Monitoring a HPC Cluster with Nagios

Piero Calucci

Scuola Internazionale Superiore di Studi Avanzati
Trieste

~~2009-04-01~~[1]
2009-04-03

[1] Try again. . . Fail better.

Outline

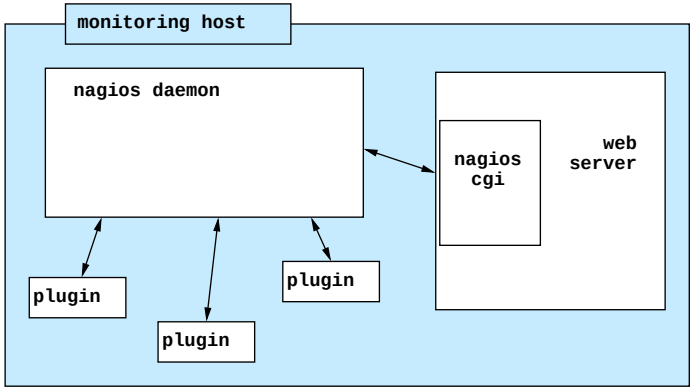
1. Nagios Concept
2. Nagios Web Interface
3. Nagios Installation for HPC Monitoring @SISSA

What is Nagios?

«NAGIOS® is a system and network monitoring application. It watches hosts and services that you specify, alerting you when things go bad and when they get better».

— *Nagios documentation*

Architecture



All the hard work is done by plugins, the nagios daemon «only» schedules them to be executed at the right time with the right parameters and collect results.

The cgi interface is entirely optional, but highly useful.

Nagios Operation

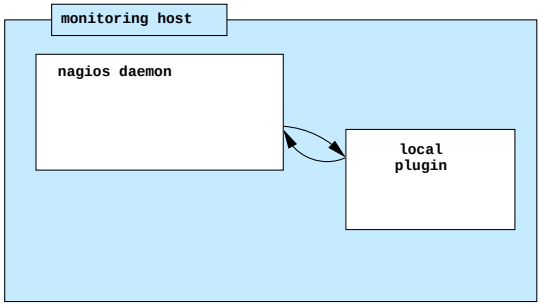
The nagios daemon

- schedules and executes active host and service checks
- accepts asyncronous passive checks
- sends out notifications on host or service state change
- executes event handlers on host or service state change
- writes and rotates log and state files

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Nagios Operation

Local Plugin Execution

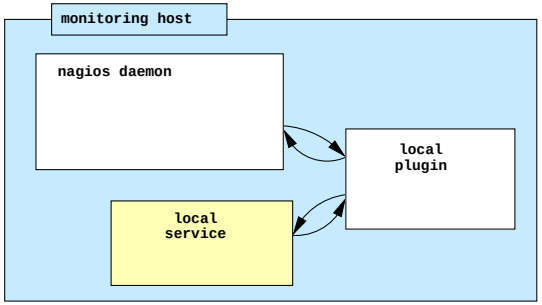


All active checks involve the local execution of some plugin.

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Nagios Operation

Local Plugin Execution: Local Service Check

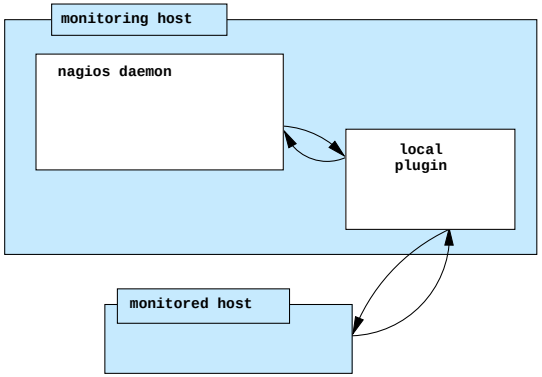


Locally executed plugins can just check for some local service. . .

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Nagios Operation

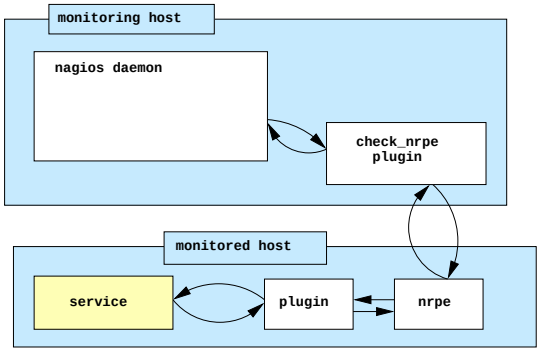
Local Plugin Execution: Remote Host Check



. . . or then can go to the network and check some remote host or service.

