Innovative Network Concepts
"Expert IP Network Consultants"

Layer-2 Multicast State Problems Caused by IPv6 Neighbor Discovery (ND)

Jeff Wheeler jsw@inconcepts.biz

The Problem

- MLD-snooping is like IGMP-snooping but for IPv6
  - It keeps unnecessary multicast traffic from being flooded to all ports on a LAN
- IPv6 Neighbor Discovery (ND) is like ARP
  - It is how hosts find each-other and their router
  - among other functions
- MLD-snooping and ND don't work well together
  - And they may break other multicast traffic
- If layer-2 multicast breaks, IPv6 breaks!

Why isn’t my network broken today?

Is it going to break in the future? If so, when? How will I know?

Why isn’t my network broken today?

- Do you have MLD-snooping enabled?
  - Probably not
- Do you have thousands of hosts with IPv6?
  - Maybe not today
- So it could break tomorrow, but “probably” or “maybe” won’t?
  - Pretty much. You need to test before something happens at 3:00 AM on Christmas.

Non-traditional multicast problem!

- This affects you even if you do not think you use multicast in any way
  - No streaming video, no financial data stream, no fancy replication, no in-house custom application
- Because IPv6 Neighbor Discovery depends on multicast, you are probably already using it today.
- Even if you have no public IPv6 addresses on any of your routers or hosts
- If not, you will be eventually
  - You'll upgrade your OS and suddenly it will configure an IPv6 link-local address *even with no public IPv6 address*!

Exposes “multicast newbies” to a whole new world of problems!

- Operators who have never used multicast, never cared about multicast, and never want to, are suddenly depending on it.
- They may not be equipped to troubleshoot problems.
- Vendor TACs are certainly not equipped.
- Industry is unprepared.

Background: IPv6 Neighbor Discovery

Comparison to the familiar IPv6 ARP

ND and ARP compared

- ARP – RFC826 November 1982
  - ARP Ethertype 0x0806
  - Requests are simply broadcasted
  - Replies are unicasted
- ND – RFC4861 September 2007 (originally RFC1970 August 1996)
  - 5 types of ICMPv6 packets (RS, RA, NS, NA, Redir)
  - Requests are sent to one of 2^24 IPv6 layer-3 multicast addresses ("Solicited-Node Addresses") computed based on the neighbor being solicited – looks like `FF02::1:FF`*`xx:xxxx`*
    - Which are then mapped to MAC destination `3333:FF`*`xx:xxxx`* for layer-2 multicasting
    - Which requires all IPv6 hosts to join multicast groups in order for Neighbor Discovery to work
  - Replies are unicasted
- ND is far more complex than ARP

Solicited-Nodes address – What is it?

- ND host solicitations go to `FF02::1:FF`*`xx:xxxx`*
  - Maps to MAC address `3333:FF`*`xx:xxxx`*
- RFC4291 §2.7.1 says, all hosts are "required to compute and join (on the appropriate interface) the associated Solicited-Node multicast addresses for all unicast and anycast addresses that have been configured for the node's interfaces"
- In a typical hosting / datacenter environment, that means every server or VM will join at least:
  - One L2 group for the link-local IPv6 address
  - One L2 group for each IPv6 address assigned
    - Unless those addresses have identical last-24-bits (unusual)

Why?

- IPv6 ND packets, unlike ARP, are not broadcast to every single host on the LAN
  - Actually they are, unless you have MLD-snooping on your LAN. MLD-snooping is like IGMP-snooping but for IPv6.
  - But your NIC card is capable of filtering out the packets the host is not interested in, assuming it has a large enough layer-2 multicast address table
- This is good because ARP who-has packets go to every host on the LAN; ND packets don't have to!
  - No longer do all your VMs have to process each packet!
  - This might be hundreds or even thousands of packets **per hour**, so it's a huge CPU savings, right? Well… no.

