Scaling the Address Resolution Protocol for Large Data Centers (SARP)

draft-nachum-sarp-04

Youval Nachum	Marvell
Linda Dunbar	Huawei
Ilan Yerushalmi	Marvell
Tal Mizrahi	Marvell

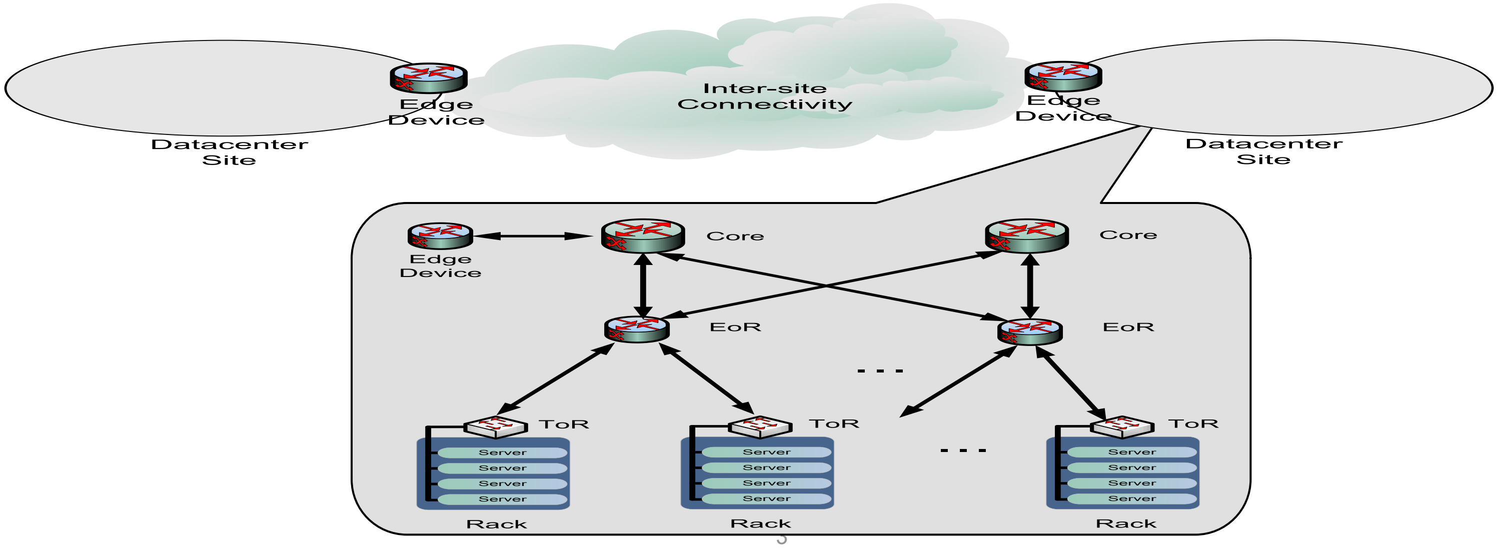
IETF Meeting 86, March 2013

History of this Draft

- **March 2012 – draft 00.**
- **Discussion in ARMD mailing list.**
- **July 2012 – IETF 84 – presented in INTAREA WG.**
  - Main feedback: need to equally address IPv4 and IPv6.
- **October 2012 – draft 03.**
  - More details about SARP with IPv6.
- **March 2013 – draft 04:**
  - Address issues discussed at mailing list

Perceived issues associated with subnets spanning across multiple L2/L3 boundary router ports:

- ARP/ND messages are flooded to many physical link segments which can reduce bandwidth utilization for user traffic;
- the ARP/ND processing load impact on L2/L3 boundary routers;
- intermediate switches exposed to all host MAC addresses which can dramatically increase their FDB size;
- In IPv4, every end station in a subnet receives ARP broadcast messages from all other end stations in the subnet. IPv6 ND has eliminated this issue by using multicast.

