



Installation, Integration and Maintenance of 1000s of nodes with three useful steps

Christopher Maestas

Jonathan Atencio

Nathan Baca

November 14, 2008

© 2004 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice



3 steps to a useful large system deployment



- Discovery through DHCP snooping
- Single System Image Management using oneSIS
- Load balanced HPC High Availability with HP Service Guard

Discovery through DHCP Snooping

Discovery through DHCP Snooping



- DHCP Snooping
- DHCP Option82
- Implications
- Use Cases
- Caveats
- HP Procurve Switch Configuration
- DHCP Server Configuration

DHCP Snooping

- Prevent rouge dhcp servers on network by only trusting
 - Switch-to-switch connections
 - DHCP Servers

DHCP Option 82

- Remote-id
 - Switch port
- Circuit-id
 - Switch mac address
 - Subnet vlan ip address
 - Management vlan ip address

Implications

- Client ip address is assigned based on switch and port
- No need to discover macaddresses
- No need to track macaddresses
- Integrate new and replacement nodes for efficiently

Use Cases

- PNNL Chinook
- SNL TACO

PNNL Chinook

- 2323 Compute Nodes
- 14 DHPC Servers
 - 1 per CU
- 240 Switches

SNL TACO

- 19 Spare nodes
- Rack the spare node and it is configured automatically

DHCP Snooping Caveats

- current implementation is switch mac address dependent
 - if a switch is replaced, to generate a new dhcp config for nodes connected to that switch
- trusted ports can not in turn be served a dhcp ip address based on dhcp snooping information

HP Procurve Configuration

- Enable dhcp-snooping

```
dhcp-snooping # enable dhcp snooping
dhcp-snooping vlan 1 # default vlan
dhcp-snooping trust 24 # trust dhcp server
dhcp-snooping trust Trk1-11 # trust uplinks
```

DHCP Server Configuration

- Class
- Pool
- Putting it all together

Class

- Match remote id and circuit id

```
class "cu1n1" {  
    match if(option agent.remote-id =  
        00:1b:3f:5c:06:c0) and  
        (binary-to-ascii (10, 8, ".",  
            option agent.circuit-id) = "0.1");  
}
```

Pool

- Allow members of a class

```
pool {  
    #compute  
    option host-name cu1n1;  
    range 172.30.1.1;  
    allow members of "cu1n1";  
}
```

Additional Information

- [Configuring Advanced Threat Protection](#)

Single System Image Management using oneSIS

Single system image management using oneSIS: What



- oneSIS – <http://www.onesis.org>
- enable diskless (stateless) technology
 - diskless used to fight against distribution initialization scripts
- image + configuration file
 - manage local disks as well

Single system image management using oneSIS: HowTo



- Image with sprinkles
 - XC, HP, EPEL, RPMFORGE, PNNL
- Image distribution to other nodes
- Configuration file inclusion for rhel4/5

Single system image management using oneSIS: Example Use Cases



- Configuration file overview
 - TYPES: director, associate director, managers, logins, computes, io-routers
 - DISKS
 - director cluster is the only diskfull install in the system.

Load balanced HPC High Availability with HP Service Guard

HP Serviceguard

Serviceguard

- What is it?
 - HP's high availability solution (XC)
 - Guardian of essential services for HPC & data centers
- Configuration
 - Redundancy, Redundancy, Redundancy
 - Node Failure
 - Load Balancing at scale
- Use Cases
 - 2-node Serviceguard cluster
 - Serviceguard cluster pool

What is Serviceguard?

High Availability Cluster

- A High Availability (HA) cluster is a managed collection of servers and storage devices. HA clusters provide high availability of service, applications, and data. And help achieve performance scalability.

Serviceguard

- Serviceguard for Linux (SGLX) is high availability clustering software designed by HP to protect mission-critical applications running on Linux from various software and hardware failures. (HPUX)

What is it really?

- Redundant configuration, nodes, network, storage, hardware.
- Cluster addressable by a network alias.
- Nodes communicate to each other with a regular heartbeat message.
- A node can go down but the network alias and any services can be passed on to another node in the cluster.
- Cluster handles node failure with reformation.

Configuration: Packages with services

- Packages contain all the processes, resources, and network aliases that provide a service.
- Resources can be:
 - Nodes
 - Volume groups and/or disk groups
 - File systems
- Each package has a configured *primary* node as well as *adaptive* nodes that it can use in case of failure.
- Packages are distributed across the cluster to maximize availability and performance.
- Serviceguard provides the mechanism to seamlessly transfer a package from one node to another.
- Node failure does not lead to service failure.