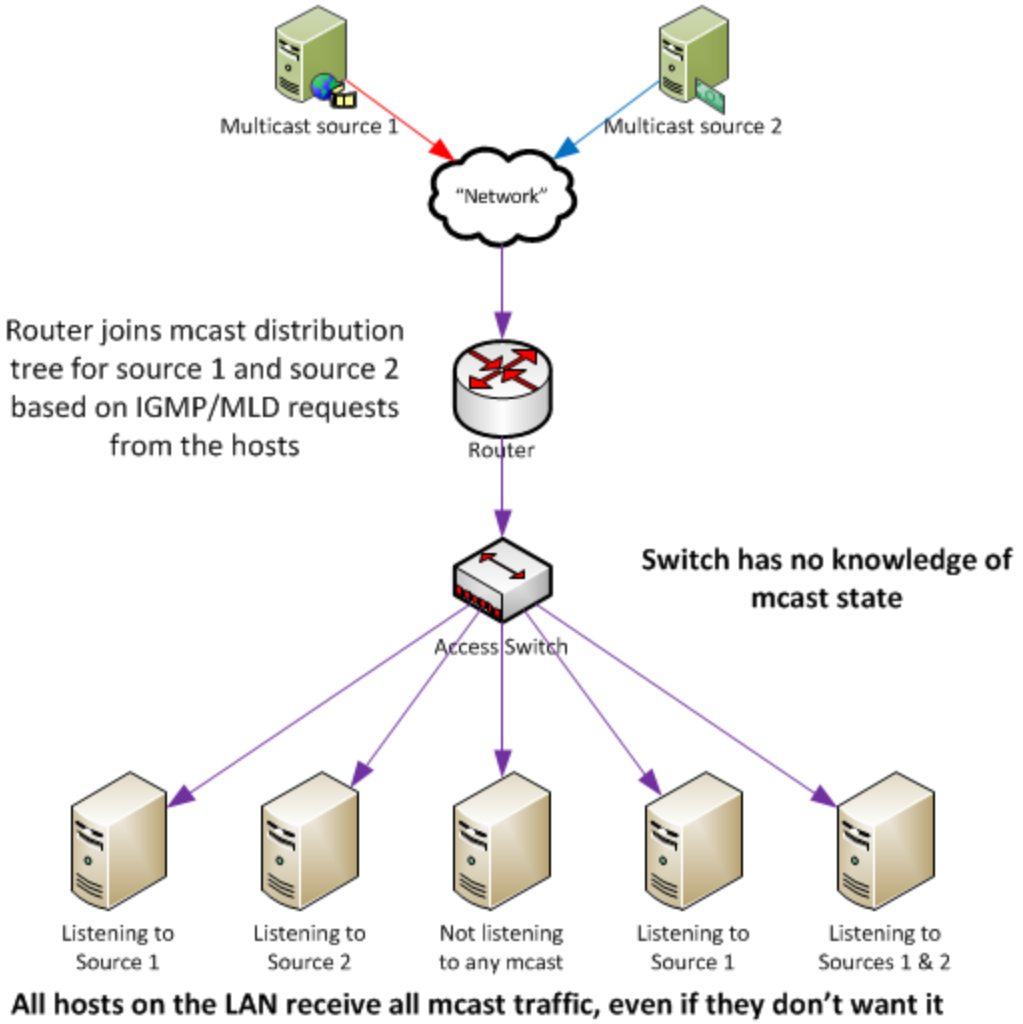
Background: MLD Snooping

Comparison to the familiar IGMP Snooping

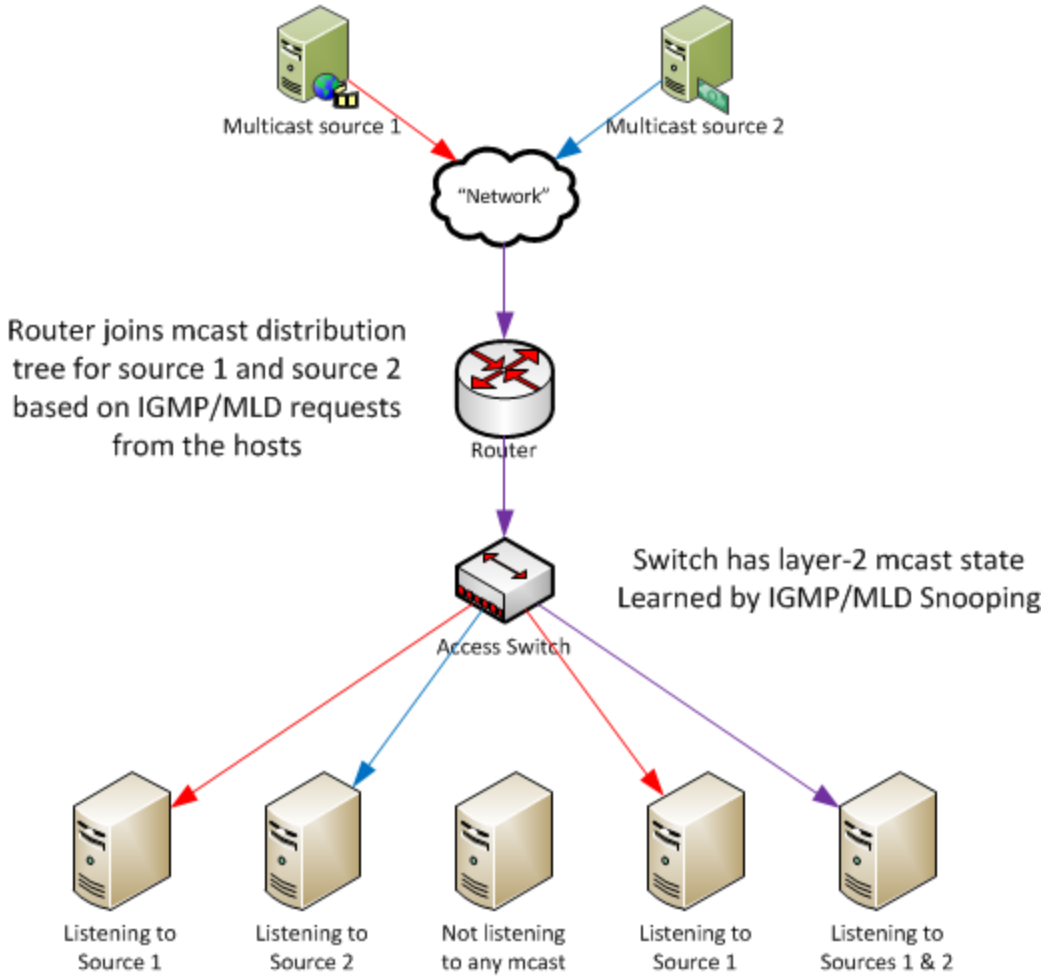
IGMP and MLD Snooping
What do they do?

- Both are mechanisms for layer-2 switches to learn about multicast group memberships
- Switches use this information to avoid flooding multicast traffic to all ports in a LAN
- Without IGMP or MLD snooping, all hosts on a LAN (or VLAN) will receive all multicast frames

LAN without IGMP/MLD Snooping



LAN with IGMP/MLD Snooping



IGMP and MLD Snooping Compared

- MLD (v1 RFC2710 Oct'99) (v2 RFC3810 June'04)
  - Is for IPv6
  - Encapsulated inside ICMPv6 packets
    - Types 130, 131, 132, 143
  - IPv6 Router Alert option
- IGMP (v1 RFC1112 Aug'89) (v2 RFC2236 Nov'97) (v3 RFC3376 Oct'02)
  - Is for IPv4
  - IP Protocol number 2 (TCP is 6, UDP is 17, etc.)
  - Router Alert option
- MLDv2 and IGMPv3 are quite similar
- MLDv1 is comparable to IGMPv2

Solicited-Nodes and MLD-snooping, how they can work together

- **They can't**, the whole idea is stupid
  - This was not predicted in the 90s when IPv6 was invented
- A modern datacenter TOR switch today (48x10GE 4x40GE street price $9k USD) scales to about 1000 - 3500 multicast groups
- Some switches support "fully-provisioned" multicast, but not all
  - Meaning every port could join every group (up to max groups) while others also have a listener limitation

How will I run out of 1000 groups?