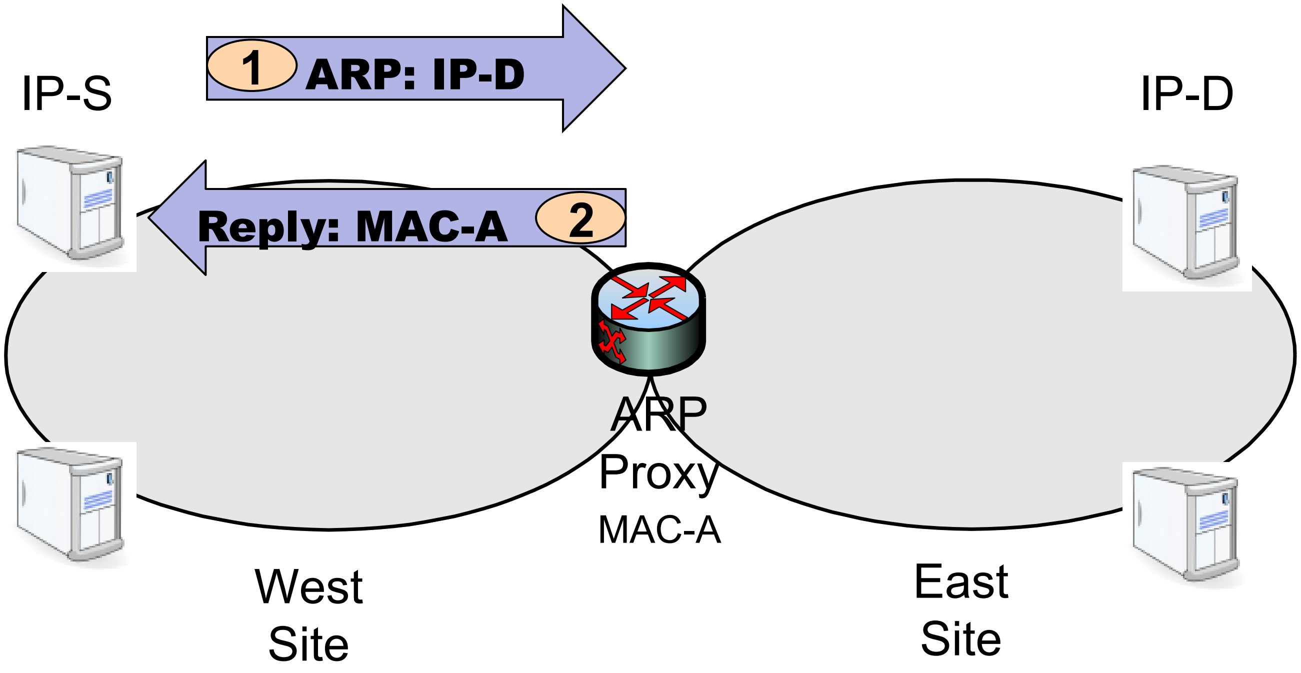


Real Impacting Issues?

- **As majority of servers move towards 1G/10G links, the traffic taken by ARP/ND broadcast/multicast becomes less significant**
  - ARP/ND messages are flooded to many physical link segments which can reduce bandwidth utilization for user traffic;
- **the ARP/ND processing load impact on L2/L3 boundary routers;**
  - [ARMD-Statistics] has shown that the major impact of large number of mobile VMs in Data Center is to the L2/L3 boundary routers.
  - Dual stack makes it worse
- **intermediate switches being exposed to all host MAC addresses which can dramatically increase their FDB size;**
- **Today's servers only need <2% CPU to process 2000/s ARP i.e. impact to Server is insignificant**
  - In IPv4, every end station in a subnet receives ARP broadcast messages from all other end stations in the subnet. IPv6 ND has eliminated this issue by using multicast.

