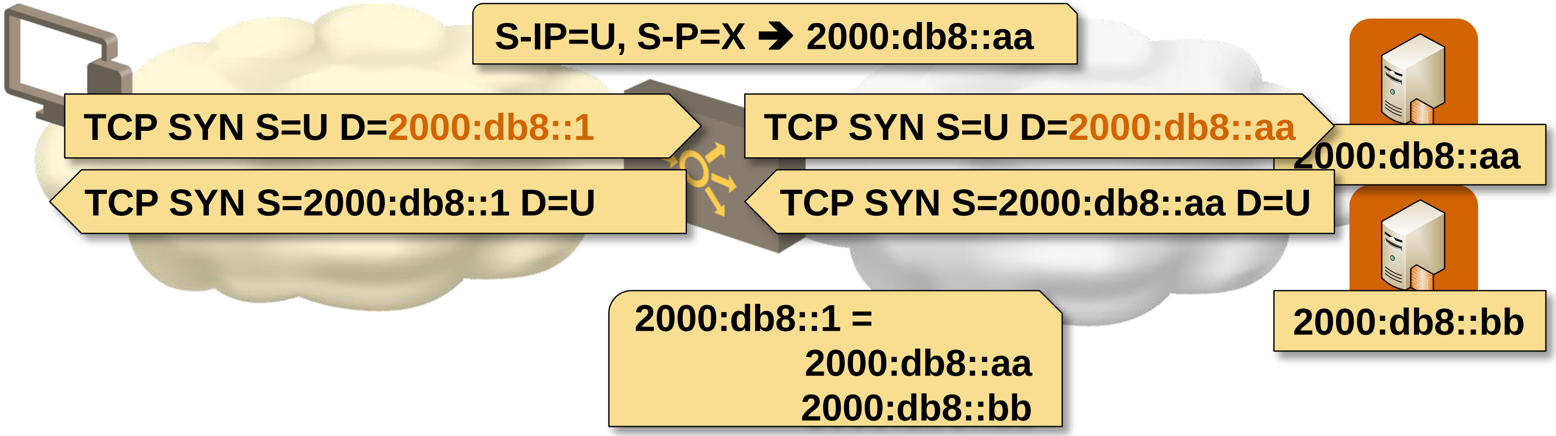
Local Endpoint HA Solutions

IPv6 Server Clusters



- Almost identical to IPv4 solution
- Each cluster node has a “regular” IPv6 address
- Primary node (per service) owns service IPv6 address
- Node availability checked with a keepalive protocol between cluster members
- Backup node takes over services and IPv6 addresses of a failed primary node
- Backup node sends unsolicited neighbor advertisement (equivalent to gratuitous ARP) to purge ND caches in all adjacent nodes

Load Balancers

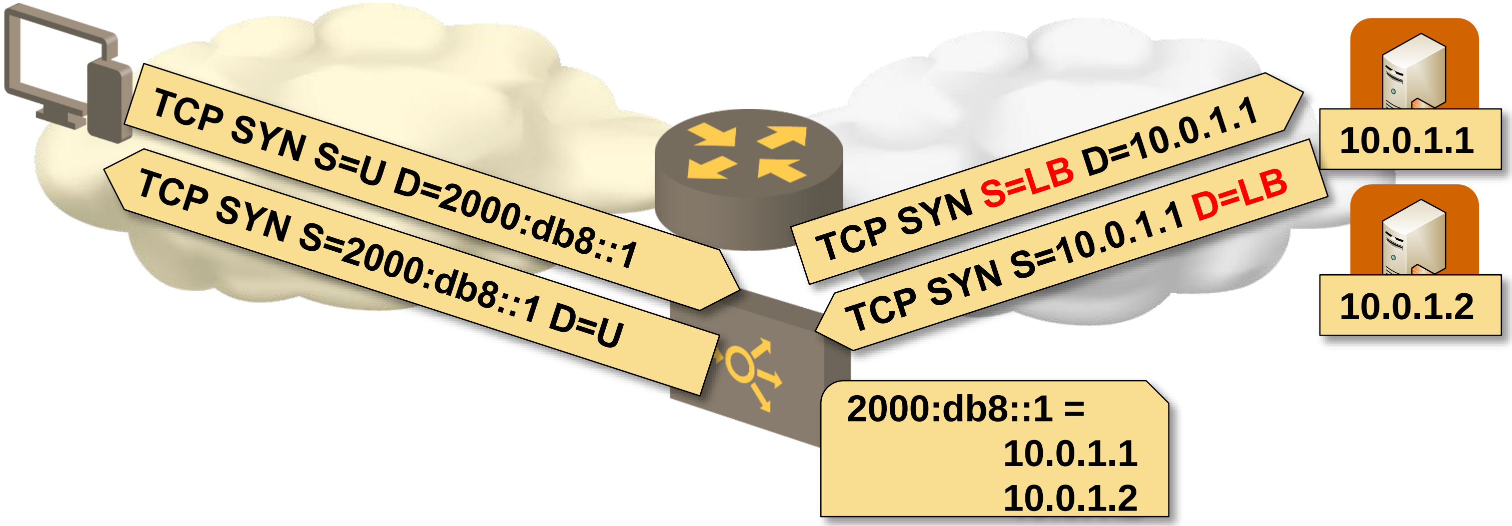


SLB66 is almost identical to SLB44

- Load balancer in the forwarding path (destination NAT)
- SNAT for out-of-path load balancer (source + destination NAT)
- Direct server return (shared destination address, no NAT)

SLB is needed due to TCP and Socket API limitations

Load Balancers – Protocol Translation (SLB64)



Make IPv4 content available to IPv6 clients

- Virtual IP address = IPv6 address
- Server pool = IPv4 or IPv6 addresses
- Source and destination addresses must be in the same address family
  ➔ Source NAT is mandatory

