

Danger of Proxy ARP in IX environment

version 0.3

Maksym Tulyuk maksym.tulyuk@ams-ix.net RIPE 63. Vienna 03 November 2011

Content



- 1. The Outage
- 2. Analysis and improvements
- 3. Useful tools and procedures

Danger of Proxy ARP in IX environment

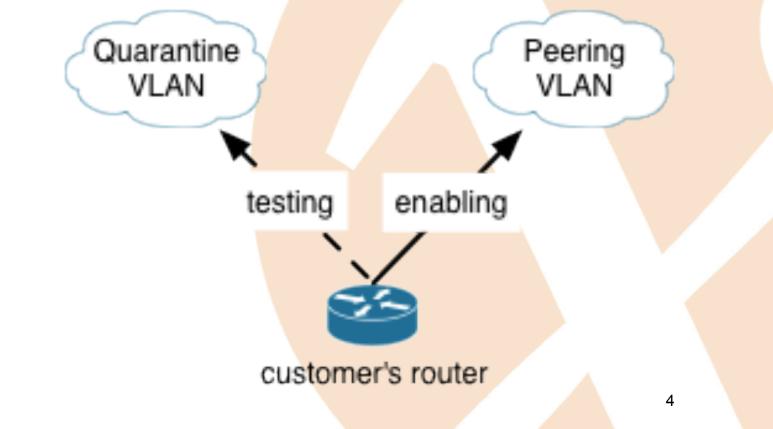


1. The Outage

Maksym Tulyuk maksym.tulyuk@ams-ix.net RIPE 63. Vienna 03 November 2011

1.1. Customer in production

- 1. A customer's router was moved from the Quarantine network to the Production network
- 2. This is a standard procedure was done many times



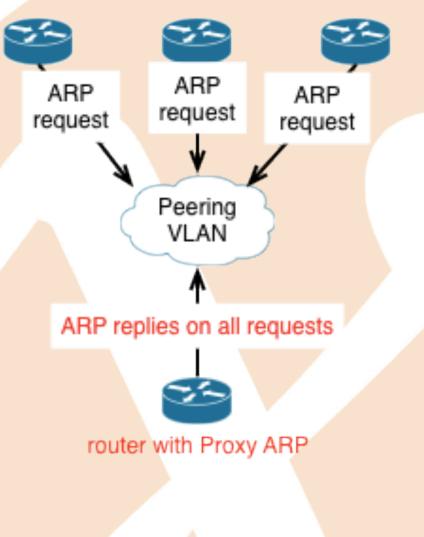


1.2. Proxy ARP in action

Proxy ARP - router sends ARP replies from own mac address

Proxy ARP conditions

- 1. Proxy ARP was enabled on a customer's router
- 2. The router had no IP address from IX range
- 3. The default route was set on the router



5

1.3. BGP sessions down

Flapping of BGP sessions was only one visible thing of the accident

Aug 11 11:15:19 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:15:42 BGP: Peer 195.69.xxx.xxx DOWN Aug 11 11:16:31 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:16:58 BGP: Peer 195.69.xxx.xxx DOWN Aug 11 11:17:06 BGP: Peer 195.69.xxx.xxx DOWN Aug 11 11:19:19 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:21:42 BGP: Peer 195.69.xxx.xxx DOWN Aug 11 11:23:54 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:24:30 BGP: Peer 195.69.xxx.xxx DOWN Aug 11 11:26:30 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:26:51 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:27:48 BGP: Peer 195.69.xxx.xxx UP (ESTABLISHED) Aug 11 11:28:02 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:28:15 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:28:36 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:29:06 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:29:12 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) Aug 11 11:29:17 BGP: Peer 195.69.xxx.xxx UP (ESTABLISHED) Aug 11 11:30:43 BGP: Peer 195.69.xxx.xxx DOWN (TCP Connection Closed by Remote)

(Rev Notification:Hold Timer Expired) (Rcv Notification:Hold Timer Expired) (Rcv Notification:Hold Timer Expired) (Rev Notification:Hold Timer Expired) (Rev Notification:Hold Timer Expired)

amsterdam internet exchange

ams

Aug 11 11:31:02 BGP: Peer 195.69.xxx.xxx DOWN (Rcv Notification:Hold Timer Expired) (TCP Connection Closed by Remote) <.xxx DOWN (TCP Connection Closed by Remote) <.xxx DOWN (Rev Notification:Hold Timer Expired)</p> <.xxx DOWN (TCP Connection Closed by Remote)