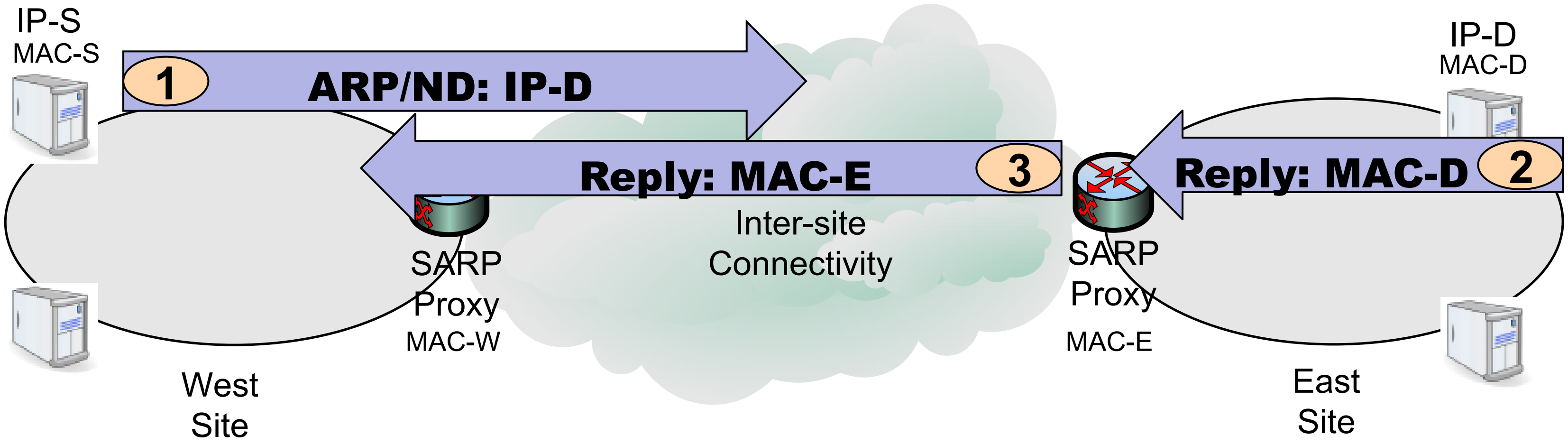
Background – Proxy ARP

- **Proxy ARP (RFC 1027, RFC 1009, RFC 925).**
- **Proxy ARP responds based on IP subnet.**
  - Assumption: IP subnet implies location.

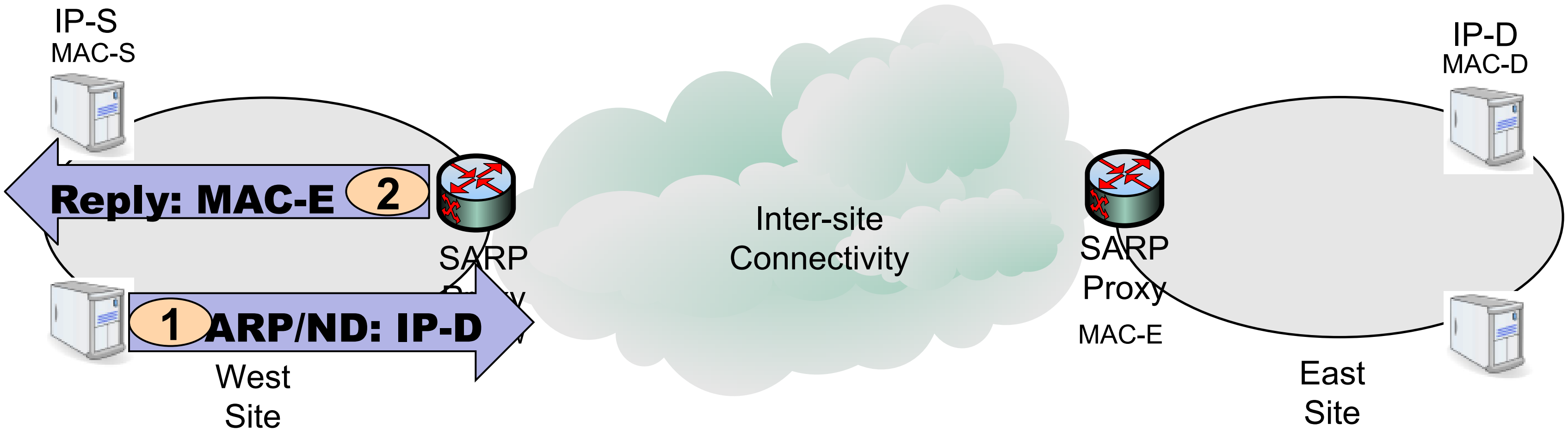


SARP

- **Edge devices: proxy SARP.**
- **IP subnet does not imply location.**
- **MAC-W / MAC-E imply location.**