Nagios Operation

Nagios Remote Plugin Executor

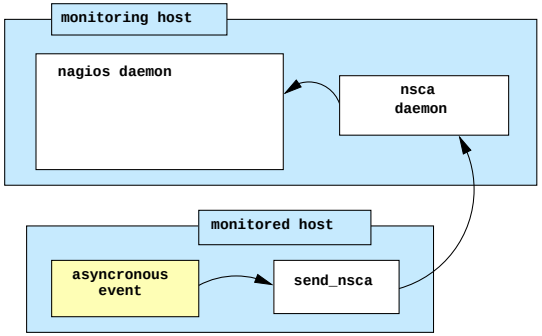


NRPE allows execution of plugins on remote hosts. Remote plugin results are reported to nagios by the locally executed `check_nrpe` plugin.

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Nagios Operation

Nagios Service Check Acceptor



The NSCA daemon relays to Nagios asyncronous notifications sent by `send_nsca` (this is how passive checks work).

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Host Checks

- a host is basically anything that can be given a name and an address
- hosts can be
  - UP
  - DOWN
  - UNREACHABLE (the host may well be up and running, but something in the network in between is broken)
- host checks are executed
  - at regular intervals
  - on-demand when a service on the host changes state
  - on-demand when required by reachability or dependency logic

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Service Checks

- a service is any monitored «thing» associated with a host
- service state can be
  - OK
  - WARNING
  - CRITICAL
  - UNKNOWN
- service checks are executed
  - at regular intervals
  - on-demand when required by dependency logic

State Types

Each host or service state can be of SOFT or HARD type.

SOFT type states are considered «uncertain» or «transitioning» and are checked with a different (usually higher) frequency until a specified maximum retry count is reached – they then become HARD states.

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Flap Detection

- When a host or service changes its state «too frequently» it is detected as being flapping
- Flapping hosts do not trigger notifications in order to avoid filling up mailboxes
- Flap detection threshold is configurable, and flap detection can be disabled entirely. However, the defaults are good enough

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Notifications

- notifications can be sent out whenever a HARD state transition occurs and when a host or service remains in a non-OK hard state for a specified time
- notification can enabled or disabled for each host or service
- notification times and contact groups can be set up so that only the right person is contacted, only when he is on duty

Monitoring a HPC Cluster with Nagios
Piero Calucci
Nagios Concept
Web Interface
HPC

Event Handlers

- event handlers are executed (if defined) when a host or service changes state and for each retry in SOFT states
- they are given all state information: state, type, retry count
- they can do basically anything, as long as they are given sufficient permissions, including
  - restarting a failed service or host
  - changing nagios configuration by writing to the command pipe (adaptive monitoring)

Tactical Overview

Hosts

4 Down	0 Unreachable	164 Up	0 Pending
1 Scheduled			
4 Acknowledged			

Services

37 Critical	3 Warning	0 Unknown	1392 Ok	91 Pending
2 Unhandled Problems	3 Unhandled Problems		64 Disabled	91 Disabled
35 on Problem Hosts				

Monitoring Features

Flap Detection	Notifications	Event Handlers	Active Checks	Passive Checks
Enabled	Enabled	Enabled	Enabled	Enabled
All Services Enabled	472 Services Disabled	All Services Enabled	155 Services Disabled	All Services Enabled
No Services Flapping	2 Hosts Disabled	All Hosts Enabled	All Hosts Enabled	All Hosts Enabled
All Hosts Enabled				
No Hosts Flapping				

Monitoring a HPC Cluster with Nagios

Piero Calucci

Nagios Concept

Web Interface

HPC

Host State

Host Information
Last Updated: Fri Mar 20 12:51:53 CET 2009
Updated every 90 seconds
Nagios® 3.0.6 - www.nagios.org
Logged in as *nagiosadmin*

View Status Detail For This Host
View Alert History For This Host
View Trends For This Host
View Alert Histogram For This Host
View Availability Report For This Host
View Notifications This Host

Host
c007
(c007)