Typical Steps

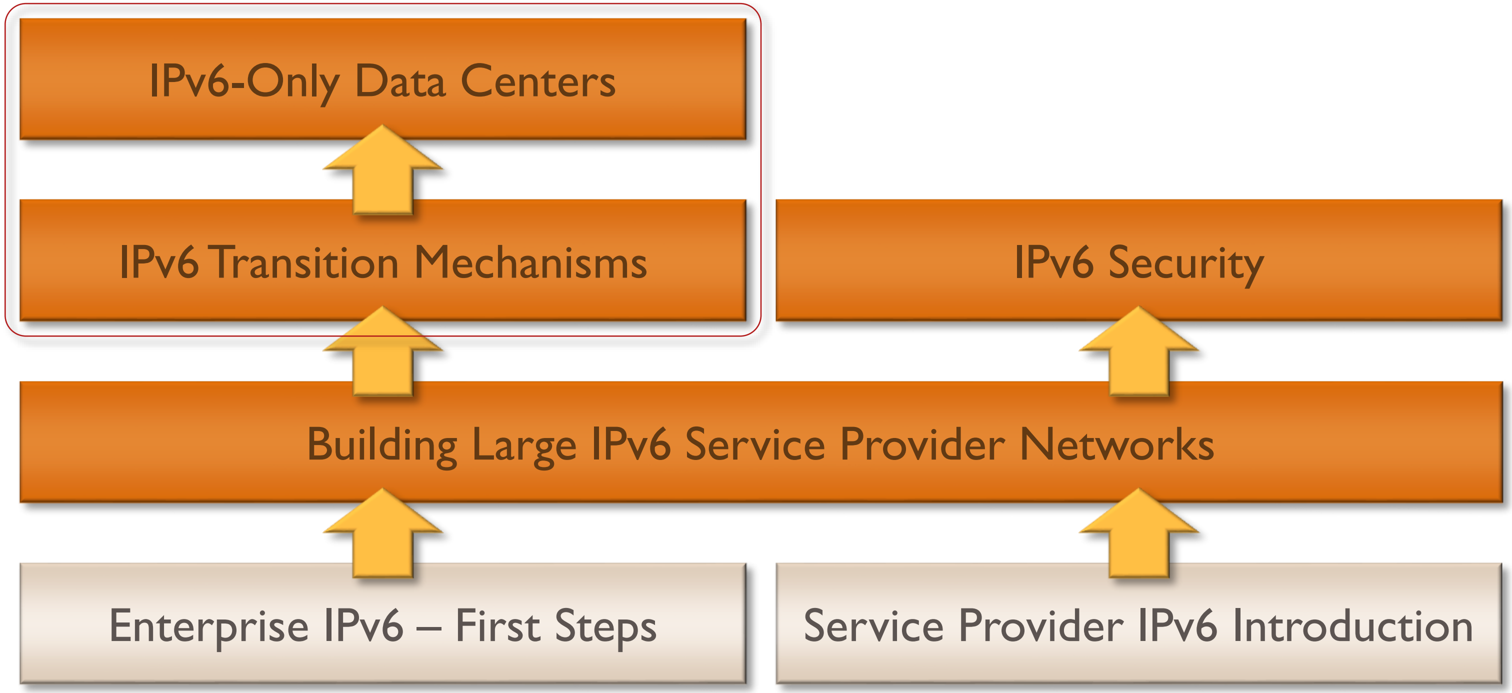
• IPv4 only	Losing control of user experience
• NAT64	Why are we having performance issues?
• SLB64	Darn, we lost client IP addresses
• Dual-stack servers	Ouch, this is complex
• IPv6-only servers with SLB46	
• IPv6-only data center with NAT46	
• No IPv4	... in a universe far far away

Let Me Recap

- IPv4 only
- NAT44 in DMZ
- SLB64, IPv4-only servers
- SLB64, SLB66, dual-stack servers
- SLB66, IPv6-only servers
- NAT46, SLB66, IPv6-only servers
- IPv6 only

How many migrations do you want to do in the next 5 years?

IPv6 Webinars on ipSpace.net



Availability

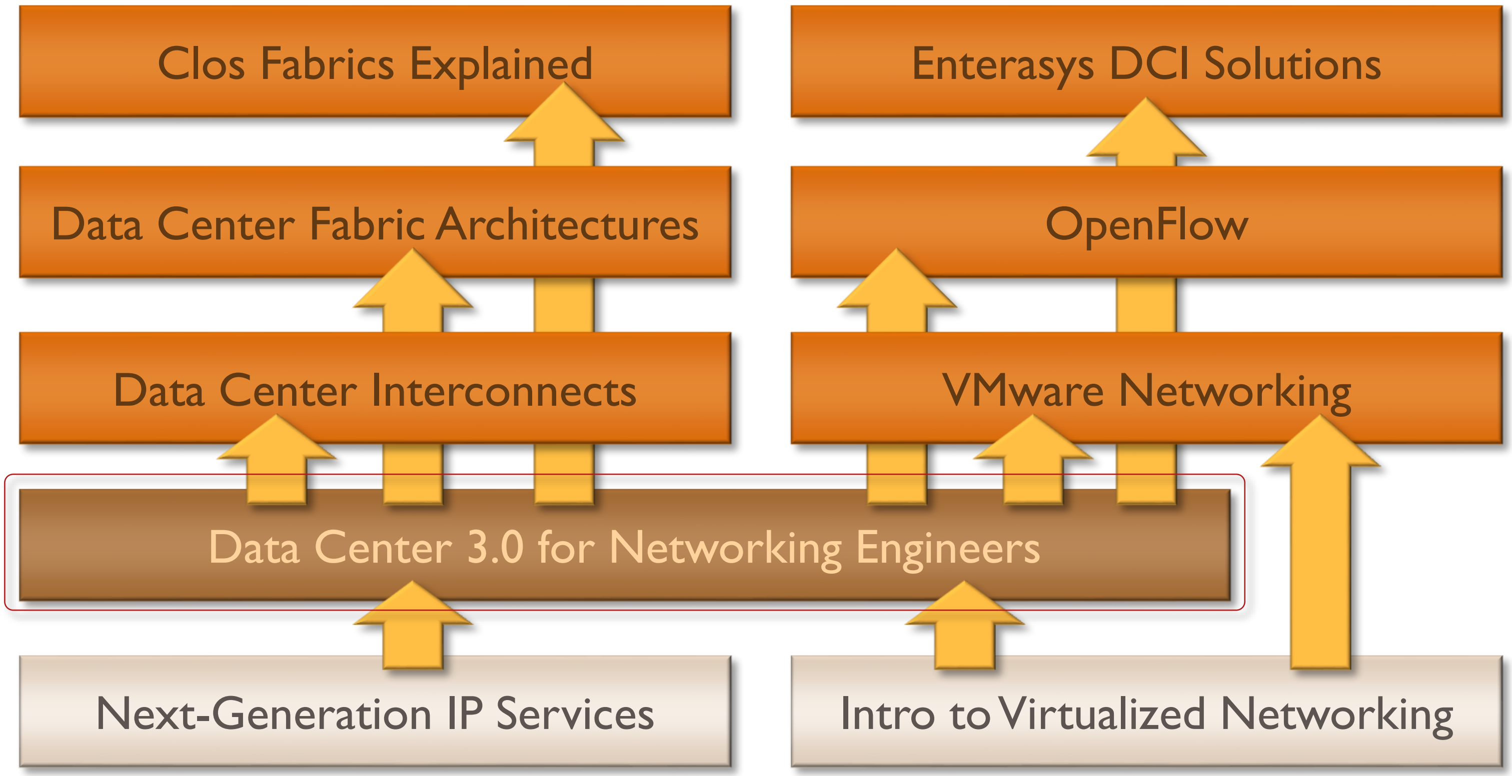
- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

