THE SUPERCOMPUTER COMPANY

Overview of Node Health Checker

Jason Sollom
May 23rd, 2011

Introduction to Node Health Checker

- Node Health Checker is a System Management tool that can be configured to deal with "unhealthy" compute nodes. It can
  - **Reboot** unhealthy nodes, thereby erasing any problems
  - **Sequester** unhealthy nodes, preventing subsequent applications from running on and failing on them
  - "**Dump**" unhealthy nodes to an off-node location for later analysis and debugging

“Dumping” a node

- Dumping command: ldump
- Copies the node's memory to a file
- Can specify different amounts of memory to copy
  - Kernel pages
  - All used pages
  - All pages

Compute Node Availability

- Compute nodes are available if they are in the **UP** state.
- Compute nodes are **not** available if they are in any other state.
  - **DOWN**: failed to boot or experienced a hardware problem.
  - **SUSPECT**: Detected a problem; Monitoring to see if it recovers.
  - **ADMINDOWN**: Detected a problem. Not monitoring it.
  - **UNAVAIL**: Node will be rebooted.

Acronym

- NHC = Node Health Checker
- ALPS = Application Level Placement Scheduler
- SMW = System Management Workstation

Without NHC

- Unhealthy nodes can go undetected.
- An unhealthy node can be assigned to one application after another.
- An unhealthy node can cause an application to fail, wasting much computing time.
- System Administrators have to discover unhealthy nodes on their own.

With NHC

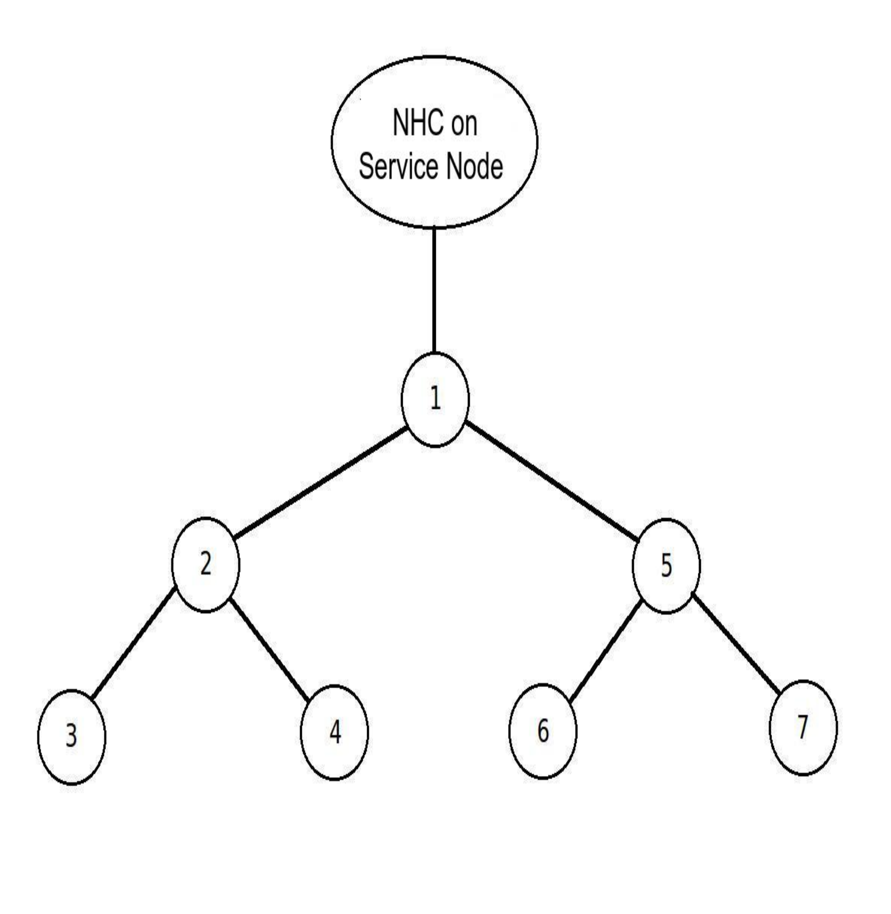
- Node health checking occurs automatically.
- When an “unhealthy” node is detected, it is removed from the pool of available compute nodes.
  - NHC detects common problems.