Aug 11 11:14:23 bqpd[6123]: neighbor 195.69.xxx.xxx: state change Established -> Idle, reason: NOTIFICATION received Aug 11 11:14:24 bgpd[6123]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0 Aug 11 11:14:24 bqpd[6123]: neighbor 195.69.xxx.xxx: state change Established -> Idle, reason: NOTIFICATION received Aug 11 11:14:38 bgpd[12947]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0 Aug 11 11:14:38 bqpd[12947]: neighbor 195.69.xxx.xxx: state change Established -> Idle. reason: NOTIFICATION received Aug 11 11:14:45 bgpd[12947]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0 Aug 11 11:14:45 bgpd[12947]: neighbor 195.69.xxx.xxx: state change Established -> Idle, reason: NOTIFICATION received Aug 11 11:14:52 bgpd[6123]: neighbor 195.69.xxx.xxx: state change Idle -> Active, reason: Start Aug 11 11:14:53 bgpd[6123]: neighbor 195.69.xxx.xxx: state change Idle -> Active, reason: Start Aug 11 11:15:07 bapd[12947]: neighbor 195.69.xxx.xxx: state change Idle -> Active. reason: Start Aug 11 11:15:15 bgpd[12947]: neighbor 195.69.xxx.xxx: state change Idle -> Active, reason: Start Aug 11 11:15:46 bgpd[6123]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0 Aug 11 11:15:46 bapd[6123]: neighbor 195.69.xxx.xxx: state change Established -> Idle, reason: NOTIFICATION received Aug 11 11:16:06 bqpd[12947]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0 Aug 11 11:16:06 bapd[12947]: neighbor 195.69.xxx.xxx: state change Established -> Idle. reason: NOTIFICATION received Aug 11 11:16:06 bapd[12947]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0 Aug 11 11:16:06 bgpd[12947]: neighbor 195.69.xxx.xxx: state change Established -> Idle, reason: NOTIFICATION received Aug 11 11:16:16 bapd[6123]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0 Aug 11 11:16:16 bqpd[6123]: neighbor 195.69.xxx.xxx: state change Established -> Idle, reason: NOTIFICATION received Aug 11 11:16:16 bgpd[6123]: neighbor 195.69.xxx.xxx: state change Idle -> Active, reason: Start Aug 11 11:16:17 bapd[6123]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0 Aug 11 11:16:17 bqpd[6123]: neighbor 195.69.xxx.xxx: state change Established -> Idle, reason: NOTIFICATION received Aug 11 11:16:28 bapd[12947]: neighbor 195.69.xxx.xxx: received notification: HoldTimer expired, unknown subcode 0

Aug 11 11:14:23 bapd[6123]; neighbor 195.69.xxx.xxx; received notification: HoldTimer expired, unknown subcode 0

6



1.4. Troubleshooting

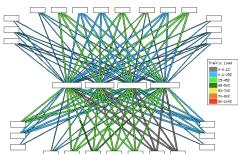
Hypotheses during outage

- 1. Platform issue?
- 2. BGP issue?
- 3. Cisco issue?
- 4. Route-server issue?

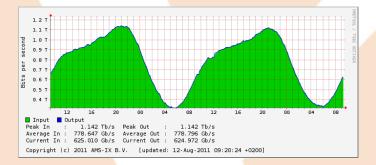
1.4.1. Platform issue?



- 1. No interfaces down it's not Physical layer
- 2. No packet losses it's not Data Link layer
- 3. Small drops on graph it's not Network layer
- 4. Only BGP sessions down something on Transport or BGP layer?



