Find the latest webinars at:

News, Views, Tips, Forums, Jobs, Resources, and More!

OpenVMS Cluster Load Balancing

Presented by

Paul Williams

www.parsec.com | 888-4-PARSEC

To Download this Presentation, please visit:
http://www.parsec.com/public/ClusterLoadBalancing.pdf

To E-mail Paul
williams@parsec.com

www.parsec.com | 888-4-PARSEC

Outline

- Load Balancing Mechanisms
- Batch and Print Queues
- TCP/IP
- DECnet
- Local Area Transport (LAT)
- Host Based Volume Shadowing
- MSCP Server
- Lock Manager
- Questions and Answers

Evaluating Load Balancing Mechanisms

What happens when?

- A new request is made
- A node fails
- Resources are exhausted on a node
- A node is returned to service

Load Balancing Goals

- Never direct a request to a non-functional node
- Direct requests to the node which can provide the best level of service
- Direct requests to other nodes prior to scheduled downtime
- Make failover and recovery transparent to user

Load Balancing Mechanisms

- Failover
  - All requests go to a single node while it is up
- Round Robin
  - Balanced based only on number of requests serviced
- Load Based
  - Balances requests based on ability of serving nodes to handle the work

OpenVMS Queue Manager

- Maintains all queues, forms and characteristics
- Manages all jobs in each queue
- Must run on one node in a VMScluster
- Default is any node
- Failover is automatic and transparent to users

```
$ start /queue /manager /on=(class2,class3,*)
$ show queue /manager /full
Master file:  STAFF_DISK:[COMMON]QMAN$MASTER.DAT;

Queue manager SYS$QUEUE_MANAGER, running, on CLASS2::
  /ON=(CLASS2,CLASS3,*)
  Database location:  STAFF_DISK:[COMMON]
```

Generic Batch Queues

- Directs batch jobs to available batch execution queues
- Balances load based on ratio of job count to job limit
- Restarting of batch jobs is not transparent and is controlled by the user

```
$ initialize/queue/batch class2_batch/start/job_limt=5/on=class2::
$ initialize/queue/batch class3_batch/start/job_limt=3/on=class3::
$ initialize/queue/batch -
_$ class_batch/generic=(class3_batch,class2_batch)
$ show queue/batch/full
Batch queue CLASS2_BATCH, idle, on CLASS2::
  /BASE_PRIORITY=4 /JOB_LIMIT=5 /OWNER=[SYSTEST,SYSTEM]
  /PROTECTION=(S:M,O:D,G:R,W:S)

Batch queue CLASS3_BATCH, idle, on CLASS3::
  /BASE_PRIORITY=4 /JOB_LIMIT=3 /OWNER=[SYSTEST,SYSTEM]
  /PROTECTION=(S:M,O:D,G:R,W:S)

Generic batch queue CLASS_BATCH, stopped
  /GENERIC=(CLASS3_BATCH,CLASS2_BATCH) /OWNER=[SYSTEST,SYSTEM]
  /PROTECTION=(S:M,O:D,G:R,W:S)
```

Autostart Print Queues

- Provides failover for printers
- If node providing printer service fails, queue automatically fails over to a different designated node
- Practical for LAT, telnet and other network based printers
- Restarting of job is automatic
- Job restarts at the beginning

```
$ initialize /queue /start laser /processor=tcpip$telnetsym -
_$ /autostart=(class2::"dds24:3101",class3::"dds24:3101")
$ show queue /full laser
Printer queue LASER, idle, on CLASS2::"dds24:3101", mounted form
  DEFAULT
   /AUTOSTART_ON=(CLASS2::"dds24:3101",CLASS3::"dds24:3101")
  /BASE_PRIORITY=4 /DEFAULT=(FEED,FORM=DEFAULT) Lowercase
   /OWNER=[SYSTEST,SYSTEM] /PROCESSOR=TCPIP$TELNETSYM
  /PROTECTION=(S:M,O:D,G:R,W:S)
```

TCP/IP Considerations

- Presentation based on TCP/IP Services for hp OpenVMS
  - Current version v5.4
- Other brands offer different features and implementations
  - Multinet
  - TCPware

TCP/IP Cluster Alias

- One node of the cluster answers to the address
- If that node fails, another node begins responding
- Users and applications must initiate a new request after failover
- Perform the following on each node and then restart TCP/IP

```
$ tcpip
TCPIP> show configuration interface

 Interface: LO0
   IP_Addr: 127.0.0.1         NETWRK: 255.0.0.0         BRDCST:

 Interface: WE0
   IP_Addr: 10.100.0.16       NETWRK: 255.255.255.0     BRDCST: 10.100.0.255

TCPIP> set configuration interface we0 /cluster=class -
_TCPIP> /c_broadcast_mask=10.100.0.255 /c_network=10.100.0.17
```