More information @ http://www.ipSpace.net/IPv6

Data Center Webinars on ipSpace.net



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

More information @ http://www.ipSpace.net/DC

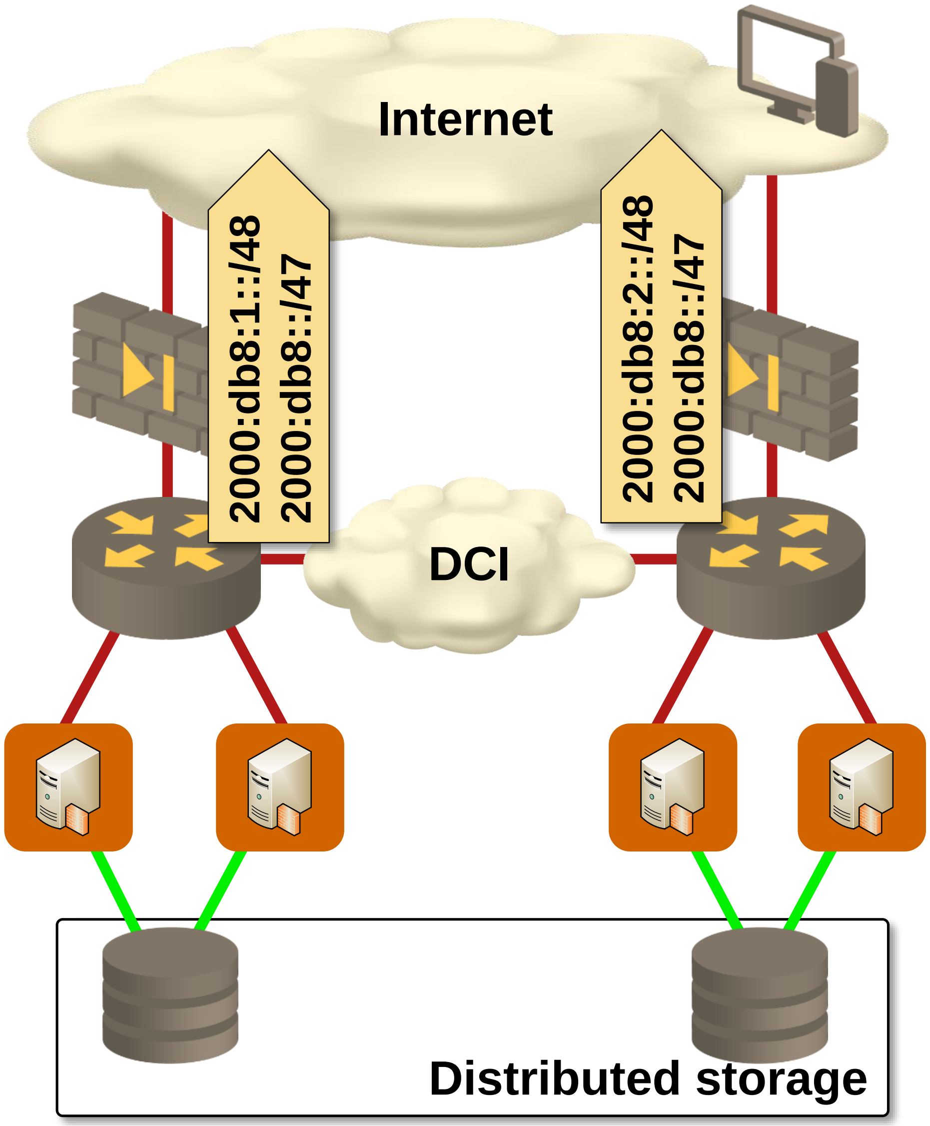
Redundant Network Connectivity

External Connectivity: Specific+Summary Prefix

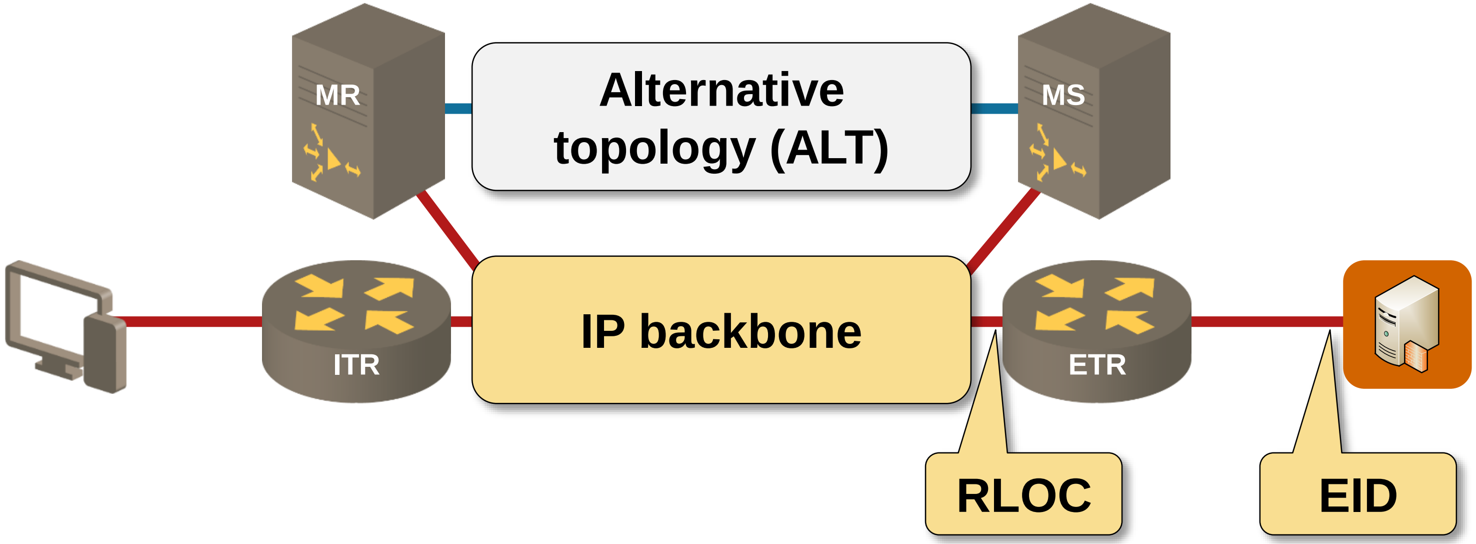
- Each data center advertises its own prefix
- Both data centers advertise a summary prefix for backup purposes

Results:

- Traffic flows are optimal