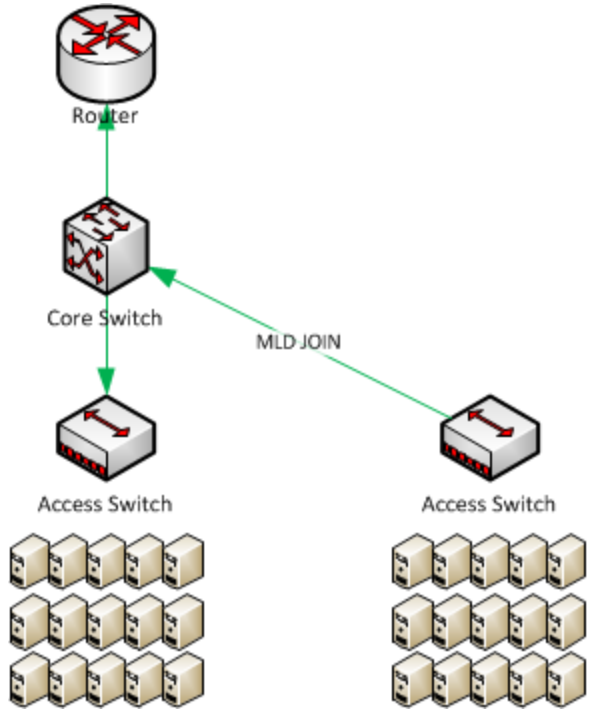
- If your layer-2 domain contains 1000 VMs (one rack of modern servers)
- Then just link-local IPv6 addresses will produce 1000 groups
  - Remember, solicited-nodes have a 24 bit collision space; so the chance of having two random link-local addresses (based on last 24 bits of MAC address) collide is about 1 / 2^24 or one chance in 16 million
  - VMs will have a link-local address even if no public IPv6 addresses are in use!
- Any VMs who actually use a public IPv6 address will consume yet more groups
  - If you have a lot of subnets and most customers use the first few addresses, like ::2, ::3, ::20, then you won't use up very many groups
  - But if you configure /64 subnets, and customers decide to pick different addresses, then each one could potentially create another group

How big are your layer-2 multicast state domains?

Is it one rack? My whole datacenter?
Something in-between?

No MROUTER, big domain!

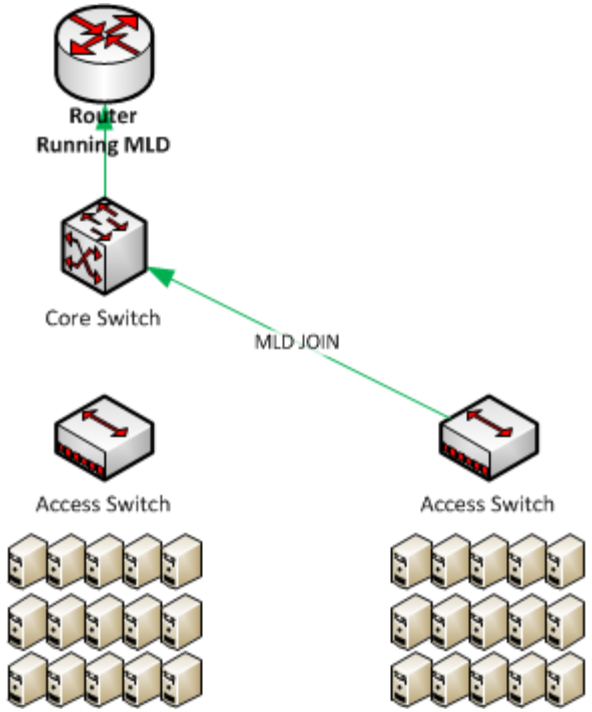
- With mld-snooping but no MLD router enabled on the LAN, there could be no promiscuous ports
- This means MLD join state will be replicated to all switches on the LAN



- Just 40 servers per TOR × 25 racks = 1000 groups
- Even with no public IPv6 addresses in use!
- Good thing you don’t need mld-snooping yet!

Configure MROUTER, smaller domain!

- Router is running MLD
- Switches should detect it is an MROUTER and find that they have a promiscuous (uplink) port
- MLD state won't be replicated to unrelated TOR switches



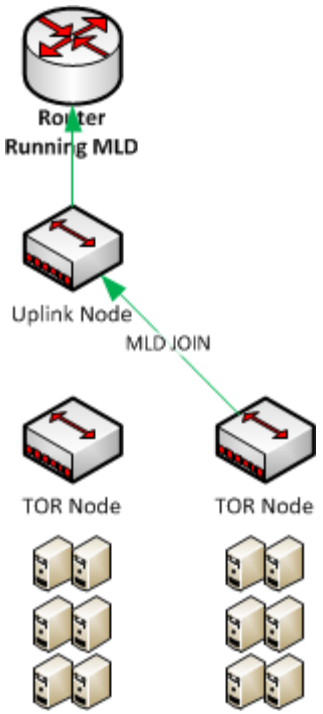
- But your core switch will still learn all MLD state
- Eventually it will run out of state table space
- What if you are using a "datacenter fabric?"
  - The fabric switches that connect to your routers probably have equally-small L2 mcast state tables. This means they will be exhausted rapidly and problems will start happening.