Balancing TCP/IP with DNS

- May be implemented in any DNS
- Round-robin scheduling used in BIND version 8
- Random cyclic scheduling used in BIND version 9
- Enter multiple A records for a single host name in the DNS

```
class           IN      A       10.100.0.47
                IN      A       10.100.0.30
                IN      A       10.100.0.72
```

Using Load Broker with DNS

- Metric server runs on each cluster member
- Load broker runs on one or more nodes
  - May run on multiple nodes of the cluster
  - Could be a single node not part of the cluster
- Load broker polls metric servers for current load rating
- Load broker sends updates to DNS with current address list
  - Least loaded systems are listed first
  - Unresponsive systems are removed from the list
- Requires DNS to allow dynamic update from load broker
- DNS uses normal round-robin scheduling on list of addresses
- TTL (Time To Live) set to limit length of time addresses are cached
  - Tradeoff between DNS load and speed of fail-over
  - Some clients ignore TTL or set a minimum time

Metric Server

- Calculates rating based on
  - Count of interactive users
  - Interactive user limit
  - Previous rating to smooth out peaks
  - Rating defined by system manager with logical name tcpip$metric_cpu_rating
  - Amount of free memory
  - System parameter FREEGOAL
- Enable metric server with tcpip$config.com
  - Server components menu

Configuring Load Broker

- Enable dynamic updates on master DNS from load broker
- Enable load broker with tcpip$config.com
  - Server components menu
- Configure load broker
  - Files in sys$sysdevice:[tcpip$ld_bkr]
  - Copy template configuration file
  - tcpip$lbroker_conf.template
  - tcpip$lbroker.conf
  - Restart after changing configuration

Sample Load Broker Configuration File

sys$sysdevice:[tcpip$ld_bkr]tcpip$lbroker.conf

```
cluster "class.parsec.com"
{
  dns-ttl 45;
  dns-refresh 30;
  masters {
    10.100.0.53;
  };
  polling-interval 9;
  max-members 3;
  members {
    10.100.0.100;
    10.100.0.53;
    10.100.0.54;
    10.100.0.80;
    10.100.0.129;
    10.100.0.130;
  };
failover 10.100.0.200;
};
```

PARSEC Group
Our Trainers Consult. Our Consultants Train.

Cisco LocalDirector

- Intelligent layer 2 (IP stack) bridge
- Load balances TCP/IP traffic across multiple servers
- Directed mode
  - Uses Network Address Translation (NAT)
- Dispatch mode
  - Multiple servers respond to the same address after they have been configured which is shown on next slide
  - LocalDirector replaces MAC address with that of the target server

Configuring Multiple Addresses on Interface

- @sys$manager:tcpip$define_commands
- ifconfig <interface> alias <address>/<mask>
- netstat –n "-I" <interface>

```
$ @sys$manager:tcpip$define_commands
$ netstat -n "-I" we0
Name  Mtu   Network     Address                Ipkts Ierrs    Opkts Oerrs  Coll
WE0   1500  <Link>      aa:0:4:0:12:4          80669     0    21890     0     0
WE0   1500  10.100.0/24 10.100.0.18             80669     0    21890     0     0
WE0   1500  10.100.0/24 10.100.0.17             80669     0    21890     0     0
$ ifconfig we0 alias 10.100.0.13/24
$ netstat -n "-I" we0
Name  Mtu   Network     Address                Ipkts Ierrs    Opkts Oerrs  Coll
WE0   1500  <Link>      aa:0:4:0:12:4          80799     0    21963     0     0
WE0   1500  10.100.0/24 10.100.0.18             80799     0    21963     0     0
WE0   1500  10.100.0/24 10.100.0.17             80799     0    21963     0     0
WE0   1500  10.100.0/24 10.100.0.13             80799     0    21963     0     0
```

DECnet Alias

- Distributes incoming DECnet connections among cluster nodes
  - DECnet IV – Round-robin distribution based on connection count compared with maximum links allowed
  - DECnet Plus – Round-robin distribution based on connection count compared with specified weight
- Not transparent
  - User or application must re-connect after node failure