- DCI heavily loaded during external connectivity failures ➔ use DNS-based load balancing
- Stateful firewalls in forwarding path will break TCP sessions after external link failure/recovery



Introduction to LISP

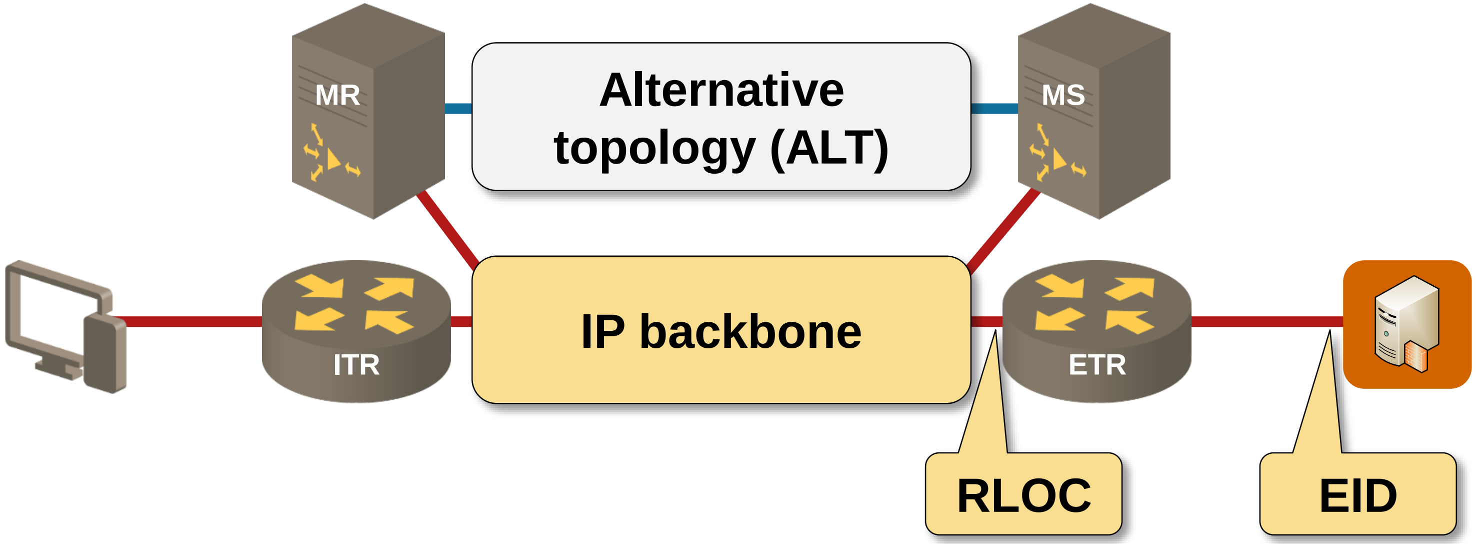


LISP = Locator/Identity Separation Protocol

- Maps host IP prefix (EID) into transport IP address (RLOC)
- EID is fixed, RLOC can change
- Host-to-host traffic is UDP-encapsulated between ITR and ETR
- Global EID-to-RLOC mapping service