SARP Cache



IP-S
MAC-S

Reply: MAC-E 2

1 ARP/ND: IP-D

SARP

West
Site

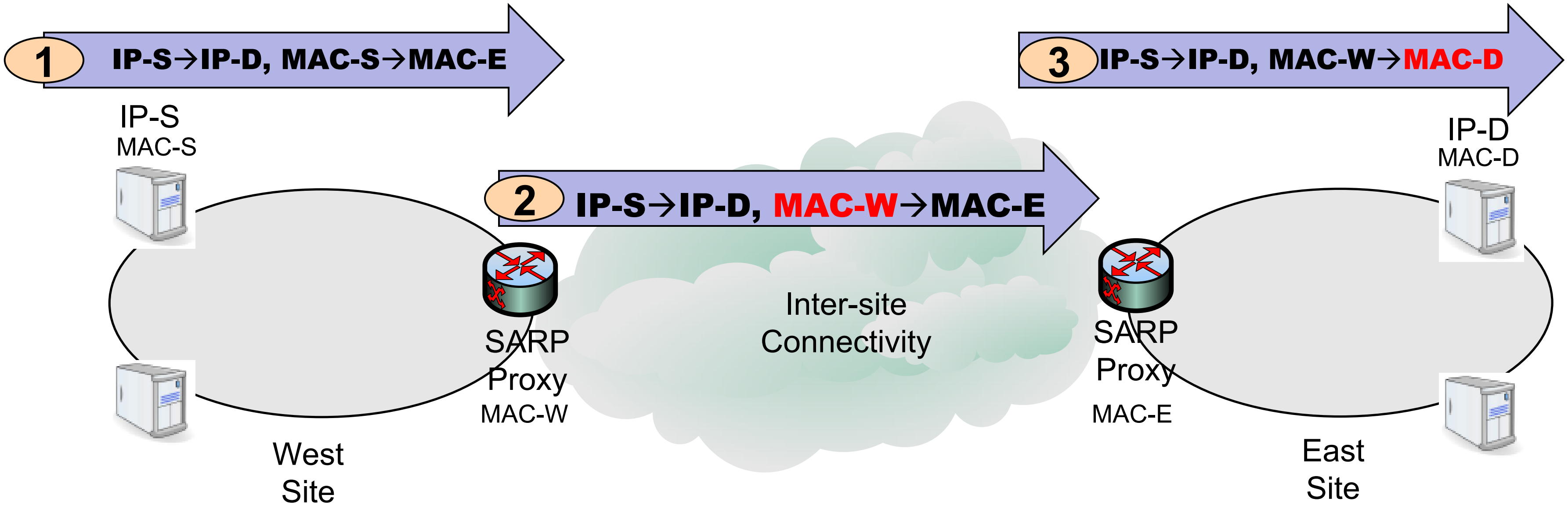
Inter-site
Connectivity

SARP
Proxy
MAC-E

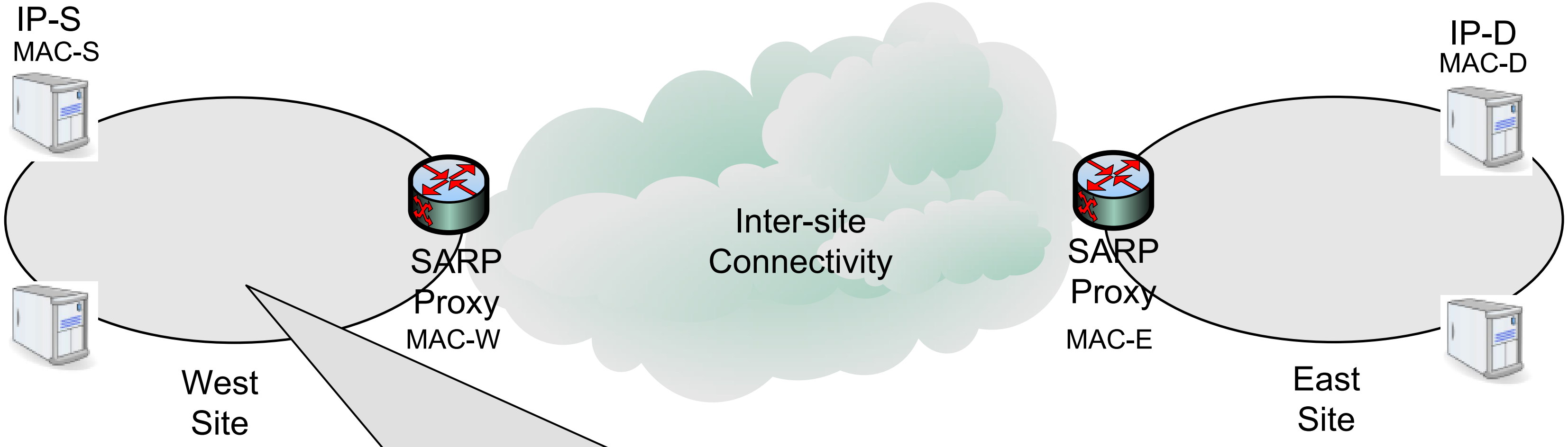
IP-D
MAC-D

East
Site

SARP – Data Plane



