Track 2: Operations: Data Center Architectures and Technologies

Disaster Recovery

SANOG 2006 Tutorials: 1st August 2006

Zeeshan Naseh
Day Agenda

• **Part 1 - Data Center Designs and Services** (Zeeshan Naseh)
  - Data Center Architectures and Technologies Overview
  - Content Switching & Application Optimization

• **Part 2 - L2 Switching Protocols** (Bilal Khawaja)

• **Part 3 - Fiber Channel and Storage Area Networks** (Asim Khan)

• **Part 4 - Data Center Disaster Recovery** (Zeeshan Naseh)
Data Center Disaster Recovery Agenda

• Distributed Data Centers
• Topologies
• DNS-BASED: Technology and Products
• ROUTE HEALTH INJECTION: Technology and Products
• HTTP REDIRECTION: Technology and Products
• Overcoming the Inherent Limitations
• Real World Deployments
DISTRIBUTED DATA-CENTERS
Why Distributed Data Centers?

- Avoid single point of failure
- High availability of applications and data for customers, partners and employees
- Scalability
- Load distribution: better use of global resources
- Disaster recovery
- ISP redundancy
- Better response: proximity to clients
- Optimal content routing

BUSINESS CONTINUITY And More …
Business Resilience

Ability of a business to adapt, change and continue when confronted with various outside impacts

Business Continuance

Ensuring business can recover and continue after failure or disaster: recovery of data and resumption of service

Disaster Recovery

Mitigating the impact of a disaster
Disaster Recovery

• Mechanisms used to react to a local failure by redirecting all requests to an alternate location

• Relies on data-replication

• Relies on applications being able to receive connections at any time in any location

• Typically refers to a topology with a “warm standby” data-center that only receives client requests when the primary fails
Disaster Recovery

SLB → DR Site

Content Switch

Primary Data Center

DR Data Center
Geographic Server Load Balancing

- Techniques used to distribute client traffic to servers across remote locations
- Very often deployed in conjunction with local load balancing (content switching)
- Business Continuance is the key driver
- Often associated to DNS-based deployments
- DNS is not the only solution (and has specific limitations!)
- Can rely on dedicated products or leverage content switches functions
Geographic Server Load Balancing

SLB

GSLB

Content Switch

Data Center 1

Data Center 2
TOPOLOGIES
Distributed Data Center Topology
Distributed Data Center Technologies

FCIP Link (FC frames tunneled over IP network)

IP Network

IP Storage Services Module for FCIP (GigE) in MDS9000 switch
Distributed Data Center Technologies

“Content Routing” for site selection (DNS, L3, RHI, HTTP redirection)
Distributed Data Center Technologies

“Content Switching” for load balancing, health monitoring, load/health reporting

Production Data Center

Back-up Data Center

IP Network

FCIP Link

IP Storage Services Module for FCIP (GigE) in MDS9000 switch
“Storage” & “Optical” for back-end data-base sync / async replication
Disaster Recovery Data Center Configurations

Warm Standby

Hot Standby
Disaster Recovery
Warm Standby

Site selection intelligence

Clients

Primary Data Center (Active)

Secondary Data Center (Standby)

Application A
Application B
While end users are serviced by Standby data center, begin logistics of recovering primary data center.
Warm Standby
Data Center Redundancy

• Advantages
  Simple design, typical Phase I deployment
  Easy to build and maintain
  Simple configuration

• Disadvantages
  Under utilization of resources
  Delay in failover with manual switchover
  No load sharing
Disaster Recovery
Hot Standby

Site selection intelligence

Primary Data Center (Active)

Secondary Data Center (Active)

Application A
Application B
Application C

Clients
Disaster Recovery
Hot Standby

Site selection intelligence

While end users are serviced by the Standby data center, begin logistics of recovering the primary data center.