	No TRANSPORT NAME	Rig Tal Mag House and	No. To black the born of the	Mig Taking at a set of the	RA 14 M & R & M & M & M & M & M & M & M & M &	No. TALANCE MANYAN	No Territoria de la caración
79.07	3		2	1			2
X	12	12	1	12	1	12	1 -
	Rt ffe ffe ffe	Re ofe die die	ta din bie die	Rt die He die	tet tin He ife	tta file file file	Ra dia Ma dia
National Action 10		Colorester 1	No feed and a strain of	and the restor data strained at	ala ha crimita si si si	afa frank an la k srightly	Mig Formation data and an
1		1 5	1	15	18		15
18	V		18 1	18			12
Aus also tite alle		the sin tin sin	an die He die	an die Un fin	the Ale His site		na ifa fia fia
		NUC USE UAL DESCRIPTION					
No. Toronton and A strength	- Nata and Arabit			and the second s		Ming Top and an At & Strappents	and the balance of a set of
10 1	18	3.0	2	12			
	" and and and and	I X .				" NOT NOT NOT NOT"	ANY ANY ANY
ALL ALL ALL ALL ALL	an An He He		ate alle alle alle	at de lle lle	ter die He He	tes die His His	Rt Jo Jo Jo
. No. 100 (100 (100 (100 (100 (100 (100 (100	Mig Terrology Date of a real	RATE OF BUILDED		Mig Terrology (27 to 12 million)	RATE OF BUILDING	10 10 10 10 10 10 10 10 10 10 10 10 10 1	Ma harolar to to the
1	15	1			1.5	1-	18
13 1	1	1	X	12			1.5
Ra Ha Ha Ha	RA LES MA MA	Rez Alia Alia Alia	~	Ra Ja Ha Ha	Res alle Min Alle	And Alia His His	na dia ilia ilia
				And Mile Ster Ster Ster		And Barry Mars Street	
-	and has served as a served as	-	and the last of the last of the				Mig Taxantal All A series
18 18	1		18	V	-		18
" was not not not"	" NOT NOT NOT NOT	**** *** *** ****	" NAS NOT NOT NOT		" and all and all	" was not not not"	And \$10 \$25
tta din lin din	na dia Madia	tes sin sin sin	tta Bin Hin Hin			tes Ma Els Ma	no dia Un dia
Martin of Article III	Market of Articles and Articles	RATINGTONICS	Martin of Article III	Market of Articles of Articles		Martin of the state of	Ma hardness of a set of
18	18			10			18
18	1	12	12		I X	13	12
Ra Ha Ha Ha	RE ALS HE ALS	tez alle alle alle	te fin lin fin	RE ME ME ME		tes the life file.	Re die He die
No local de la companya de	No local di sense di secondo della solo	En la seconda de la compositione de la composition de la compositi	and a second sec	No local di la segui di la segui di	and have placed as an of	ACCR A REPORT OF A REPORT OF A REPORT	No beneficial and a
12	15	1 -	a statement of the second seco		12		1 -
1	1			-		V	-
An An He Me	an die He die	an die Uie Ale	AND ARE MELATE	An An He He	tes die Hie Hie		to Ja He He
							37.220.52P.225
No TO MANDA MARK	No Torona or to come	1	No TO MARK OF MILE	Mig Torrestanter & Kit Mills	84740400000	No TO LA DALLA	
-	1		1	18			3.0
tes dis Ma His	RE ALS HE HE	Ra fin fin fin	Ra file IIs IIs	Ra dia Ma dia	Re die die die	ALL DIE DIE DIE	I X



1.4.2. BGP issue?



- At least, 111 BGP peers were down available statistics only from our equipment (route-servers, routers)
- 2. 88 peers with route-servers were down
- 3. 57 peers with our routers were down
- 4. Most of mac addresses are ... Cisco

1.4.3. Cisco issue?



- 1. 90% affected routers were Cisco
- 2. Cisco's "bad fame" Monday's presentation
- 3. RIPE "experiment": https://labs.ripe.net/Members/erik/ripe-ncc-and-duke-university-bgp-experiment
- 4. 10% routers of **other** vendors

1.4.4. Route-server issue

- 1. Our route-servers were unstable
- /var/log/messages: "arp info overwritten for [IP] by [MAC] on [interface]" – mac address is the same all the time
- 3. Call from a customer: "router with [MAC] is hijacking IP addresses!"