When and Where NHC is Launched

- Boot/Reboot
  - Launched from the compute node
  - Run-level script
- Application Termination (ALPS)
  - Launched from service node
- Manual Launch (System Administrator)
  - Launched from service node
  - This is rare.

NHC Fan-out Tree

- NHC uses a binary fan-out tree to contact nodes.
- NHC works around nodes it cannot contact.



Two Modes:

- Normal
- Suspect

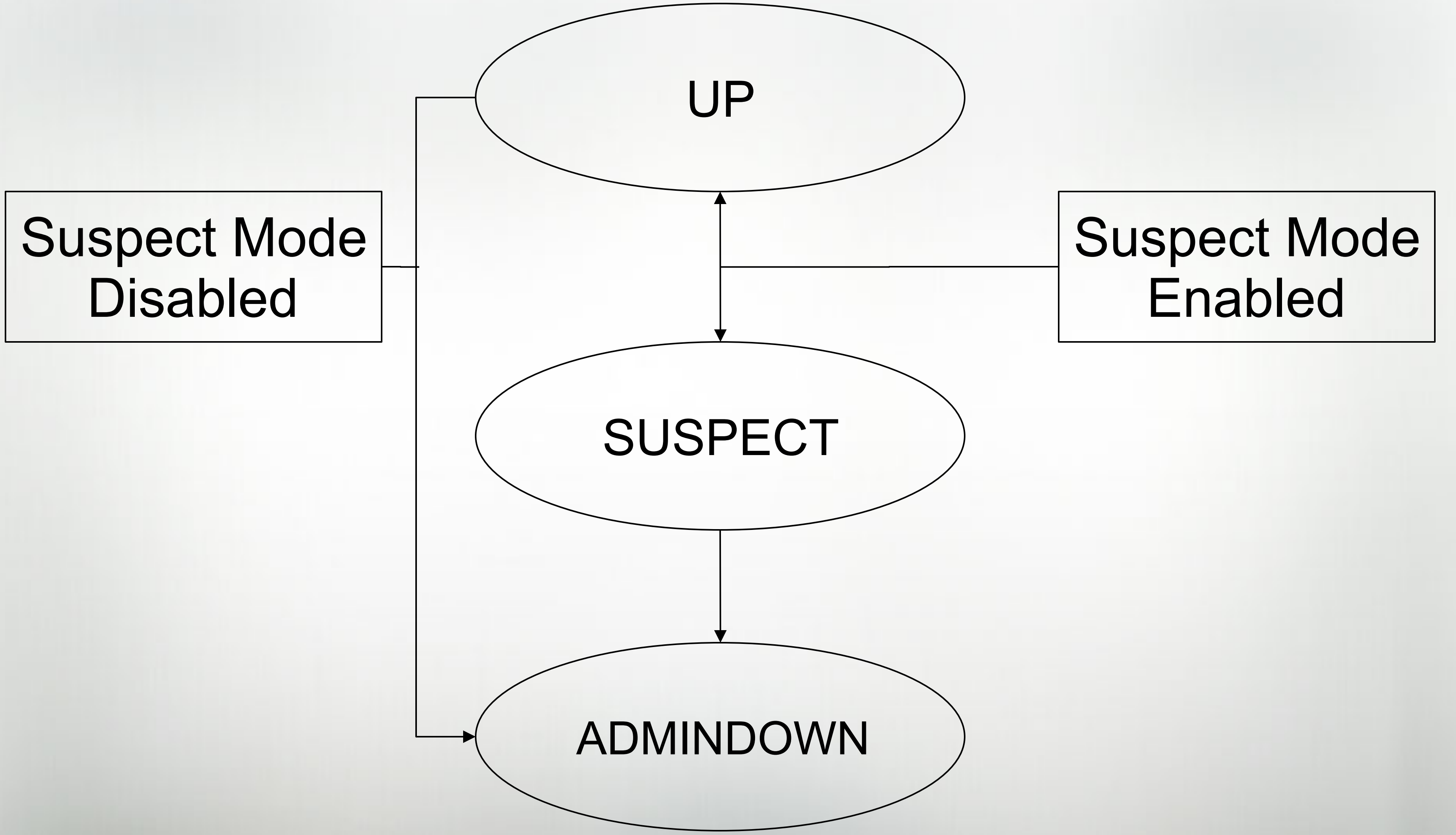
Normal Mode

- All NHC tests run once on each node.
- Failure indicates an unhealthy node.
- If Suspect Mode is disabled, unhealthy nodes are immediately dealt with.

