ISNIC rótarorsök

ISNIC incident post-mortem for virtual host outage on 2021-01-23

At 21:53 UTC, Saturday 23th, one of three physical servers in ISNIC’s production virtual cluster suddenly rebooted, bringing a large part of ISNIC services down with it.

ISNIC administrators were notified that the ISNIC website was responding with an error message at 10:00 UTC, Sunday January 24th. At 10:20 UTC service was mostly restored, including registry, EPP, whois and other `.IS` top level domain related functions.

The availability of the `.IS` zone was not affected, as we run geographically distributed authoritative nameservers for the zone. RIX was not affected.

The HFP (Hosting, Forwarding and Parking) service was in a read-only state (no changes could be made to hosted domains) from the start of the event until roughly 14:30 Sunday.

Some Svarbox customers were completely without service during this event. Other customers may have experienced degraded (slow or unreliable) service. Fortunately, most of Svarbox usage takes place during Icelandic business hours, so the real impact of the event was minimal. Service was fully restored by 13:59 UTC on Sunday.

The `teljari.is` service was not affected.

Timeline

Timestamp	Event
2021-01-23 21:59 UTC	A physical server in the production cluster reboots uncommanded
2021-01-24 10:00 UTC	ISNIC administrators are alerted to `isnic.is` website being down
2021-01-24 10:20 UTC	Services mostly restored, with the exception of Svarbox and HFP
2021-01-24 13:59 UTC	Svarbox service fully restored
2021-01-24 14:30 UTC	HFP service fully restored

Cause

We believe the reboot happened due to a hardware problem, though we have not yet been able to pinpoint a particular device fault.

In order to ensure the integrity of data, the virtual machines running in the cluster are not configured to start automatically in case of a physical server reboot; instead we were counting on our internal monitoring system to notify us of such a major event and could then do a controlled startup.

Two faults caused the monitoring system to fail to notify our on-call staff. The first was that both of our internal DNS resolvers were running on the failed server, so both were shutoff and none of our servers could resolve DNS hostnames. The second fault was that the monitoring system was also running on the failed server.

The extended HFP read-only state was caused by the fact that the hidden master DNS server VM was on the malfunctioning server, and when rebooted the `bind9` process failed to start, likely due to unsatisfied dependencies on external services. This was reported by the Zabbix monitors, but initially went undetected by staff due to information overload.

The extended Svarbox unavailability was caused by the fact that two Svarbox backend servers, and one Svarbox haproxy, were hosted on the malfunctioning VM host. This caused unavailability for any customers assigned to Svarbox clusters 2 or 3. When rebooted, these machines failed to start up properly; the `haproxy` and `svarboxd` services failed to start, likely due to unsatisfied dependencies on external services. This was also reported by the Zabbix monitors, but initially went undetected by staff due to information overload.

Action items/lessons learned

We have documented several necessary changes needed to minimize chances of an event like this causing such a long outage, some technical, others procedural.

All virtual machines have been moved off the affected server. We are in the process of ordering new equipment to replace the problematic server.

Our internal monitoring system has already been moved outside of the main cluster, to a dedicated server. One of our DNS resolvers will also be moved off of the main cluster.

We have deployed simple external monitoring which verifies that the internal monitoring server is active and able to send SMS alerts to the on-call personnel. This external monitor is run by a third party and does not rely on any ISNIC infrastructure, to ensure that internal outages cannot prevent on-call staff from being alerted in a timely fashion.

We are adding new tests to our monitoring system to ensure we get alerted if both parts of a redundant pair of VMs are running on the same virtual host, so we can respond proactively.

A checklist is under development to help on-call personnel focus their efforts during an outage. This will prevent mistakes and delays caused by information overload and unclear priorities. This outage was particularly problematic because our internal monitors could not be relied upon immediately after the event, so having a low-tech “offline” guide to manual troubleshooting would have made a big difference.