Primary Data Center (Failed / Inactive)

Secondary Data Center (Active)

Application A
Application B
Hot Standby
Data Center Redundancy

- **Advantages**
  - Good use of resources due to load sharing
  - Ease of management

- **Disadvantages**
  - Complex, typical Phase II deployment
  - Data mirroring in both directions
  - Managing two active data centers
Site Selection Mechanisms

- Site selection mechanisms depend on the technology or mix of technologies adopted for request routing:
  1. DNS-based
  2. Route Health Injection and L3 routing
  3. HTTP redirect

- Health of servers and applications need to be taken into account

- Optionally, also other metrics (like load and distance) can be measured and utilized for a better selection
DNS-BASED: TECHNOLOGY AND PRODUCTS
DNS—Domain Name System

• Applications, like browsers, connect to servers using server names

• The operating system DNS resolver contacts the configured DNS server to get the IP address

• Applications use the address provided by the resolver

• When multiple addresses are provided, applications can behave differently: use first IP, use random IP, use first IP and move to the next one if unsuccessful
**DNS—Query**

**User Datagram Protocol, Src Port: 1302 (1302), Dst Port: domain (53)**

**Domain Name System (query)**
- Transaction ID: 0x002a
- Flags: 0x0100 (Standard query)
  - 0... ..... ..... = Response: Message is a query
  - .000 0... ..... = Opcode: Standard query (0)
  - ..... .0. ..... = Truncated: Message is not truncated
  - ..... ....1 ..... = Recursion desired: Do query recursively
  - ...... ....0 .... = Non-authenticated data is unacceptable

Questions: 1
Answer RRs: 0
Authority RRs: 0
Additional RRs: 0

**Queries**

- `www.cisco.com: type A, class inet`
  - Name: `www.cisco.com`
  - Type: Host address
  - Class: inet
DNS—Query Response

User Datagram Protocol, Src Port: domain (53), Dst Port: 1302 (1302)
Domain Name System (response)
Transaction ID: 0x002a
Flags: 0x8580 (Standard query response, No error)
Questions: 1
Answer RRs: 1
Authority RRs: 2
Additional RRs: 2
Queries <--snipped-->
Answers
www.cisco.com: type A, class inet, addr 198.133.219.25
  Name: www.cisco.com
  Type: Host address
  Class: inet
  Time to live: 1 day
  Data length: 4
  Addr: 198.133.219.25
Authoritative nameservers <--snipped-->
Additional records <--snipped-->
Iterative Requests

If the name server does not have a cached response, it provides a pointer to another name server.
Recursive Requests

If the name server assumes the full responsibility for providing a full answer
DNS-Based Site Selection

- The client DNS resolver (implemented as part of the client OS) typically sends a recursive query.
- The client D-proxy typically performs iterative queries, then returns the final result to the client.
- The device which acts as “site selector” is the authoritative DNS server for the domain hosted in multiple locations.
- The “site selector” sends keepalives to servers or content switches in the local and remote locations.
- The client connects to the selected location.
- All the devices involved might cache the information.
DNS-Based Site Selection

Client

http://www.cisco.com/

Data Center 1

Data Center 2

DNS Proxy

Root DNS for /

Root DNS for .com

Authoritative DNS
cisco.com

Authoritative DNS
www.cisco.com

Keepalives

© 2006 Cisco Systems, Inc. All rights reserved.
DNS-Based Site Selection

Root DNS for /

DNS Proxy

1. Client
http://www.cisco.com/

2. TCP:80

3. Authoritative DNS
cisco.com

4. Root DNS for .com

5. Authoritative DNS
www.cisco.com

6. Data Center 1

7. Data Center 2

8. Keepalives

9. Keepalives

10. Data Center 1
Data Center 2
GSLB Methods

1. Ordered List
   - Prefers first entry in the list.
   - Uses next VIPs when all previous VIPs are overloaded or down
   - Used for active-standby scenarios

2. Static Proximity Based on Client’s DNS Address
   - Maps IP address of client’s DNS proxy to available VIPs

3. Round Robin
   - Cycles through available VIPs in order

4. Weighted Round Robin
   - Weighting causes repeat hits (up to 10) to a VIP
GSLB Methods

5. Least Loaded
   Least connections on CSM and least loaded on CSS
   Load communicated through Content and Application Peering Protocol (CAPP) User Datagram Protocol (UDP)

6. Source Address and Domain hash
   IP address of client’s DNS proxy and domain used
   Always sticks same client to same VIP

7. DNS Race
   Initiates race of A-record responses to client
   Finds closest SLB to client’s D-proxy

8. Drop
   Silently discards request
Advantages of the DNS Approach

• Protocol independent: works with any application
• Easy to implement, with minimal configuration changes in the DNS authoritative server
• Can take load or data center size into account
• Can make the decision based on source IP (D-proxy)
Limitations of the DNS-Based Approach

- Visibility limited to the D-proxy (not the client)
- DNS caching in the D-proxy
- DNS caching in the client application (browsers defaulting to 15 or 30 minutes timeouts)
- D-proxy ignoring TTL
- Order of multiple A-record answers can be altered by D-proxies
ROUTE HEALTH INJECTION: TECHNOLOGY AND PRODUCTS
Route Health Injection