How big is your state-domain?

- Remember, it extends at least from your host (or VM) to that host's default router
- If your top-of-rack switch is not also the default router, then your VLANs might span some "core switch"
- If your VLANs are trunked to all rack switches, then your entire datacenter could be learning mcast groups for EVERY HOST THAT HAS IPv6
  - Including every VM!
- This will happen if your router is not configured as an IGMP/MLD-speaker. You probably won't configure that unless you are running PIM for layer-3 multicast!

What about “fabric switches?”

- Some datacenter “fabric switches” do not offer a powerful “core switch” to uplink to your layer-3 routers
- You simply use a 1U TOR-like switch to provide connectivity to “the Internet,” or the rest of your enterprise network



- Your uplink nodes will run out of L2 mcast state
- Again, good thing you don’t need mld-snooping yet

What happens when I run out of the 1000 groups my switch supports?

- Ask your vendor
- They might not know for sure, or the salesman might not tell the truth
- So you will need to test your switch anyway!
- Don’t believe them if they just say “it will just flood to all ports, no packets will be lost”
- Because what happens when a group has been created, but a second member wants to join it?
  - Will it evict that group from the state table so it can be flooded?
  - Or will it **silently drop packets**?

Review of MLD-Snooping and ND

They can interact badly.

I said Solicited-Nodes and MLD-snooping can't work together. Why?

- If you might run out of groups, then the best thing which your switch could possibly do is flood traffic to all ports
  - (worst thing is lose packets)
- This defeats the purpose of MLD-snooping
  - Which is to avoid sending multicast traffic to switch ports that are not listening to that group
- It also uses up groups which might be useful for other things
  - For example, multicast stream of financial data

Doesn’t this defeat the whole purpose of Solicited-Nodes?

- Almost. If your NIC card has a large enough multicast group filter, it can still filter out uninteresting ND packets (based on L2 group)
- So even though the NIC receives packets, it does not copy them to the host (or guest VM) and interrupt the CPU to process it

Will this break MLD-snooping or IGMP-snooping all together?

- Yes. If you use up all your groups by trying to filter out a few thousand packets per hour, then you won't have any left for useful things like that important financial data stream
- The combination of IPv6 Solicited-Nodes and MLD-snooping does more harm than good

Can this be fixed?

Worked around?

Can it be fixed? Worked around?

- Configure mld-snooping “flood nd” knob
  - Your switch vendor needs this feature!
- Disable MLD snooping all-together
- Eliminate Solicited-Nodes from IPv6 ND

FIX: MLD-snooping “flood ND” knob

- Your layer-2 switch MLD-snooping implementation could have a knob to
  - Ignore state for MAC address `3333:FFxx:xxxx`
  - Because that’s what Solicited-Nodes addresses, `FF02::1:FFxx:xxxx`, maps to
- This means that MLD-snooping will not work for multicast addresses `FFxx::FFxx:xxxx`
- It will still work normally for other addresses

FIX: MLD-snooping “flood ND” knob Technical Disadvantages

- Don’t expect to use any IPv6 multicast groups that are addressed `::FF`*`xx:xxxx`*
  - Except for ND, because of this work-around
  - Essentially 1/256th of the IPv6 multicast address space becomes special-purpose
  - Simply number IPv6 multicast groups `::00`*`xx:xxxx`* to `::FE`*`xx:xxxx`*
- But MLD work normally for other addresses

FIX: Disable MLD-snooping

- This means all IPv6 multicast traffic will be flooded to all ports on the LAN
  - Probably what you are doing today, but you need to check!
- You won't want to do this if you ever plan to have busy IPv6 multicast groups on your LANs
- Did you know that VXLAN, and other overlay technologies, will depend on multicast?
- It is in your interest to ask your vendor to provide a working mld-snooping implementation that is not subject to this problem!