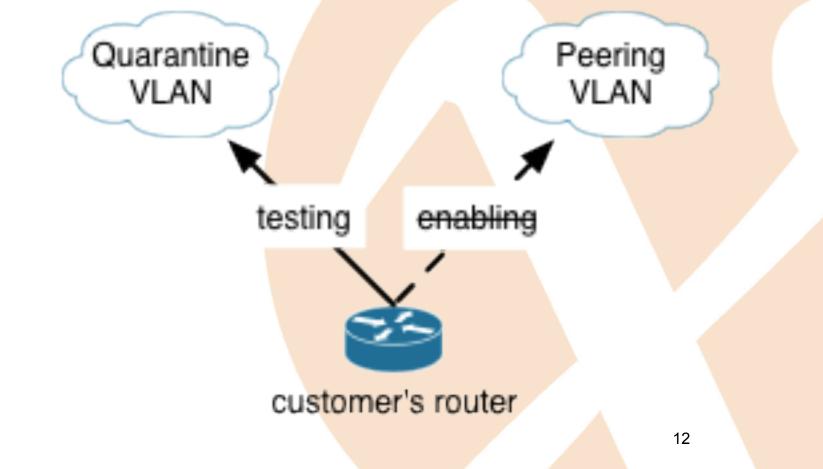
#sh arp mac-address yyyy.yyyy.yyyy

	IP Address	MAC Address	Туре	Age	Port
1	195.69.xxx.xxx	.	Dynamic	4	1/2
2	195.69.xxx.xxx	уууу .уууу .уууу	Dynamic	5	1/2
3	195.69.xxx.xxx	уууу .уууу .уууу	Dynamic	24	1/2
4	195.69.xxx.xxx	уууу .уууу .уууу	Dynamic	5	1/2
5	195.69.xxx.xxx	уууу •уууу •уууу	Dynamic	0	1/2
6	195.69.xxx.xxx	уууу .уууу .уууу	Dynamic	24	1/2
7	195.69.xxx.xxx	уууу •уууу •уууу	Dynamic	4	1/2
8	195.69.xxx.xxx	уууу .уууу .уууу	Dynamic	5	1/2
9	195.69.xxx.xxx	уууу •уууу •уууу	Dynamic	13	1/2
10	195.69.xxx.xxx	уууу .уууу .уууу	Dynamic	19	1/2
11	195.69.xxx.xxx	уууу .уууу .уууу	Dynamic	8	1/2
12	195.69.xxx.xxx	уууу •уууу •уууу	Dynamic	5	1/2

1.5. Disable port



- 1. The port with the broken router was disconnected
- 2. It was disconnected 3 (three) times good team work ©

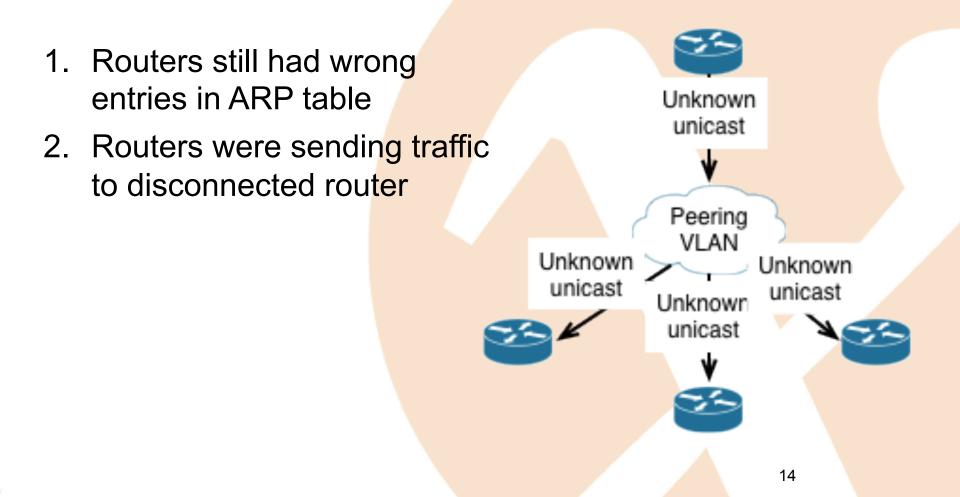


1.6. Calls from customers

- 1. 10 calls from customers (1 call per 4 min)
- 2. No network tickets were sent 😕
- 3. Confirmation that the right router was disconnected