The Main Idea

- Rely on L3 protocols for request routing
- Advertise the same VIP address from two or more different data centers
- For Disaster Recovery advertise the preferred data center’s VIP with better metrics
- The upstream routers select the best route
- The content switches at each location provide server and application health monitoring
- In case of virtual server failure at the primary site, the route is withdrawn and network converges
Route Health Injection

Client A

Router 13

Router 11

Client B

Router 10

Location A
Backup Location for VIP x.y.w.z

Location B
Preferred Location for VIP x.y.w.z

Very High Cost

Low Cost

Preferred Location for VIP x.y.w.z

Backup Location for VIP x.y.w.z
Route Health Injection

Client A \(\rightarrow\) Router 13 \(\rightarrow\) Router 12 \(\rightarrow\) Location A

Backup Location for VIP x.y.w.z

Location A

Very High Cost

Router 12 \(\rightarrow\) Router 11 \(\rightarrow\) Client B

Client B

Preferred Location for VIP x.y.w.z

Location B

Low Cost

Router 10
Products for Route Health Injection
CSM + MSFC

• The Content Switching Module (CSM) can be configured to “inject” a 32-bit host route as a static route in the MSFC routing table.

• The CSM injects or remove the route based on the health of the back-end servers (checked with L3-7 probes or inband health monitoring).

• Out of band CSM – MSFC communication.
module ContentSwitchingModule 5
  variable ADVERTISE_RHI_FREQ 3

vlan 3 client
  ip address 10.3.3.20 255.255.255.0
  alias 10.3.3.21 255.255.255.0

vserver RHIVIP
  virtual 100.100.100.100 tcp www
  vlan 3
  serverfarm FARM1
  advertise active
  persistent rebalance
  inservice

Router#show ip route
Codes: C - connected, S - static, I - IGRP, R - RIP, M - mobile, B - BGP

  100.0.0.0/32 is subnetted, 1 subnets
S   100.100.100.100 [1/0] via 10.3.3.21, Vlan3

  10.0.0.0/24 is subnetted, 1 subnets
C   10.3.3.0 is directly connected, Vlan3
S*  0.0.0.0/0 [1/0] via 10.20.196.193
RHI on CSS using OSPF

Topology

- OSPF can be used on the Content Services Switch to achieve Route Health Injection
- CSS advertises host routes to the neighbor for the Active Virtual IP addresses
- In Box-to-Box Redundancy, the active CSS forms adjacency and advertises operational VIPs
- In VIP/Interface Redundancy both the CSSs forms adjacency but only active advertises
- In VIP/Interface Redundancy advertise always when redundant-vip not configured
RHI on CSS using OSPF
Box to Box Redundancy

ip redundancy
app
app session 10.33.133.2

ospf router-id 10.115.1.12
ospf as-boundary
ospf area 0.0.0.2
ospf enable
ospf advertise 10.115.1.101 255.255.255.255 metric 200
ospf advertise 10.115.1.102 255.255.255.255 metric 200
ospf advertise 10.116.1.101 255.255.255.255

MSFC
O E2 10.116.1.101/32 [110/1] via 10.115.1.12, 00:04:15, Vlan115
O E2 10.115.1.101/32 [110/200] via 10.115.1.12, 00:04:15, Vlan115
O E2 10.115.1.102/32 [110/200] via 10.115.1.12, 00:04:15, Vlan115

Circuit VLAN115
description "CLIENT VLAN"
redundancy
ip address 10.115.1.12 255.255.255.0
ospf
ospf enable
ospf area 0.0.0.2
!
owner znaseh
content app1
vip address 10.115.1.101
protocol tcp
port 8081
add service s1
active
Advantages of the RHI Approach

• Support legacy application, that do not rely on a DNS infrastructure

• Very good re-convergence time, especially in Intranets where L3 protocols can be fine tuned appropriately

• Protocol-independent: works with any application

• Robust protocols and proven features
Limitations of the RHI Approach

• Relies on host routes (32 bits), which cannot be propagated all over the internet
• Requires tight integration between the application-aware devices and the L3 routers
• Internet deployments require route summarization (more on this later …)
HTTP REDIRECTION: TECHNOLOGY AND PRODUCTS
HTTP Redirection