FIX: Eliminate Solicited-Nodes

- Solicited-Nodes could just be eliminated
  - After all, if your LAN scales up to so many hosts that it is even useful to filter out the traffic, it scaled up to too many hosts for MLD to work
  - And in many environments, NIC cards might not have enough multicast filter entries
    - Common 10GE server adapter has just 128
- Just send ND Neighbor Solicitations to
  - IPv6 All-Nodes address `FF02::1`
  - Better yet, just broadcast them like ARP
  - Because the complexity of Solicited-Nodes Addresses has extremely limited value and high cost

FIX: Eliminate Solicited-Nodes How could we transition to this?

- Add a bit or Option to ND Router Advertisement so all hosts on the subnet will know about it
  - In about 10 years everyone will be updated… :-/
- This has been done before to correct deficiencies in ND, including deficiencies that the designers of ND thought were features (in the 1990s)
  - For example, they thought the lack of any "default router preference" mechanism was GOOD and even documented it in the ND RFC
  - Turns out they were wrong; now that feature has been added back to ND

FIX: Eliminate Solicited-Nodes Technical Disadvantages

- The multicasting scheme for ND has some particular advantages for wireless networks
  - If the access point takes advantage of it
- NIC card won't be able to filter unwanted packets
  - It might not be doing this today, anyway

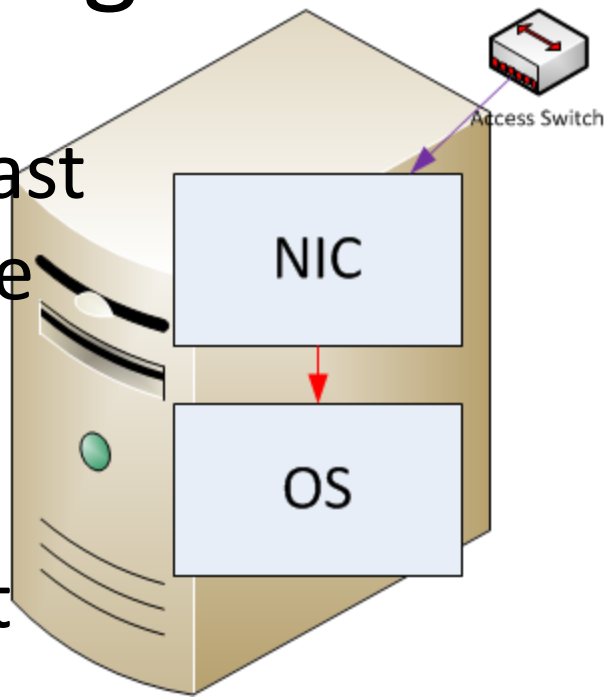
Conclusion

- Changing the way ND works is possible but would take a long time to be deployed
- Your vendor can give you an MLD-snooping knob by fiat, and this could be standardized
  - `(config)# mld-snooping nd-workaround`
  - Should be the default for all routers
    - But operators should be able to disable it if they want
    - `(config)# mld-snooping no-nd-workaround`

Concerns for server guys?

What about their NICs and hypervisors? Should they care about this problem?

How does the NIC filtering work?



- The host NIC has its own multicast state table, which is setup by the operating system.
- NIC filters out unwanted mcast traffic as long as the NIC's mcast state table has enough space.
- Unfortunately, state-of-the-art NICs only have ~128 filters. Once they are used up, the OS and NIC must compromise and allow unwanted traffic to reach the OS (kernel.)

NIC filtering state table wasted

- The more NIC filter table rows that are used up by filtering out a tiny amount of IPv6 ND traffic, the fewer rows are available for busy multicast groups.
- A smart OS could realize this and allow all of `3333:FF`*`xx:xxxx`* if the server has both a lot of IPv6 addresses, and needs some multicast groups for another reason
  - Financial data stream
  - Storage replication
  - Whatever

Issue for hypervisors?

- Hypervisors should take responsibility for IPv6 Neighbor Discovery listening and responses
  - This is a security feature
  - Which comes with added, beneficial side-effect!
- Now the guest VMs won't be awakened every time an IPv6 ND packet arrives that may or may not be interesting to the VM
  - Which might be expensive if you have to wake up 64 VMs at the same time
- Alternately, hypervisor could be aware of the layer-3 multicast group memberships for the IPv6 `FF02::1:FF`*`xx:xxxx`* address space

DRAFT

Comments / Questions?
Feel free to contact me

Jeff Wheeler jsw@inconcepts.biz