VMware Host 43 Restart

Problem Impact Analysis # Draft

Event Occurrence: 2018-04-23 Approx. 2200
Event Resolved: 2018-04-23 Approx. 2359

Background

The OIT Production Cluster consists of 16 physical servers hosting over 500 guest virtual servers and provides the vast majority of OIT's on-site hosting capacity. Mission-critical services such as Active Directory, Banner, Blackboard, Alertus, and more run on its hardware.

Break Down of the Problem

On Sunday, April 22, at 2:36:52 pm, VM43 stopped responding to management commands. The on-call infrastructure engineer examined the host with VMware support who concluded that the host would need to be rebooted as the hostd service was unresponsive and could not be restarted. An emergency reboot was scheduled for **ADD THIS**

Target State / Goal

VM hosts should never stop responding to management commands. On-premises infrastructure should be resilient enough that even if a physical host does fail the guest VMs are protected or resume operations quickly and automatically. Additionally, guest VMs should be distributed so that redundancy is preserved with members of the same cluster not running on the same host.

Root Cause Analysis

VMware support determined that bad firmware on the storage connections resulted in high latency, dropped packets, and bad paths which consumed host resources until the hostd service became unresponsive.

Develop Countermeasures

1. Update host firmware.
   This countermeasure was already identified following the 20180324 VMware Host Restarts event but was paused following the 20180406 VM Host 16 Failure when the engineer applying the upgrades rebooted the wrong host. It was decided to wait for the arrival of our new engineer, who has more VMware experience, to resume applying these upgrades.
2. Implement Distributed Resource Scheduler (DRS) anti-affinity rules.
   Anti-affinity rules tell DRS which servers are not allowed to run on the same host. vCenter automatically enforces the rules so that if a host fails only one of the servers in a cluster are impacted.

Implementation of Countermeasures

-

Follow Up / Review