1.7. Unknown unicast flood

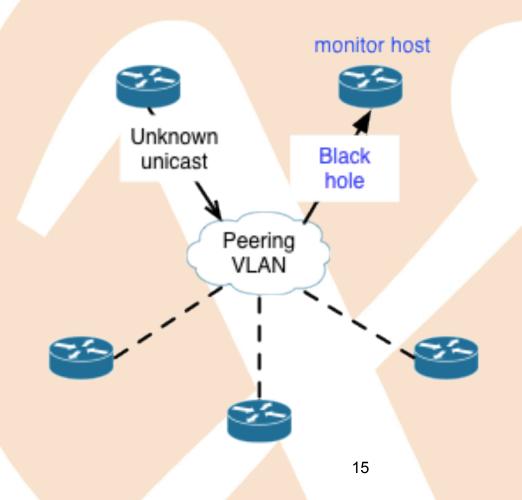
Disabling the broken router caused "unknown unicast" flood



1.7. Unknown unicast flood

Solved by adding wrong mac-address on a monitor host

- 1. It was quickly detected
- 2. Didn't affect any customer's routers
- 3. Made black hole





1.8. Update mac tables

Despite disconnecting the broken router, all affected routers had wrong mac addresses in their ARP tables

Two ways to fix: 1) Reactive approach

- 1. Do nothing TCP/IP already has mechanism to fix
- 2. When timer "ARP cache timeout" is expired, ARP table will be automatically updated
- 3. AMS-IX recommends to set **4 hours** for "ARP cache timeout"
- 4. In the worst case, it would take 4 hours until full restoration



1.8. Update mac tables

Despite disconnecting the broken router, all affected routers had wrong mac addresses in their ARP tables

Two ways to fix: 2) Proactive approach

- 1. Send spoofed ARP request (unsolicited ARP reply, gratuitous ARP, ARP request)
- 2. Proper mac addresses were got from ARP sponge (part of our monitoring system)

1. The Outage



- 1. Customer in production
- 2. Proxy ARP in action
- 3. BGP sessions are down
- 4. Troubleshooting
- 5. Disable port
- 6. Calls from customers
- 7. Unknown unicast flood
- 8. Update mac tables



1. The Outage

- 1. Customer in production
- 2. Proxy ARP in action
- 3. BGP sessions are down
- 4. Troubleshooting
- 5. Disable port
- 6. Calls from customers
- 7. Unknown unicast flood
- 8. Update mac tables

Downtime: **40 min** Why so long?

Danger of Proxy ARP in IX environment



2. Analysis and improvements: Root causes

Maksym Tulyuk maksym.tulyuk@ams-ix.net RIPE 63. Vienna 03 November 2011

2. Root causes of outage amsix

- 1. No tests for Proxy ARP were made
- 2. No confirmation permanent IP address was set

The issue happened because 3 conditions had been met:

- 1. Proxy ARP on customer's router
- 2. No IP address from IX range
- 3. Customer's router had default router

If any of them wasn't met, the issue wouldn't have happened

2.1. Proxy ARP tests



Issue: the new router wasn't checked for Proxy ARP

Solution 1: Test all new connected routers if Proxy ARP is disabled on them.

Fix: proxy ARP **test was added** to internal procedure; all engineers were informed **Solution 2**: Periodic check all connected routers if Proxy ARP is disabled on them

Fix: tool checks connected routers once per day

2.2. Confirmation



Issue: no confirmation permanent IP address was set

Idea:

- 1) Customer sets up an IP address from IX range
- 2) Engineer checks if the IP address was set
- 3) The router is moved to the production network

Discussion...

Danger of Proxy ARP in IX environment



2. Analysis and improvements: Lengthy troubleshooting

Maksym Tulyuk maksym.tulyuk@ams-ix.net RIPE 63. Vienna 03 November 2011

2. Lengthy troubleshooting

- 3. Duty engineer didn't have enough information
- 4. Customers weren't informed on time
- 5. No roles between engineers
- 6. No logs from kernel

In short

- 1. Internal procedures were not clear enough
- 2. Tools didn't provide enough functionality

2.3. Full information



Issue: duty engineer didn't have enough information because enabling the new link was done by engineer from **the previous shift**

Fix: everyone has to inform duty engineer what and when he is going to do

Discussion: tool to deal with planned works

2.4. On time information



Issue: customers weren't informed on time

Solution: a tool allows to open and send a network ticket quickly and easily

Development:

- 1. Option in the old interface to send network ticket quickly
- 2. New web interface under testing



2.5. Outage procedure

Issue: no roles between engineers caused unnecessary delays

Development and discussion...