The Main Idea

- Leverages the HTTP redirect function
  HTTP return codes 301 and 302
- Incoming client requests are redirected to the selected location
- The balancing decision happens after the DNS resolution and the L3 routing of the initial request has been completed
- Can be used in conjunction with other site selection mechanisms
HTTP-Redirect

After the DNS Resolution

http://www.cisco.com/

GET/HTTP/1.1
Host: www.cisco.com
HTTP/1.1 302 Moved
Location: www2.cisco.com

GET/HTTP/1.1
Host: www2.cisco.com
HTTP/1.1 200 OK

http://www2.cisco.com/

http://www1.cisco.com/
HTTP-Redirect Example

```
10.20.211.100.80 > 10.20.1.100.34589: FP 1:56(55) ack 287 win 2048 (DF)
0x0000  4500 005f 763c 4000 3e06 dd6c 0a14 d364 E.._v<@>.>l...d
0x0010  0a14 0164 0050 871d 7b57 aead ec1d 6b04 ...d.P..{W....k.
0x0020  5019 0800 8b1a 0000 4854 5450 2f31 2e30 P.......HTTP/1.0
0x0030  2033 3031 2046 6f75 6e64 200d 0a4c 6f63 .301.Found...Loc
0x0040  6174 696f 6e3a 2068 7474 703a 2f2f 7777 ation:.http://ww
0x0050  7732 2e74 6573 742e 636f 6d0d 0a0d 0a w2.test.com....
```
Advantages of the HTTP Redirect Approach

- Visibility into the client IP address
- Visibility into the request
  Possibility to distinguish:
  http://www.example.com/app1
  http://www.example.com/app2
- Inherent persistence to the selected location
Limitations of the HTTP Redirect Approach

- It is protocol specific (HTTP)
- Requires redirection to fully qualified additional names (www2, www3, ...)
- Clients can bookmark a specific location
OVERCOMING THE INHERENT LIMITATIONS
# GSLB and Disaster Recovery

## How Does It Work?

<table>
<thead>
<tr>
<th>DNS</th>
<th>HTTP Redirect</th>
<th>Route Health Injection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device Is Authoritative DNS</td>
<td>Device Returns HTTP 302</td>
<td>L3 Routing and Reconvergence</td>
</tr>
</tbody>
</table>

## Advantages

<table>
<thead>
<tr>
<th>DNS</th>
<th>HTTP Redirect</th>
<th>Route Health Injection</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Protocol Independent</td>
<td>• Visibility into Client IP and Request</td>
<td>• Protocol Independent</td>
</tr>
<tr>
<td>• Also GSLB</td>
<td></td>
<td>• No Need for DNS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Can Work w/ DNS</td>
</tr>
</tbody>
</table>

## Caveats

<table>
<thead>
<tr>
<th>DNS</th>
<th>HTTP Redirect</th>
<th>Route Health Injection</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Hard coded IPs</td>
<td>• HTTP only</td>
<td>• Only Disaster Recovery</td>
</tr>
<tr>
<td>• Subject to DNS Implementation</td>
<td>• Bookmarks</td>
<td>• Host Routes Dropped at Edge</td>
</tr>
<tr>
<td>• DNS Caching</td>
<td>• Resource Intensive</td>
<td></td>
</tr>
</tbody>
</table>
System Level Approach

• Each application and its respective user base varies

• Complete solution can be achieved by combining multiple technologies
  DNS
  Route Health Injection
  BGP / IP Anycast
  HTTP Redirect
REAL-WORLD DEPLOYMENTS
Real-World Deployments
Coast-to-Coast Fast Disaster Recovery

• Goal

Very fast disaster recovery between 2 remote data centers; minimize possibility of data center down
Non-DNS based application; client configuration cannot be changed
Support proprietary protocol: very long connections

• Solution

*Active-standby* content switches in each data-center; RHI across data centers

*Active* and *passive health monitoring* tuned to react as fast as possible
Real-World Deployments
Coast-to-Coast Fast Disaster Recovery

Data Center 1

Data Center 2

Partner

RHI

MSFC1

CSM1

FW1

R3

10.19.30.13 1000
10.19.30.14 2000

R4

10.19.30.13 1000
10.19.30.14 2000

FW2

R1

R2


MSFC2

CSM2

OSPF

BGP

© 2006 Cisco Systems, Inc. All rights reserved.
Real-World Deployments
Coast-to-Coast Fast Disaster Recovery