LISP works for any combination of IPv4 and IPv6

LISP Terminology



ITR: Ingress Tunnel Router

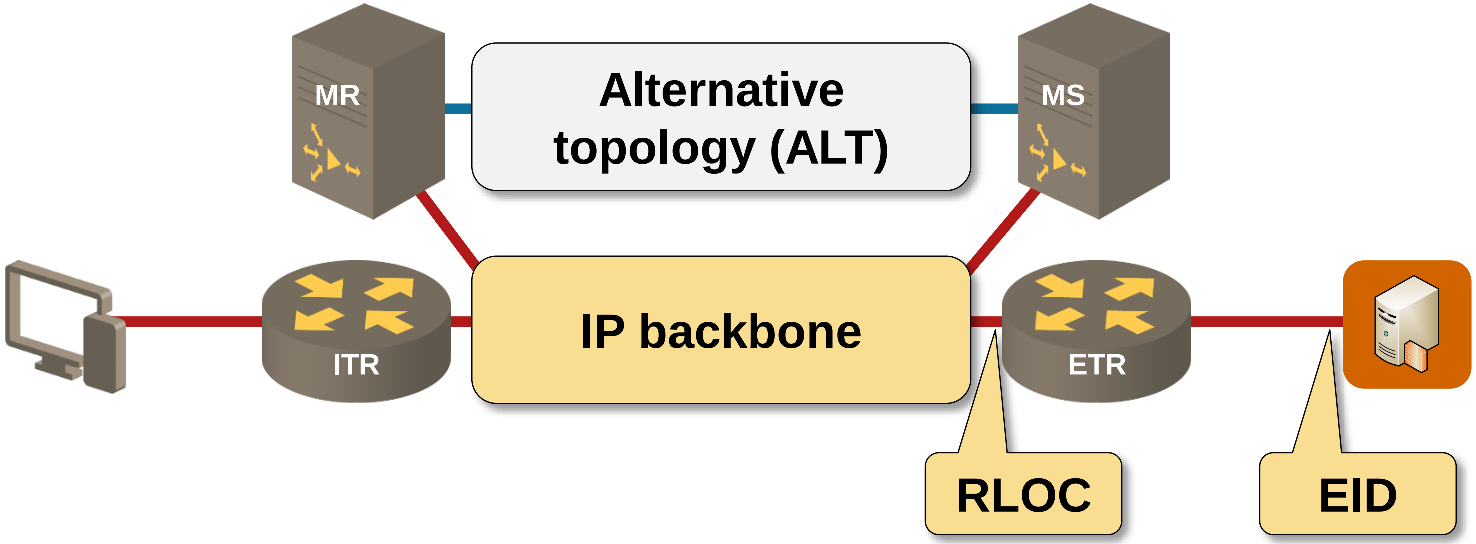
ETR: Egress Tunnel Router

MR: Map Resolver (performs EID-to-RLOC mapping for ITR)

MS: Map Server (ETR registers EID-to-RLOC mappings with MS)

ALT: Alternate topology (BGP over GRE) propagates EID-to-RLOC mapping information

A Day in Life of a LISP Packet

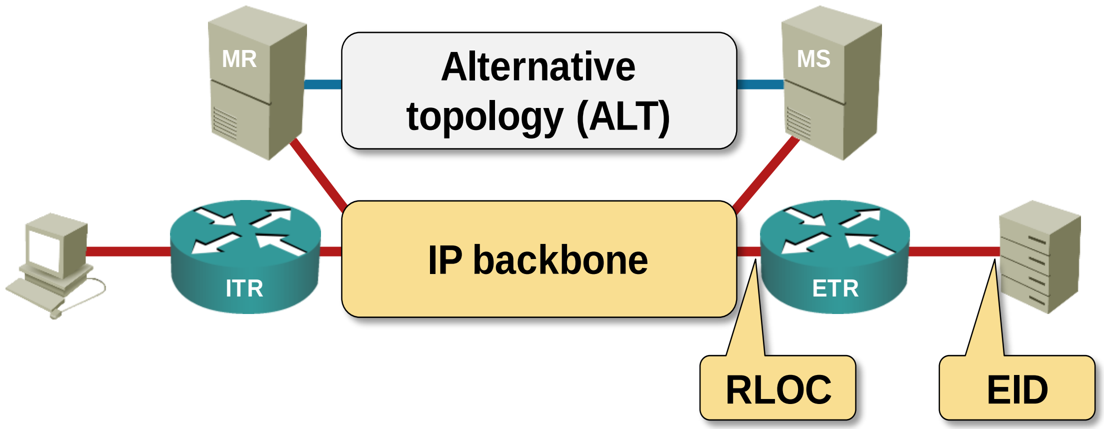


1. Host sends an IP packet to ITR
2. ITR performs EID-to-RLOC lookup in local cache
3. ITR encapsulates IP packet into LISP+UDP+IP envelope
4. ITR sends IP packet addressed to ETR RLOC into IP backbone
5. ETR receives LISP packet, decapsulates it and performs EID lookup
6. ETR forwards original IP packet toward target EID

EID-to-RLOC Mapping Service

Topology-driven actions

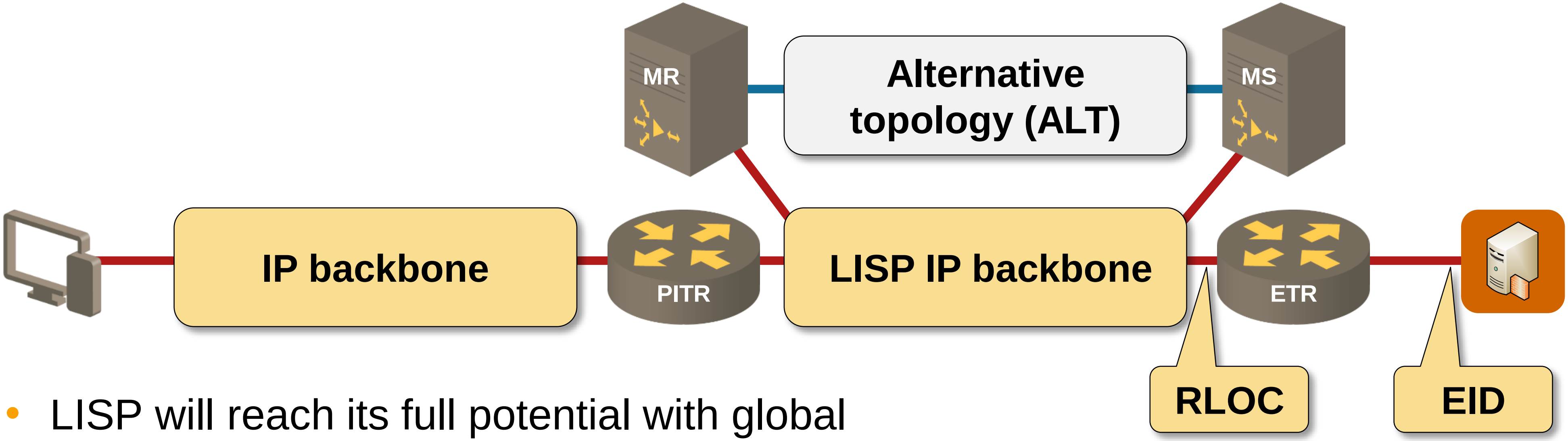
- ETR registers EID-to-RLOC mapping with MS
- Mapping is propagated throughout the ALT backbone



Data-driven actions

- ITR receives IP packet addressed to unknown EID
- ITR sends Map-Request to local MR
- MR forwards Map-Request onto ALT topology
- Map-Request reaches ETR
- ETR responds with Map-Reply (Map-Reply can be based on ITR location)
- Map-Reply reaches ITR
- ITR installs the reply into local LISP EID-to-RLOC mapping cache