2.6. Enable logs



Issue: no logs from kernel on syslog server: /bsd: arp info overwritten for [IP] by [MAC] on [interface]

Fix: route-servers **export to the syslog server** messages from BGP daemon **and** ARP messages

2. Analysis and improvements

Root causes

- 1. No tests for proxy ARP were made
- 2. No confirmation permanent IP address was set

Lengthy troubleshooting

- 3. Duty engineer didn't have enough information
- 4. Customers weren't informed on time
- 5. No roles between engineers
- 6. No logs from kernel



2. Analysis of the outage

- 1. No proxy ARP tests were made
- 2. No final confirmation from customer
- 3. Duty engineer didn't have enough information
- 4. Customers weren't informed on time
- 5. No roles between engineers
- 6. No logs from kernel

Why only 40 min?

Danger of Proxy ARP in IX environment



3. Tools and procedures

Maksym Tulyuk maksym.tulyuk@ams-ix.net RIPE 63. Vienna 03 November 2011

3.1. Team work



"Never debug alone":

- 1. 15 min 3 engineers were involved
- 2. 25 min 5 engineers were involved

Advice: if the issue is going more that 15-30 min, escalate to your colleague/2nd line



3.2. Central log server

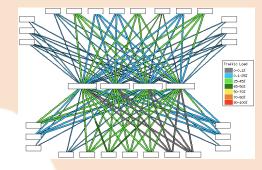


- 1. Logs from all devices are being sent to a log server
- 2. The server is available during outage
- 3. Analysing tools (sed/grep/awk/etc) are installed on server
- 4. Engineers know how to use them