Implementation Details

• Route Health Injection is used on the CSM to dynamically announce VIPs to the MSFC
• MSFC running OSPF; other switches within the data center part of OSPF domain
• MSFC running BGP to connect to edge routers; EBGP peers across firewalls [Between MSFC and Edge routers]
• EBGP runs between edge routers of primary and secondary data center
• Firewall rules adjusted to allow BGP peers; TCP port 179
• OSPF in the Core
• Hold down timer will be used to prevent a failed center from immediately coming back on-line if the servers become active (probe failed value)
Real-World Deployments
Coast-to-Coast Fast Disaster Recovery

vlan 61 client
  ip address 10.19.71.12 255.255.255.0
gateway 10.19.71.1
  alias 10.19.71.11 255.255.255.0
vlan 73 server
  ip address 10.23.71.12 255.255.255.0
  alias 10.23.71.11 255.255.255.0

!probe ICMP icmp
  interval 2
  retries 3
  receive 2
  failed 65535

serverfarm APP_1
  nat server
  no nat client
  real 10.23.71.113
    inservice
  real 10.23.71.114
    inservice
  probe ICMP

vserver APP_1_VIP1
  virtual 10.19.12.13 tcp 0
  vlan 61
  serverfarm APP_1
    advertise active
    inservice

vserver APP_1_VIP2
  virtual 10.19.30.13 tcp 0
  vlan 61
  serverfarm APP_1
    advertise active
    inservice

router ospf 25
  network 10.19.0.0 0.0.255.255 area 0
  log-adjacency-changes

  redistribute static metric-type 1 subnets route-map Internal-Static

!router bgp 6117
  no synchronization
  bgp router-id 10.19.2.1
  bgp log-neighbor-changes
  bgp scan-time 5
  network 10.19.30.13 mask 255.255.255.255
  network 10.19.30.14 mask 255.255.255.255
  timers bgp 5 15
  neighbor 10.19.2.12 remote-as 6114
  neighbor 10.19.2.12 ebgp-multihop
  neighbor 10.19.2.12 update-source loopback0
  <SNIP>
Real-World Deployments
Varying Capacity Data Centers

• Goal
  Load balance applications within and across regions using DNS information
  Load balance a single APP across Sites based on availability and load
  High-volume, resilient and scaleable server farm for both static content
  Datacenters with different server capacity
  Max connections for servers needed
  Least Loaded required (ap-kal keep alive from GSS to CSS)

• Solution
  CSS deployed within the data centers for SLB
  One GSS deployed at each datacenter with ap-kal to the CSS
Real-World Deployments
Varying Capacity Data Centers

Internet

GSS-1

GSS-2

Active

Active
Real-World Deployments
Varying Capacity Data Centers

Implementation Details

- ACLs used on the GSS for protection against unauthorized access and attacks
- NS forwarder configured for unsupported record types – for example MX, AAAA etc
- Primary balance method is Round Robin
- For applications requiring site stickiness Source IP hash is used
  - D-Proxy hopping (location cookies with SSL termination)
    - Mega Proxy and large enterprises cause uneven load on the sites
- Static Proximity used for Intranet clients
- Firewall rules updated to allow kal-ap probe (UDP/5002) to go through between GSS and CSS
- Secondary clause used within the GSS rule with keepalive type to send a response to the client in case of both data center failure. This protects against negative caching
Real-World Deployments
Varying Capacity Data Centers

!  
**app-udp**
app-udp secure
app-udp options 10.14.80.21 encrypt-md5hash somepsswd
app-udp options 10.4.92.21 encrypt-md5hash somepsswd
!
owner customer.com
content web
  add service server3
  add service server4
  protocol tcp
  port 80
  vip address 10.18.80.155
**add dns www.customer.com**
advanced-balance sticky-srcip
active

vserver HR_JOBS_80
  virtual 10.14.80.31 tcp www
  serverfarm HR_JOBS
  replicate csrp connection
  **domain jobs.hr.customer.com**
  inservice
  !

vserver HR_REVIEW_80
  virtual 10.14.80.35 tcp www
  serverfarm HR_REVIEW
  replicate csrp connection
  **domain review.hr.customer.com**
  inservice
  !
**capp udp**
secure
options 10.14.80.21 encryption md5 somepsswd
options 10.4.92.21 encryption md5 somepsswd
!
Q and A
Recommended Reading

• Designing Content Switching Solutions: ISBN: 158705213X

    By Zeeshan Naseh,
    Haroon Khan