Member of
c-nodes

10.2.10.7

Host State Information

Host Status:	UP (for 1d 0h 48m 9s)
Status Information:	PING OK - Packet loss = 0%, RTA = 0.21 ms
Performance Data:	rta=0.207000ms;3000.000000;5000.000000;0.000000 pl=0%;80;100;0
Current Attempt:	1/15 (HARD state)
Last Check Time:	03-20-2009 12:51:34
Check Type:	ACTIVE
Check Latency / Duration:	0.590 / 4.276 seconds
Next Scheduled Active Check:	03-20-2009 12:56:44
Last State Change:	03-19-2009 12:03:44
Last Notification:	N/A (notification 0)
Is This Host Flapping?	NO (0.00% state change)
In Scheduled Downtime?	NO
Last Update:	03-20-2009 12:51:44 (0d 0h 0m 9s ago)

Monitoring a HPC Cluster with Nagios

Piero Calucci

Nagios Concept

Web Interface

HPC

Service Status Detail

Current Network Status
Last Updated: Fri Mar 20 12:51:28 CET 2009
Updated every 90 seconds
Nagios® 3.0.6 - www.nagios.org
Logged in as *nagiosadmin*

View History For This Host
View Notifications For This Host
View Service Status Detail For All Hosts

Host Status Totals

Up	Down	Unreachable	Pending
1	0	0	0

All Problems	All Types
0	1

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
9	1	0	0	0

All Problems	All Types
1	10

Service Status Details For Host 'c007'

Host	Service	Status	Last Check	Duration	Attempt	Status Information
c007	EDAC memory errors	WARNING	03-20-2009 12:42:12	0d 14h 9m 16s	3/3	WARNING - several correctable memory errors
	NFS mounts from local fstab	OK	03-20-2009 12:44:38	1d 0h 46m 53s	1/3	NFS mounts OK
	NTP server	OK	03-20-2009 12:48:58	1d 0h 22m 31s	1/3	NTP OK: Offset 0.00139260292 1 secs
	PING	OK	03-20-2009 12:48:31	1d 0h 42m 57s	1/3	PING OK - Packet loss = 0%, RTA = 0.24 ms
	SSH	OK	03-20-2009 12:42:48	1d 0h 28m 41s	1/3	SSH OK - OpenSSH_4.3 (protocol 2.0)
	job events	OK	03-19-2009 11:15:51	5d 19h 16m 0s	1/3	job 22372 by smogunov/losatli
	load average	OK	03-20-2009 12:47:14	1d 0h 4m 14s	1/3	OK - load average: 0.00, 0.00, 0.00
	lustre client	OK	03-20-2009 12:47:08	1d 0h 4m 21s	1/3	lustre client OK
	pbs_mom	OK	03-20-2009 12:47:21	1d 0h 44m 7s	1/4	TCP OK - 0.000 second response time on port 15002
	reverse_ping_IB	OK	03-20-2009 12:46:02	1d 0h 45m 26s	1/3	PING OK - Packet loss = 0%, RTA = 0.39 ms

Monitoring a HPC Cluster with Nagios

Piero Calucci

Nagios Concept

Web Interface

HPC

Service Problems

Current Network Status
Last Updated: Fri Mar 20 12:50:50 CET 2009
Updated every 90 seconds
Nagios® 3.0.6 - www.nagios.org
Logged in as *nagiosadmin*

View History For all hosts
View Notifications For All Hosts
View Host Status Detail For All Hosts

Host Status Totals

Up	Down	Unreachable	Pending
164	4	0	0

All Problems	All Types
4	168

Service Status Totals

Ok	Warning	Unknown	Critical	Pending
1474	3	0	36	10

All Problems	All Types
39	1523

Display Filters:

Host Status Types:	Pending | Up
Host Properties:	Any
Service Status Types:	All Problems
Service Properties:	Not In Scheduled Downtime & Has Not Been Acknowledged & Active Checks Enabled

Service Status Details For All Hosts

Host	Service	Status	Last Check	Duration	Attempt	Status Information
a199	EDAC memory errors	CRITICAL	03-20-2009 12:42:09	10d 1h 28m 46s	3/3	CRITICAL - many correctable memory errors
c007	EDAC memory errors	WARNING	03-20-2009 12:42:12	0d 14h 8m 38s	3/3	WARNING - several correctable memory errors
m038	EDAC memory errors	WARNING	03-20-2009 12:44:44	10d 1h 26m 10s	3/3	WARNING - several correctable memory errors
m045	EDAC memory errors	WARNING	03-20-2009 12:43:54	10d 1h 27m 25s	3/3	WARNING - several correctable memory errors

What to Monitor

Hosts

- a masternode
- 160+ computing nodes
- several NFS servers
  - including a HA NFS cluster
- lustre servers

What to Monitor

Services

- «generic» services (SSH, NTP, . . .)
- HPC-specific services (maui, pbs_server, pbs_mom)
- computing node health (load average, hardware errors, . . .)