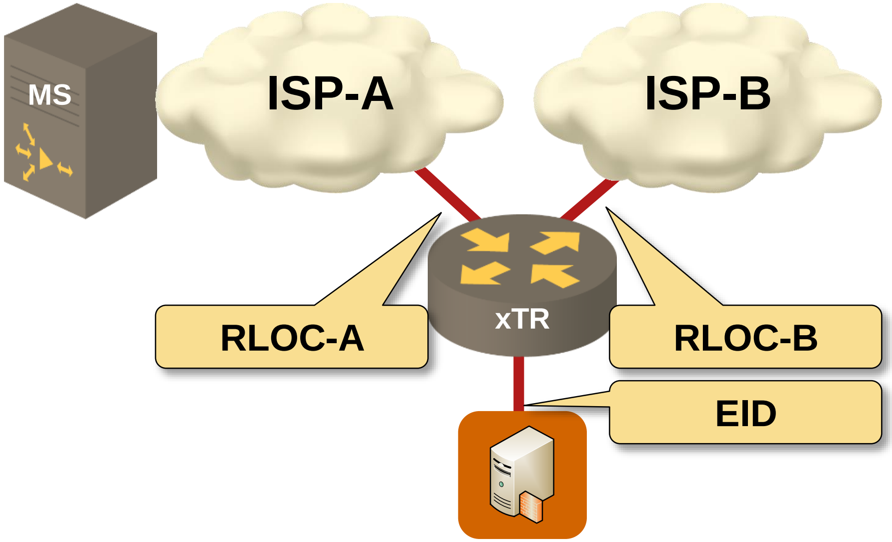
LISP Proxy Services



- LISP will reach its full potential with global deployment (every CE-router is an ITR)
- Local LISP deployment relies on proxy services
- PITR advertises EID prefixes into non-LISP IP backbone to attract traffic
- PITR performs IP-to-LISP translation
- Return traffic can flow through PITR, a dedicated PETR, or directly
- LISP and non-LISP IP traffic can use the same IP backbone

Multihoming with LISP

- Customer's xTR registers two EID-to-RLOC mappings
- RLOCs belong to ISP's PA space
- No BGP needed between customer and ISPs



Drawbacks

- Doesn't solve the fundamental problem
- Address table explosion is moved to another domain
- Requires widespread LISP deployment or external xTRs

It is easier to move a problem around than it is to solve it (RFC 1925, section 6)

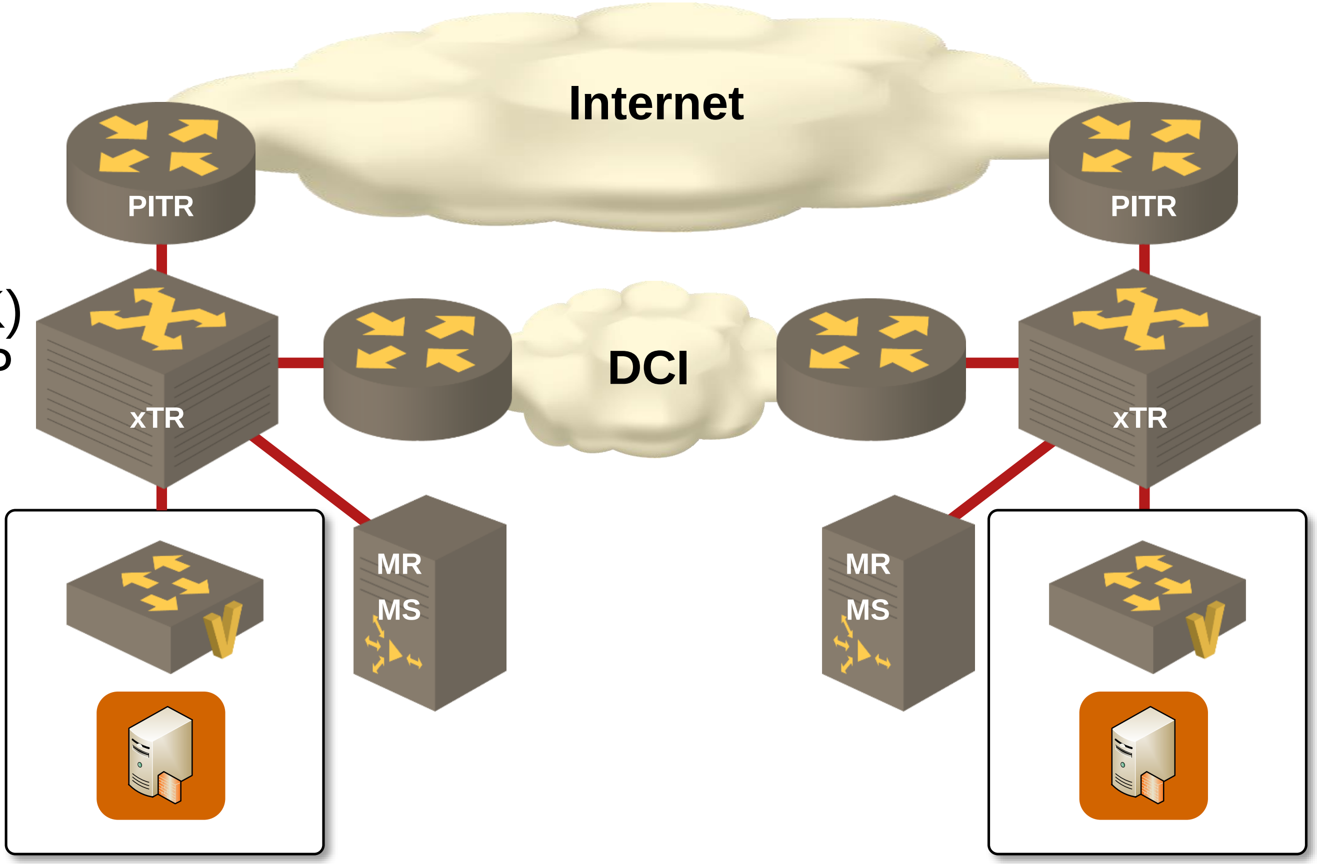
It is always possible to add another layer of indirection (RFC 1925, section 6a)