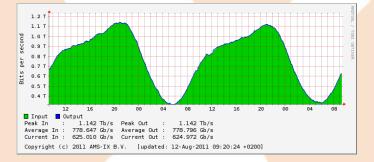
3.3. Visualisation

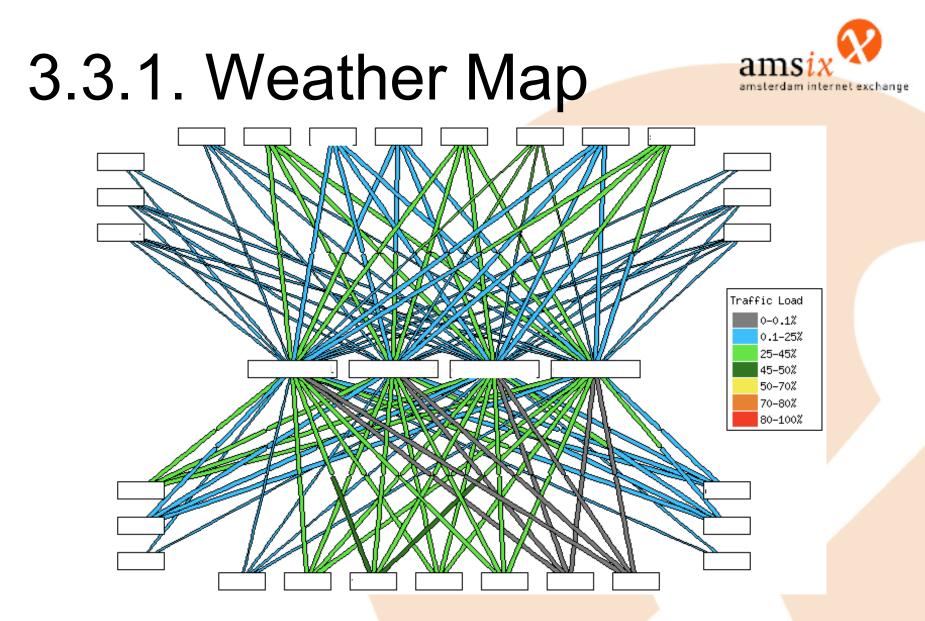


- 1. WeatherMap
- 2. Connectivity matrix
- 3. MRTG



Х		an As Us As	AN AN UN AN			A La Ma Ha Ma	Ro Ma Ba Ma
	Х						
1 : 211 / An 201 / An	I I IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	Х	I IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	I IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	I I III III	Hardin Un Ma	ha a a a a
	References and the	No Constants	Х	References and	AN AN AN AN AN	Revenue and	References
	Rt No. 10 410			Х		Victorial Control of C	
I Jan Ba Ha Ala			AND	1	Х		
A CONTRACTOR OF A CONTRACTOR A C	Re de He He			Re An He He		Х	Ref. To The State
Real of the second							Х





http://www.network-weathermap.com/

3.3.2. Real time matrix

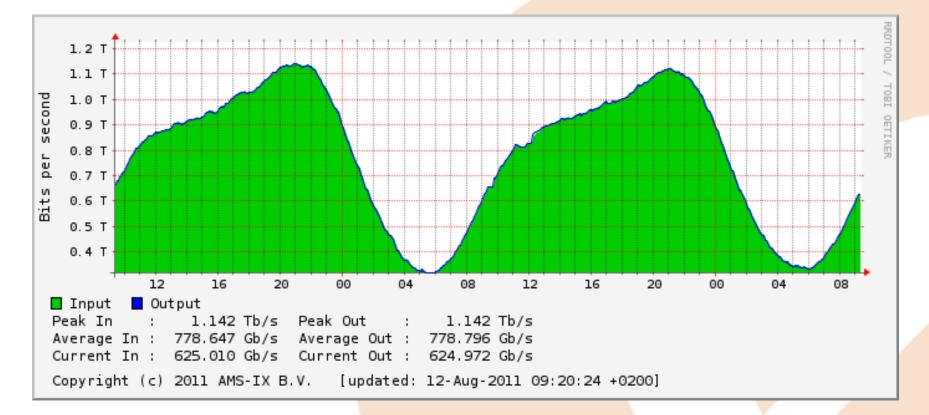


X	Aling from and app 100 to and points a big of the second	After from out any 62 to out out 62		After from out any 60 to out of 62 42 B		Aling from and app tild to add plot till a a a a a a a a a a a a a a a a a a a	Aling from and app till to and and the main second
Mag From sold are lide to real age 0.01 W	X		Mag how and an data to all and data No	Mag Fan Sal (m SB) to sal 64.52 Mag Sal Mag Sal <tr< td=""><td>Along true sol on tild to sol sol at along true sol on tild to sol sol at along true sol at along true sol at along true sol along true sol at along true sol at along true sol at along true sol at along tru</td><td>Aling transmitten die terstigte die</td><td></td></tr<>	Along true sol on tild to sol sol at along true sol on tild to sol sol at along true sol at along true sol at along true sol along true sol at along true sol at along true sol at along true sol at along tru	Aling transmitten die terstigte die	
		X					
			X				
	Adap transition (3.2) to not junctify B			X	Adap from and to: 52 to not to 107 Main Main Main Main Main Main <		
Mary from out to 10° for null age 00. Mary from out to 10° for null age 00. Mary from out to 10° for null age 00. Mary from out to 10° for null age 00. Mary from out to 10° for null age 00. Mary from out to 10° for null age 00. Mary from out to 10° for null age 00. Mary from out to 10° for null age 00. Mary from out to 10° for null age 00. Mary from out to 10° for null age 00.	Aling from out into 120 for out (an int) Big from out into 120 for out (an int) Big from out (an int)				X	Aling transmitted total 107 to read glob 83.0 Image: State of the state glob 83.0 Image: State glob 83.0 Image: Sta	
					Adaptive rule digle (1) in rule (1) Image: State of the state o	X	
					Aling from out and the list to not to b 10 ² Big from out and the list to not to b 10 ² Big from out and the list to not to b 10 ² Big from out and the list to not to b 10 ² Big from out and the list to not to b 10 ² Big from out and the list to not to b 10 ² Big from out and the list to not to b 10 ² Big from out and to b 10 ²		X