DECnet Phase IV

- Requires OpenVMS routing node as a cluster member
- Limit of 64 nodes per cluster alias
- Range of maximum links is 1 – 200
- DECnet must be restarted after configuring
- The following must be performed on each node in the cluster

```
$ mcr ncp
NCP>define node 1.100 name class
NCP>define executor alias node class alias incoming enabled alias maximum links 50
NCP>list executor characteristics

Node Permanent Characteristics as of 19-AUG-2005 10:14:50

Executor node = 1.18 (CLASS2)

Management version       = V4.0.0
Type                     = nonrouting IV
Maximum address          = 1023
Alias incoming           = Enabled
Alias maximum links      = 50
Alias node               =  1.100 (CLASS)
```

DECnet Plus

- Must have a router or a cluster member to handle a cluster alias
- Limit of 144 nodes per cluster alias
- Enable with net$configure.com shown below
- net$alias_startup.ncl is created by net$configure.com and executed at DECnet startup

```
$ @sys$manager:net$configure advanced

        [9]     Configure Cluster Alias

* Which configuration option to perform?                    [1] : 9
* Do you want to ADD or DELETE an alias?                  [ADD] :
* Fullname of Cluster Alias:                                    : .test
* Cluster Alias Phase IV Address (aa.nnnn OR AA-00-04-00-xx-xx)  : 1.813
* Selection weight for this cluster node [0 for satellites]  : 80
```

Local Area Transport (LAT)

- Distributes incoming interactive user sessions among nodes advertising the service
- Distribution is based on load rating
  - Dynamic rating calculated based on load and availability
  - Static rating set by system manager which overrides the dynamic rating
- User must create a new session after node or network failure
- Bridged protocol, not routable

LAT Services

Create LAT service on each node of the VMScluster

- One service is created by LAT startup procedure, default is node name

```
$ @sys$startup:lat$startup class2
```

- Additional services may be created in LAT$SYSTARTUP or other startup procedures for the cluster alias

```
$ lcp :== $latcp
$ lcp create service class
```

- Static rating may be set on the service
  - Valid range is 0 (lowest) to 255 (highest)

```
$ lcp create service /static_rating=100
$ lcp set service /static_rating=100
```

- Dynamic rating may be affected by setting the cpu rating
  - Valid range is 1 (lowest) to 100 (highest)
  - 0 uses default rating based on interactive job limit

```
$ lcp set node /cpu_rating=cpu-power
```

LAT Dynamic Load Rating Calculation

LATACP calculates the load rating based on

- CPU utilization
- Number of interactive processes compared to interactive login limit
- Available free memory compared to FREEGOAL
- CPU rating or interactive job limit if CPU rating is at its default value of 0

Load rating is calculated once each second with changes smoothed over time

LAT Load Rating Continued

The dynamic rating is set to 0 if

- Interactive job limit is consumed
- All units for OPENVMS-ALPHA or OPENVMS-ALPHA-USER license have been consumed

The following files in SYS$EXAMPLES may be modified to create your own load rating calculations

- LAT$RATING_CALC.C
- LAT$RATING_BUILD.COM
- LAT$RATING_DPT.MAR

Viewing LAT Service Rating

```
$ mcr latcp show node

Node Name:   CLASS2                          LAT Protocol Version:      5.3
Node State:  On
Node Ident:   Welcome to OpenVMS (TM) Alpha Operating System, Version V7.3-2

Incoming Connections:  Enabled               Incoming Session Limit:   None
Outgoing Connections:  Disabled              Outgoing Session Limit:   None
Service Responder:     Disabled              Announcements:          Enabled

Circuit Timer (msec):        80              Keepalive Timer (sec):      20
Retransmit Limit (msg):       8              Node Limit (nodes):       None
Multicast Timer (sec):       60              CPU Rating:                  0
Maximum Unit Number:       9999              Extra Datalink Buffers:      9
Queue Limit:                 24              Forward Session Limit:      16

User Groups:     0
Service Groups:  0
Service Classes: 1

Service Name     Status      Rating  Identification
CLASS2           Available     79 D   Welcome to OpenVMS (TM) Alpha Operating System,
  Version V7.3-2
CLASS            Available     79 D   Welcome to OpenVMS (TM) Alpha Operating System,
  Version V7.3-2
```

PARSEC Group
Our Trainers Consult. Our Consultants Train.