Use Cases: Chinook Configuration

2-node Serviceguard cluster

- System Monitoring & Logging
- Critical infrastructure
 - NFS exports
- 1 package (1 primary, 1 adoptive)

14-node Serviceguard cluster pool

- Diskless node image distribution
- Large Scale HPC environment
 - Large set of nodes split into smaller more manageable computational units (CUs)
 - CUs booting diskless image
- Balance image services across Serviceguard cluster pool
 - 12 packages (1 primary, 13 adoptive)
 - 2 idle adoptive nodes

Use Cases: Chinook Packages

Director 2-node cluster

- NFS package
- 800G ext3
- /shared

Node Pool CU Admin cluster

- NFS package
- read only, 1.5T xfs
- /cu1admin, /cu2admin, /cu3admin, /cu4admin,
/cu5admin, /cu6admin, /cu7admin, /cu8admin,
/cu9admin, /cu10admin, /cu11admin, /cu12admin

Chinook Serviceguard Services

Director cluster

- CU Admin Quorum Server
- OneSIS image for CU Admins
- Intel, Pathscale, XC, Totalview Licenses
- Logsurfer
- Syslog-ng
- Crond
- Speconfd
- Conserver
- Iptables
- Nagios
- Httpd

CU Admin cluster

- Syslog-ng CU Aggregation
- OneSIS image for CU Computes

Director Cluster Configuration

```
CLUSTER_NAME          cu0director

NODE_NAME             cu0director1
  NETWORK_INTERFACE   eth0
    HEARTBEAT_IP      X.X.X.X
  NETWORK_INTERFACE   eth1
    HEARTBEAT_IP      Y.Y.Y.Y
  NETWORK_INTERFACE   eth2
    STATIONARY_IP     Z.Z.Z.Z
  CLUSTER_LOCK_LUN    /dev/cciss/c0d0p1

NODE_NAME             cu0director2
  NETWORK_INTERFACE   eth0
    HEARTBEAT_IP      A.A.A.A
  NETWORK_INTERFACE   eth1
    HEARTBEAT_IP      B.B.B.B
  NETWORK_INTERFACE   eth2
    STATIONARY_IP     C.C.C.C
  CLUSTER_LOCK_LUN    /dev/cciss/c0d0p1

HEARTBEAT_INTERVAL    1000000
NODE_TIMEOUT           2000000
AUTO_START_TIMEOUT    600000000
NETWORK_POLLING_INTERVAL 2000000
MAX_CONFIGURED_PACKAGES 150
```

CU Admin Package Configuration

```
PACKAGE_NAME          culadmin
PACKAGE_TYPE          FAILOVER

NODE_NAME              culadmin1
NODE_NAME              cu0imnas1          # Standby Image Server
NODE_NAME              cu0imnas2          # Standby Image Server
NODE_NAME              cu12admin1
NODE_NAME              cu11admin1
NODE_NAME              cu10admin1
NODE_NAME              cu9admin1
NODE_NAME              cu8admin1
NODE_NAME              cu7admin1
NODE_NAME              cu6admin1
NODE_NAME              cu5admin1
NODE_NAME              cu4admin1
NODE_NAME              cu3admin1
NODE_NAME              cu2admin1

AUTO_RUN              YES
NODE_FAIL_FAST_ENABLED YES
RUN_SCRIPT            /.../culadmin.cntl
HALT_SCRIPT           /.../culadmin.cntl
SCRIPT_LOG_FILE       /.../culadmin.cntl.log
FAILOVER_POLICY       MIN_PACKAGE_NODE
FAILBACK_POLICY       MANUAL
PRIORITY              NO_PRIORITY
```

Documentation

- <http://docs.hp.com/en/B9903-90060/index.html>
- <http://www.docs.hp.com/en/ha.html>
 - High Availability NFS for Linux
 - Serviceguard for Linux Toolkits
- <http://h71028.www7.hp.com/enterprise/cache/4176-0-0-121.aspx>
 - Apache toolkit
 - MySQL toolkit
 - NFS toolkit
 - PostgreSQL toolkit
 - Oracle toolkit
 - Samab toolkit
 - Sendmail toolkit
 - Tomcat toolkit