LISP in the Data Center

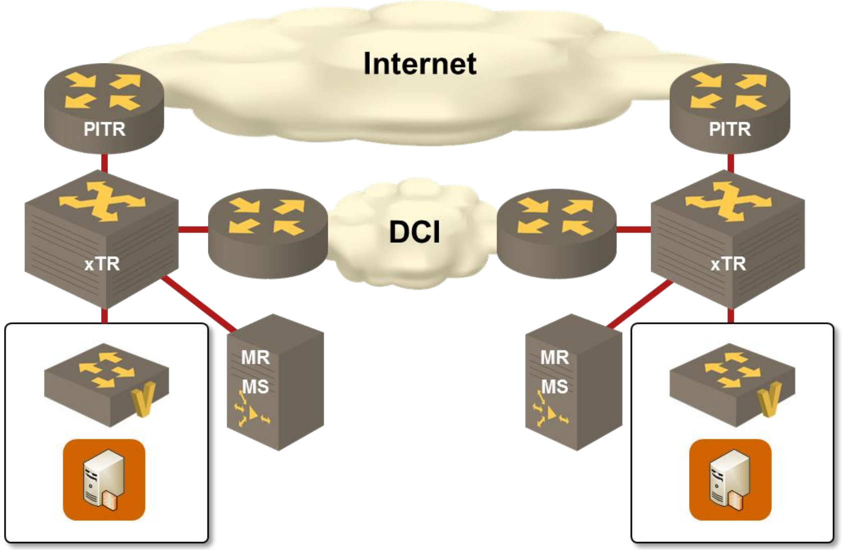
Nexus 7000 = ETR

DC edge router = ITR

- Layer-3 switch (Nexus 7K) registers off-subnet VM IP addresses with MS
- LISP mappings change after vMotion event
- L3 (LISP) transport between data centers
- No L2 DCI
- Internet multihoming is still required



DC LISP Caveats



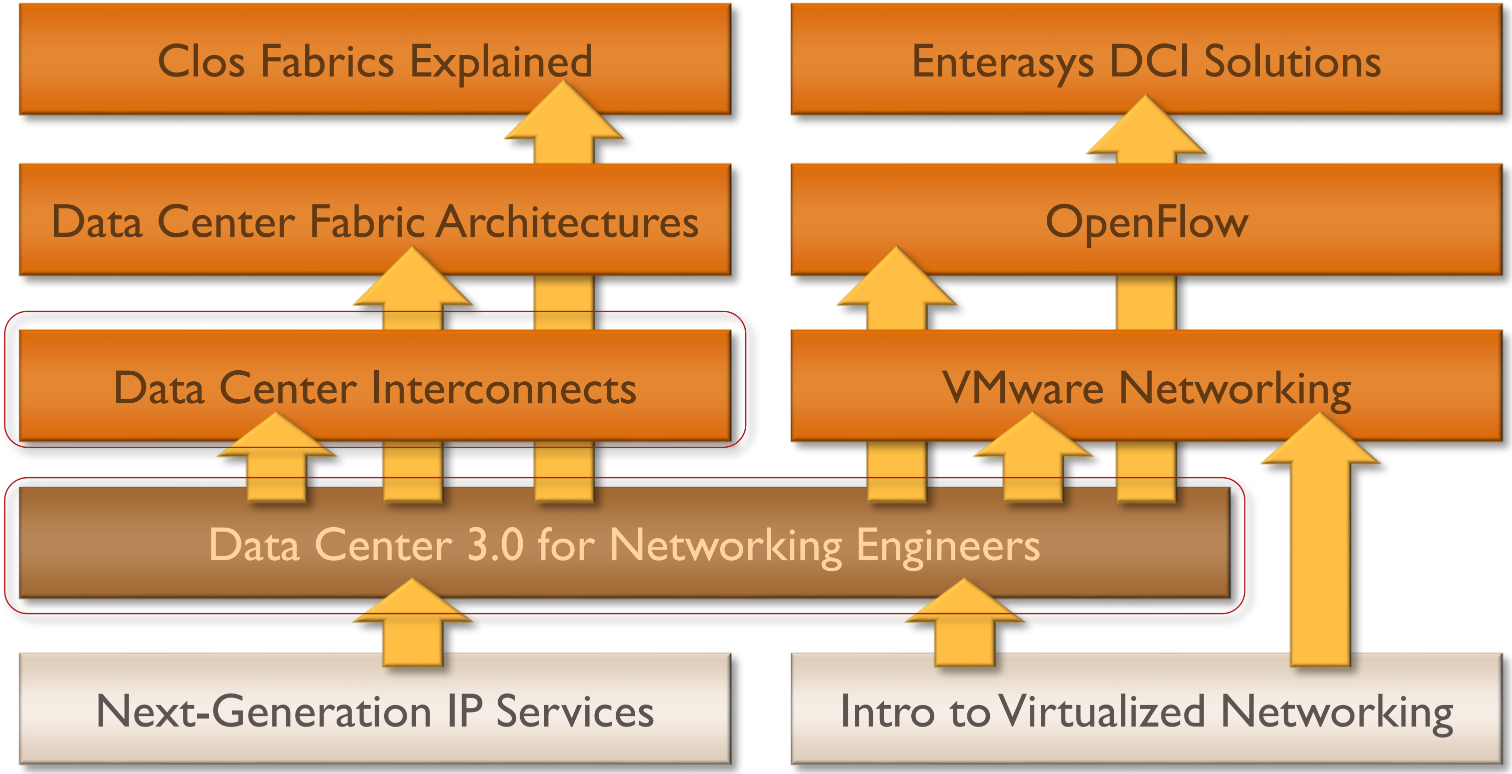
Traffic flow issues

- LISP with DC PITR does not solve the ingress traffic trombone problems
- Remote ITR is required to get optimal ingress routing
- Output traffic flow is optimal

Scalability

- EID prefix = host route (VM IP address)
- PITR EID-to-RLOC cache entry must expire soon after vMotion event
- Low TTL must be set on LISP mappings
- High volume of Map-Requests from PITRs
- Potential TCAM overflow on PITR

Data Center Webinars on ipSpace.net



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

End-to-End High Availability

Remember the Sequential Address Family Retries?