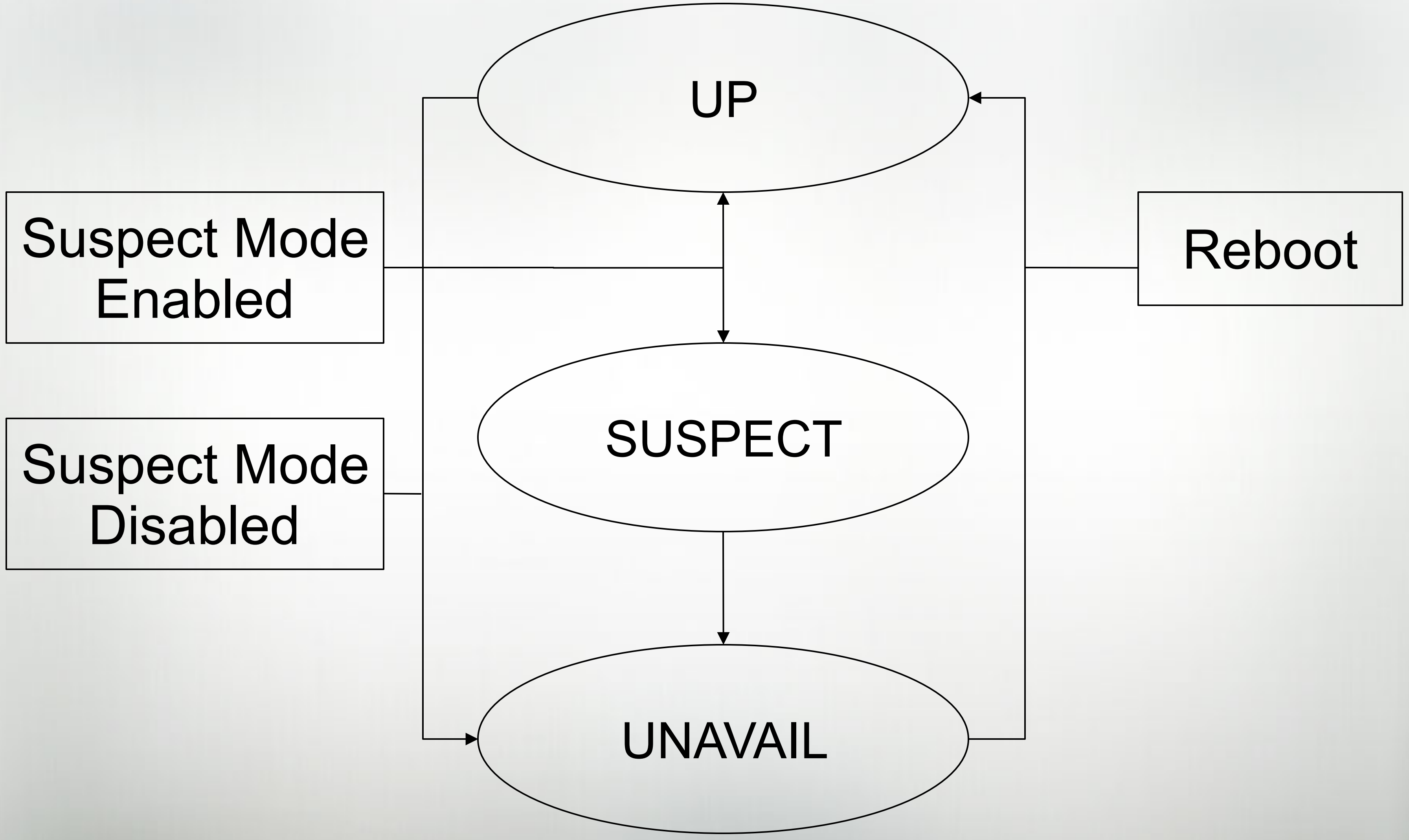
Suspect Mode

- NHC puts unhealthy nodes into a SUSPECT state and monitors them.
  - By default, the maximum length of Suspect Mode is 35 minutes.
- NHC re-runs failing tests until:
  - They pass.
  - Suspect Mode expires.

Node State Transitions



Node State Transitions



NHC Tests

- Specific Tests
- Test Actions
- Test Attributes: Timed Values

Six NHC Tests

- ALPS
- Application
- File System
- Memory
- GPU
- Plugin

ALPS Test

- Checks that the ALPS daemon is working.
- If not, ALPS could not launch applications onto the node.

Application Test

- ALPS gives NHC the Application ID (APID) for the application it should check on.
- NHC checks for processes running or hung on the node that are associated with the APID.
- If it finds any processes, the test fails.

File System

- Checks a mount point(s) on the node.
- File systems may be read-only or read-write.
- Can be configured to check
  - explicit mount points
  - all the mount points listed in /etc/fstab on the node.
- Mount points in /etc/fstab can be optionally excluded from checking.

Memory Test

- Checks the amount of non-free memory on the compute node.
- Specify a megabyte limit on non-free memory in the NHC configuration file. The test fails if this limit is exceeded.
- Only runs if the Application Test completes successfully.

GPU Test

- Runs a simple program to test the health of the GPU
- Can check the amount of non-free memory on GPU, similar to the Memory test for the processor.

Plugin Test