What to Do

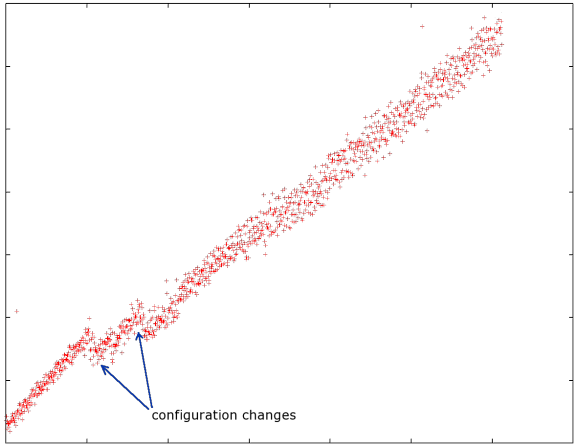
- send notifications when things go unrecoverably bad avoid sending out notifications every hour (Nagios default) – how often is too often?
- restart services and hosts when possible
  - how much does it take to declare a service crashed or a computing node dead?
  - do we trust Nagios to detect correctly?
  - do we accept the risk of rebooting a node that was just responding late?

Open Issues

- «undead» nodes
  - what are we going to do with nodes that reply to ping but nothing else?
    (they are UP from Nagios PoV)
  - how are we going to reliably detect the undead state?
- host reachability
  - cluster network topology is too simple to make Nagios reachability logic useful (no pingable gateways)
  - we have no way to check single switch port (do we?)
- multihomed nodes
  - Nagios has minimal support for multihomed hosts (or maybe I didn't understand it. . .)
  - we have no clear way to know all addresses associated with a node FQDN

Open Issues

High System CPU Time



After 3 weeks we are over 30% CPU usage on a 2x dual core opteron. . .

Open Issues

File Access Permissions Mess

On monitoring server:

- nagios needs to read its own configuration files and write log and status files
- web server needs to read some nagios config files and write to nagios command pipe
- nagios event handlers need to write to nagios command pipe
- nsca needs to read its own configuration and write to nagios command pipe

On monitored hosts:

- nrpe needs to read its own configuration and execute nagios plugins
- send_nsca needs to read its own configuration

Example

Check pbs_mom /1

```
define host {
        use             linux-server
        host_name       p001
        address         10.2.13.1
}
```

```
define hostgroup {
        hostgroup_name  p-nodes
        alias           planck nodes
        members         p001,p002,p003,. . .
}
```


Example

Check pbs_mom /2

```
define service {
        use                     generic-service
        hostgroup_name          m-nodes,c-nodes,p-nodes,. . .
        service_description     pbs_mom
        check_command           check_pbs_mom
        max_check_attempts      4
        notifications_enabled   1
        event_handler           restart_pbs_mom
        servicegroups           sg_node_batch
}
```

Example

Check pbs_mom /3

```
define command {
        command_name    check_pbs_mom
        command_line    $USER1$/check_tcp
                          -H $HOSTADDRESS$
                          -p 15002
}
```

```
define command {
        command_name    restart_pbs_mom
        command_line    /. . . /eventhandlers/restart_pbs_mom.sh
                          $HOSTNAME$ $SERVICESTATE$
                          $SERVICESTATETYPE$
                          $SERVICEATTEMPT$
}
```


Example

Check pbs_mom /4

`restart_pbs_mom.sh` is:

```
#!/bin/sh

SSH="ssh -i /var/lib/nagios/.ssh/id_dsa -t -t"
RESTART="sudo /sbin/service pbs_mom restart"

case $2 in
CRITICAL)
  case $3 in
  SOFT)
    case $4 in
    3)
      $SSH $1 $RESTART
    ;;
    esac
  ;;
  HARD)
    $SSH $1 $RESTART
  ;;
  esac
;;
esac
```


That's it

This slide intentionally left blank