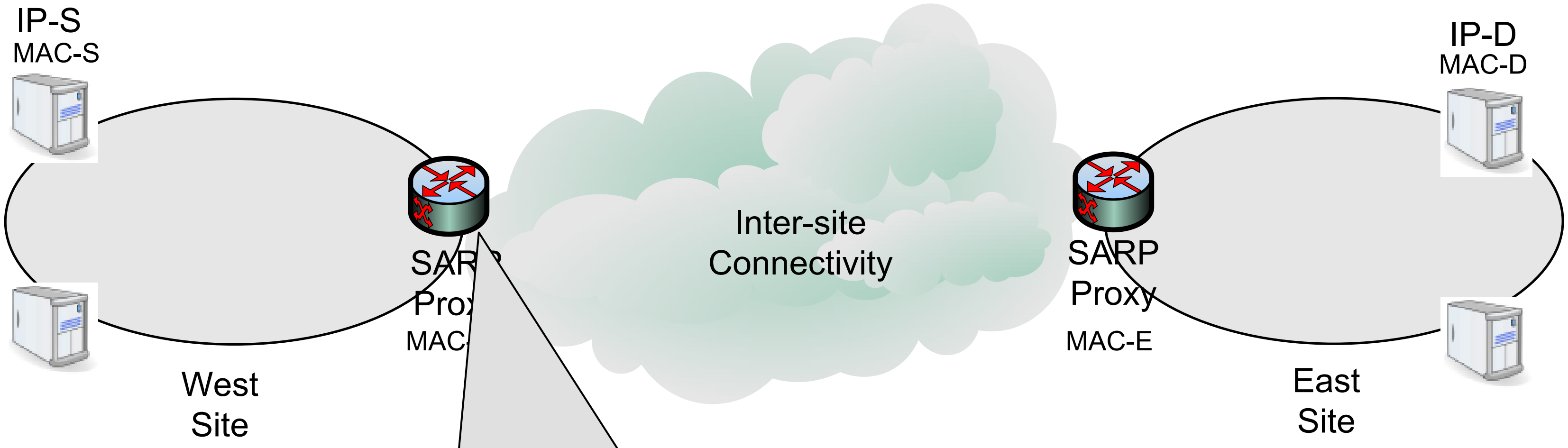
SARP – MAC Address Tables



MAC address table of bridges in the west site:

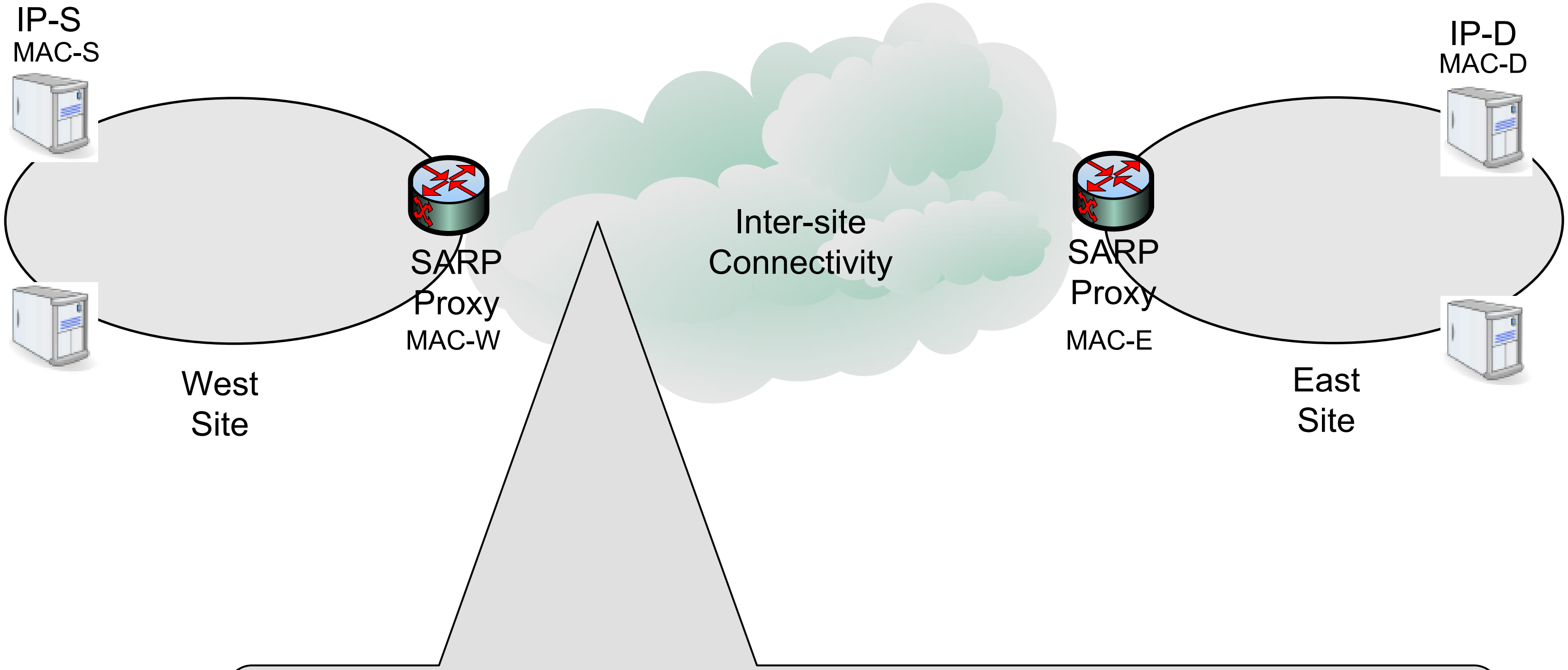
- **Local site addresses, e.g., MAC-S.**
- **Edge devices, e.g., MAC-E.**
- **No need for addresses of remote sites.**