Router-to-router performance: delay, jitter, frame loss <u>https://www.ams-ix.net/real-time-stats/</u>

3.3.3. MRTG



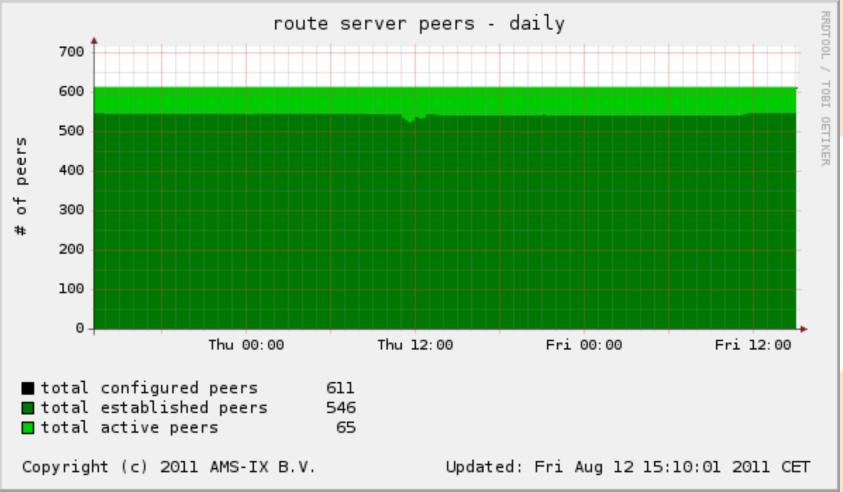


Daily graph: https://www.ams-ix.net/statistics/

38

3.3.3. MRTG



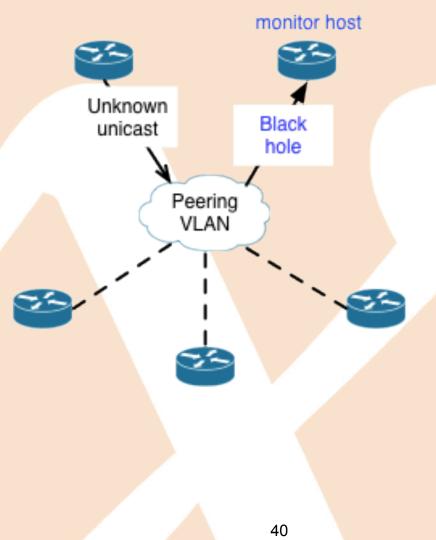


Route servers: https://www.ams-ix.net/rs-stats/

3.4. Traffic snooping



- Host connected to the production network with legal IX address stores all incoming traffic (broadcast, multicast, unknown unicast)
- 2. Already installed tools: tcpdump, tshark, etc
- 3. Black hole was made on it
- Extra: proactive traffic monitoring (STP, CDP, IGMP) through traffic analysis



3.5. ARP sponge



- 1. ARP sponge tool to reduce ARP traffic in case of "host is down/flapping"
- 2. If the sponge sees too many ARP requests it declares the host is down and starts replying with own address
- 3. ARP sponge monitors all ARP traffic, so it **knows** all mac addresses

More details about ARP sponge: <u>http://staff.science.uva.nl/~delaat/rp/2008-2009/p23/report.pdf</u> Source code: <u>http://www.ams-ix.net/downloads/arpsponge/</u>

3.6. ARP cache update



- 1. The router with proxy ARP was disconnected but other routers still have wrong ARP records
- 2. To update their ARP tables (proactive approach) ARP spoof request was sent
- 3. Another option: Gratuitous ARP or Unsolicited ARP Reply
- 4. Addresses were get from ARP sponge
- Spoofed ARP packets were injected by ARP sponge also

More details:

http://www.ams-ix.net/assets/Presentations/Euro-IX-19-ARP-Hijacking-Mitigation.pdf



3. Tools and procedures

Things helps to debug and solve the issue:

- 1. Team work
- 2. Central log server
- 3. Visualisation
- 4. Traffic snooping
- 5. ARP sponge
- 6. ARP cache update

Danger of Proxy ARP in IX environment



Questions?

Maksym Tulyuk maksym.tulyuk@ams-ix.net RIPE 63. Vienna 03 November 2011