Socket API

```
memset(&hints, 0, sizeof(hints));
hints.ai_family = PF_UNSPEC;
hints.ai_socktype = SOCK_STREAM;
error = getaddrinfo("example.com", "http", &hints, &res0);
if (error) { errx(1, "%s", gai_strerror(error)); }

s = -1;
for (res = res0; res; res = res->ai_next) {
        s = socket(res->ai_family, res->ai_socktype, res->ai_protocol);
        if (s < 0) { cause = "socket"; continue; }

        if (connect(s, res->ai_addr, res->ai_addrlen) < 0) {
                cause = "connect";
                close(s);
                s = -1;
                continue;
        }

        break;  /* okay we got one */
}
if (s < 0) { err(1, "%s", cause); }
```

Dual Stack Brokenness

	Firefox	Firefox fast-fail	Chrome	Opera	Safari	Explorer
MAC OS X 10.7.2 8.0.1	8.0.1	16.9.912.41 b	11.52	5.1.1	-	
	75s	**0ms**	**300ms**	**75s**	**270ms**	-
Windows 7	8.0.1	8.0.1	15.0.874.121 m	11.52	5.1.1	9.0.8112.16421
	21s	**0ms**	**300ms**	**21s**	**21s**	21s
Windows XP	8.0.1	8.0.1	15.0.874.121 m	11.52	5.1.1	9.0.8112.16421
	21s	**0ms**	**300ms**	**21s**	**21s**	21s
Linux 2.6.40.3-0.tc15	8.0.1	8.0.1	16.9.912.41 b	11.60 b	-	
	96s	**0ms**	**300ms**	**189s**		
iOS 5.0.1	-	-	-	-	?	-
					720ms	

Source: http://www.potaroo.net/ispcol/2011-12/esotropia.html

Dual Stack Brokenness

Traditional approach: prefer IPv6 over IPv4

- Fails miserably (after TCP timeout) in broken IPv6 environments
- No fast fallback to IPv4
- Coded in most well-written applications

Happy Eyeballs approach

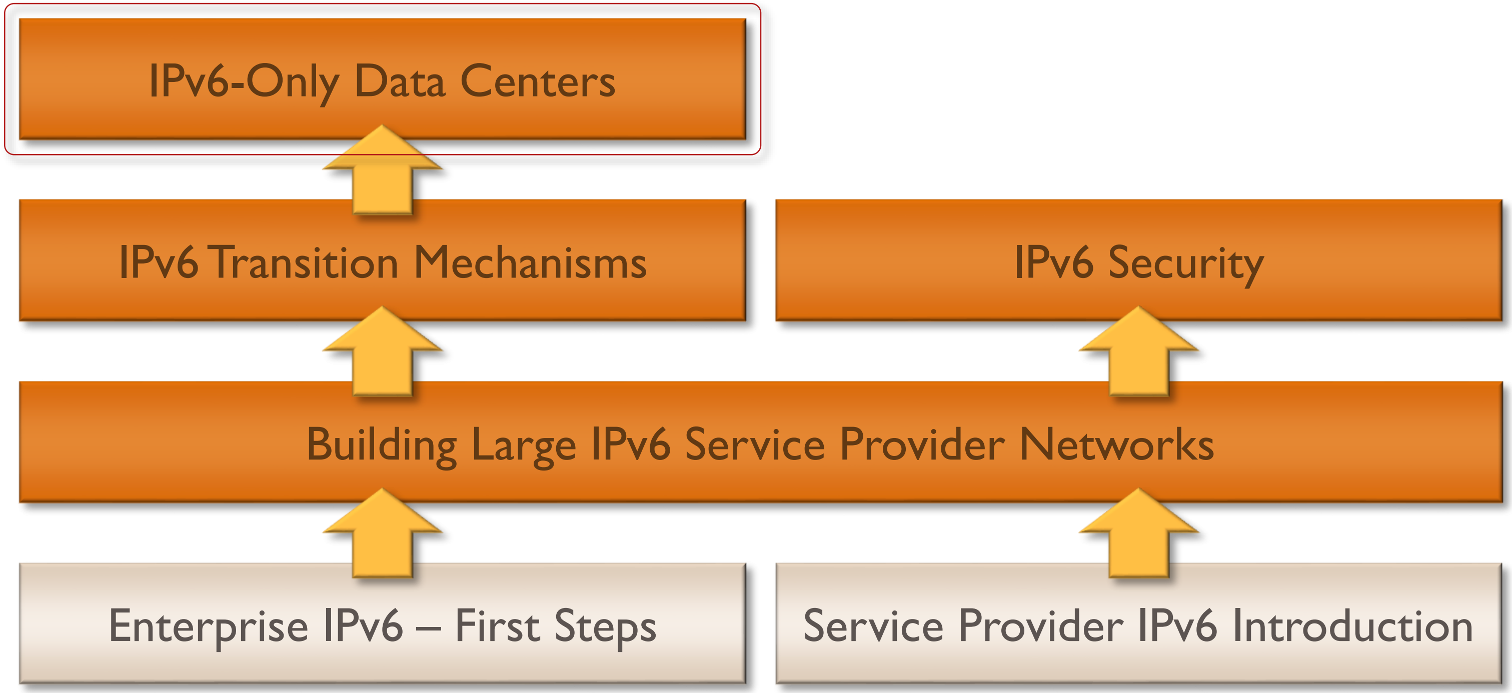
- IPv4 and IPv6 sessions established (almost) in parallel
- Inherently non-deterministic
- Tests session establishment, not data flow
- PMTUD brokenness is not detected

Network services considerations

- IPv4 and IPv6 services and filters are usually configured separately

Avoid complex dual-stack environments

IPv6 Webinars on ipSpace.net



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

Conclusions

Conclusions

- Minor differences between IPv4 and IPv6 HA solutions
- Fundamental problems are unsolved
- Dual-stack environments with happy eyeballs are inherently non-deterministic
- Avoid the complexity of dual-stack environments whenever possible ➔ consider IPv6-only data center

Questions?

Send them to ip@ipSpace.net or @ioshints