SARP – ARP Broadcast Domains



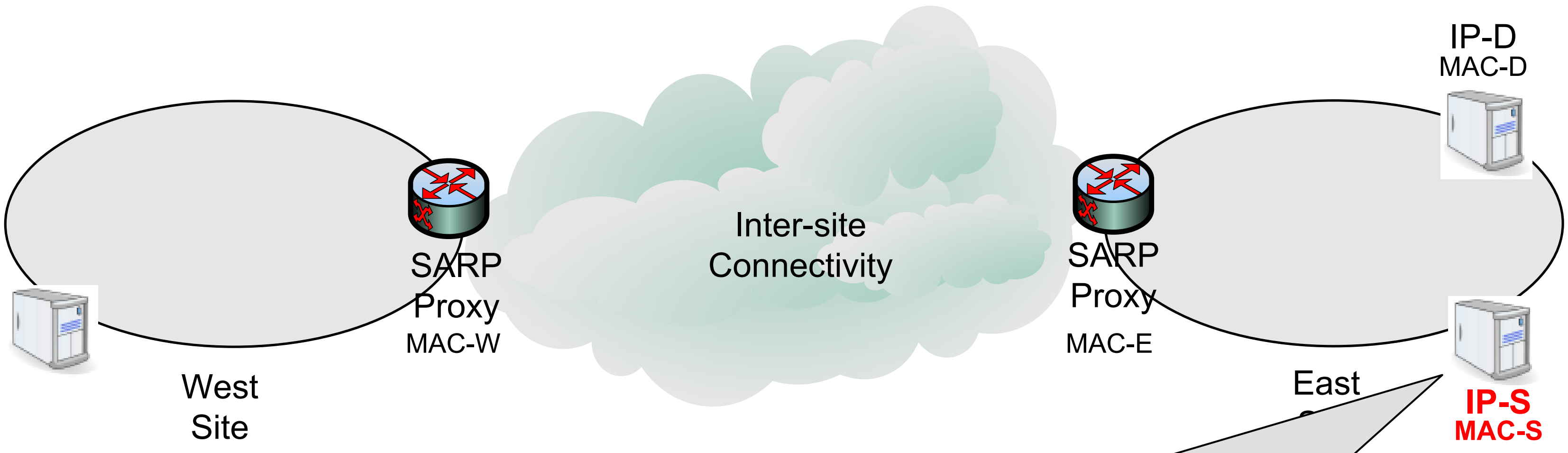
Local SARP cache limits broadcast domain for known IP addresses.

SARP over Overlay Network



SARP is agnostic to the transport technology, e.g. L2VPN.

SARP with VM Migration



- **IPv4: Gratuitous ARP is used to notify network about migration.**
- **IPv6: unsolicited neighbor advertisement is used.**

- **No need for additional control protocols.**
- **Transparent to inter-site network and protocols.**

Next Steps

- **Receive feedbacks from WG.**
- **WG adoption.**

Thanks