- Not a test, but a feature
- Runs any program accessible on the node.
  - Boot root
  - Mounted file system
- Exit code:
  - Zero: Success
  - Non-zero: Failure
- Allows NHC to be extensible, customizable.

Test Actions

- Each NHC test is assigned an action.
- Any action can be assigned to any test.
- Actions are only executed if a test fails.
- An error message is written out indicating the test failure.

Five Test Actions

- LOG
- ADMINDOWN
- DUMP
- REBOOT
- DUMP-REBOOT

Test Attribute:Timed Values

- WarnTime
- TestTime
- RestartTime

Dumping and Rebooting Nodes

- NHC sends dump and reboot requests to *dumpd* on the SMW.
- *Dumpd* is configurable:
  - Maximum amount of space allowed for dumps
  - Maximum number of dumps allowed
- NHC is configurable, too.
  - Maximum number of nodes that can be dumped per NHC invocation.

Error Reporting

- NHC reports errors to the console log.
- NHC Syntax:
- <node_health:VERSION> APID:123 (NHC_component) WARNING: ERROR MESSAGE
- Real Life Example:
- <node_health:4.0> APID:456 (Filesystem_Test) WARNING: This file was not listed in /proc/mounts: /lus/ nid00023

Service Node Crash Recovery

- If a service node should crash, once the service node is rebooted, NHC will automatically recover.
  - The in-progress NHC checks will be re-launched.
- If the service node is not rebooted, the documentation provides a manual way to recover NHC from a different service node.

NHC BoF Tomorrow

- Tuesday at 4:45
- Session 10B

Questions

CRAY

THE SUPERCOMPUTER COMPANY