MOP Boot Servers

- Boot servers service boot requests
  - From satellite nodes
  - From terminal servers
- Configuring multiple boot servers
  - Provides failover for boot requests
  - Increases load on each server
  - Fastest node completes boot request

Host Based Volume Shadowing

Load balancing shadow copies

- Limit shadow merge and full copy operations on a single node
  - This forces handling of these operations on other nodes
- System parameter SHADOW_MAX_COPY limits number of copy or merge threads on that node
  - Dynamic parameter
  - May set to 0 before mounting a disk to prevent the node from doing the shadow copy

MSCP Server

- Load balancing distributes MSCP connections from client nodes among MSCP server nodes
- Static
  - Optimal server selected when disk mounted
  - May reselect server during mount verification
- Dynamic
  - VAX v6.0 and later only
  - Not available on Alpha
  - Optimal path rechecked every 5 seconds based on capacity

MSCP_LOAD System Parameter

- Controls MSCP server load balancing
  - 0 – no MSCP serving
  - 1 – MSCP serving with default capacity for system
  - >1 – Use specified value as estimated server capacity
    - Set to 4 times the estimated number of I/Os per second the system can handle
- MSCP server calculates available capacity
  - Server capacity less recent I/Os per second handled
  - Clients select server with highest available capacity
- System manager should bias load toward systems with fastest cluster interconnect to hosts with direct access to the disk

Lock Manager

- Used to synchronize access to many resources
  - Used heavily by RMS and file system
- The master of a resource is the first node to request a lock on the resource
  - Subsequent lock requests made to the lock master
  - The lock master keeps track of all locks for all nodes for that resource
  - Non-master nodes keep track of only their locks for that resource
- The director of resource points to the master
  - Initial lock requests from each node hashes the resource name and uses it as a directory lookup to determine which node masters the lock
  - Subsequent lock requests from that node for that resource communicate directly with the master

LOCKDIRWT System Parameter

LOCKDIRWT system parameter affects distribution of directory lookups and lock mastering

- Controls lock director lookup
  - 0 – no directory lookup on this node
  - Same on all nodes
    - Each node shares equally in lock directory functions
  - Different values on each node
    - Nodes with larger values handle proportionally larger amounts of directory lookup
- Special case – if all nodes have 0, then treat as though all nodes have 1

Lock Remastering

Lock remastering can only occur if there is interest in a resource by more than one node

Reasons for remastering

- Sole interest
- Higher LOCKDIRWT
- Higher activity of nodes with equal LOCKDIRWT

PE1 System Parameter

PE1 can control resource migration

- The following values may be set
  - 0 – Resource migration is enabled without restrictions
  - > 0 – Resources with up to the number of locks specified can migrate
  - < 0 – Resource migration is disabled

Monitor DLOCK

OpenVMS Monitor Utility
DISTRIBUTED LOCK MANAGEMENT STATISTICS
on node CLASS2
20-JAN-2003 14:21:28.01

		CUR	AVE	MIN	MAX
New ENQ Rate	(Local)	72.00	71.00	70.33	72.00
	(Incoming)	2.66	1.22	0.33	2.66
	(Outgoing)	286.00	282.88	279.66	286.00
Converted ENQ Rate	(Local)	0.33	1.77	0.33	2.66
	(Incoming)	0.00	0.33	0.00	0.66
	(Outgoing)	643.33	636.11	628.00	643.33
DEQ Rate	(Local)	72.33	71.00	70.00	72.33
	(Incoming)	0.00	0.33	0.00	1.00
	(Outgoing)	286.33	283.66	281.00	286.33
Blocking AST Rate	(Local)	0.00	0.00	0.00	0.00
	(Incoming)	0.00	0.00	0.00	0.00
	(Outgoing)	0.00	0.00	0.00	0.00
Dir Functn Rate	(Incoming)	21.33	148.77	21.33	240.33
	(Outgoing)	144.00	141.77	140.00	144.00
Deadlock Message Rate		0.00	0.00	0.00	0.00

Monitor RLOCK

OpenVMS Monitor Utility
DYNAMIC LOCK REMASTERING STATISTICS
on node CLASS2
17-JAN-2003 15:49:47

	CUR	AVE	MIN	MAX
Lock Tree Outbound Rate	0.66	0.00	0.00	0.66
(Higher Activity)	0.66	0.00	0.00	0.66
(Higher LCKDIRWT)	0.00	0.00	0.00	0.00
(Sole Interest)	0.00	0.00	0.00	0.00
Remaster Msg Send Rate	5.33	0.03	0.00	5.33
Lock Tree Inbound Rate	0.00	0.00	0.00	0.00
Remaster Msg Receive Rate	0.00	0.00	0.00	0.66

Find the latest webinars at:

News, Views, Tips, Forums, Jobs, Resources, and More!

Question & Answer

Presented by

Paul Williams

www.parsec.com | 888-4-PARSEC | williams@parsec.com

To Download this Presentation, please visit:
http://www.parsec.com/public/ClusterLoadBalancing.ppt

To E-mail Paul
williams@parsec.com

www.parsec.com | 888-4-PARSEC