Layer 4-7 Server Load Balancing

Security, High-Availability and Scalability of Web and Application Servers
Foundry Overview

- Mission:
  Performance, High Availability, & Feature Leadership for Multilayer Switching (L2, L3, L4-7)

- Total Worldwide Customers: 6,000+
- Product & Corporate Awards: 50+

5th Consecutive Year of Net Profitability
Agenda

- Need for Load Balancing
- Load Balancing Technology
- Load Balancing Applications
- Benefits of Load Balancing
- Foundry Layer 4-7 Overview
Server Farm Challenge - #1

Poor Availability and Manageability

- Proliferation of IP applications complicates manageability and decreases availability
- Exponential cost of downtime and redundancy
- Disruptive server failover and maintenance
Server Farm Challenge - #2

Scalability Requires Bigger Servers

- Replace installed servers with larger and expensive ones
- Stranded capital in redundant servers
- Poor scalability with high cost and low ROI
- Compromise application and server security
- Super computer next?
Consequences of Current Approaches

- High risk of network and application downtime
- Business growth limited by server size
- Wasted capital and resources on redundant capacity
- Poor application performance, availability and security
Agenda

- Need for Load Balancing
- Load Balancing Technology
- Load Balancing Applications
- Benefits of Load Balancing
- Foundry Layer 4-7 Overview
Load Balancing Solution

- Leverage multiple commodity servers to create unlimited virtual capacity
- Scale server farms and applications to serve millions of clients
- Add network intelligence to improve server farm security and efficiency
Benefits of Load Balancing

- All servers utilized to maximum (ROI $$)$
- Ability to transparently add servers on-demand
- Full redundancy and transparency during failures
- Maximum application up time and massive scalability
- Superior service response time and performance
Load Balancing Overview

- Load Balancer receives all client requests
- Selects “best” server using real-time health and performance information
- Utilizes all available servers simultaneously
- Intelligently distributes load among servers
Load Balancing Fundamentals

VIP = Virtual IP
NAT = Network Address Translation

Client Message
Source IP = Client IP
Destination IP = Load Balancer VIP

VIP = 192.1.1.1
GW IP = 10.1.1.1

After NAT
Source IP = Load Balancer VIP
Destination IP = 10.1.1.10

• Private and secure server IP
• Clients use load balancer VIP
• Load Balancer performs NAT

GW = Gateway IP

Default Gateway = Load Balancer IP
Stateful Load Balancing

- Load balancer identifies session boundaries
- For duration of session, each client connection is
  - Assigned session entry in load balancer session table
  - Bound to same real server
  - All packets on connection sent to same real server

<table>
<thead>
<tr>
<th>Src. IP</th>
<th>Dest. IP</th>
<th>Src. Port</th>
<th>Dst. Port</th>
<th>Server</th>
</tr>
</thead>
<tbody>
<tr>
<td>188.1.1.100</td>
<td>192.1.1.1</td>
<td>100</td>
<td>80</td>
<td>RS1</td>
</tr>
<tr>
<td>188.1.1.100</td>
<td>192.1.1.1</td>
<td>101</td>
<td>80</td>
<td>RS2</td>
</tr>
<tr>
<td>188.1.1.101</td>
<td>192.1.1.1</td>
<td>200</td>
<td>80</td>
<td>RS1</td>
</tr>
</tbody>
</table>
Application and Server Health Checking

- Periodic health check requests sent to server
- Servers and applications removed when checks fail
- Health checks customizable
  - Layer 2/3 (ARP, Ping)
  - Layer 4 (TCP connections and UDP messages)
  - Layer 7 (HTTP, Application Specific, SSL, Scripted)
Delayed Binding Concept & Benefits

- Load balancer delays server-side connection
- Server selection based on “content” of client data packets
  - URL, Session ID, Cookie
- Optimal server utilization and highest availability

![Diagram showing the process of delayed binding]

1. TCP SYN
2. TCP SYN ACK
3. TCP ACK
4. HTTP Request
5. Select “best” server using L7 content
6. Complete Connection
7. Data Exchange
Layer-7 Content Switching

- Avoid replicating content and applications on all servers
- Increase overall server utilization and response time
- Use URL and HTTP hdr content to select “best” servers
  - URL full, prefix and suffix match
  - Browser type, device type and language code

**Diagram:**
- IP Hdr
- TCP Hdr
- HTTP Hdr
- Language Code

**URL Switch:**
- www.foo.com/*/gif
- /home.foo.com/*/htm
- www.foo.com/*/bin

**Network:**
- IP Network

**Servers:**
- English Clients
- Japanese Clients
- Servers for English
- Servers for Japanese

**HTTP Language Code:**
- English
- Japanese
Layer-7 XML Tag Switching

- Load balance to “right” server cluster using XML tags
- Optimize application performance for partners, suppliers and customers over extranet
- Extend load balancing to any XML-based application

```
<?xml version="1.0"?>
<root node>
  <node1 attr1=INV attr2=WIP>
  </node1>
</root node>
```

```
<?xml version="1.0"?>
<root node>
  <node1 attr1=TTKT attr2=WIP>
  </node1>
</root node>
```
Session Persistence Concept & Benefits

- Transaction spans multiple TCP or UDP connections
- Requires same server to handle all connections
- Load balancing at the “transaction” boundaries

Diagram:

1. Connection to Browse Book 1
2. Connection to Add Book 1 to Cart
3. Connection to Browse Book 2
4. Connection to Add Book 2 to Cart
5. Connection to Checkout Cart

Transaction persistence maintained
Session Persistence Mechanisms

• Layer 4 TCP connection persistence
  ➢ Source IP & port, Destination IP & port

• UDP session persistence using Layer 3/4
  ➢ Source IP & port, Destination IP & port
  ➢ Inactivity timeout used to age sessions

• Layer 7 Cookie switching
  ➢ Cookie inserted in the HTTP message
  ➢ All requests with same cookie switched to the same server
  ➢ Load balancers can insert cookies when servers do not

• SSL Session ID switching
Direct Server Return (DSR) - One Arm

- Reply traffic from server bypasses load balancer
- Load balancer processes only inbound requests
- Ideal for throughput intensive applications
  - Real-time streaming
  - Bulk data
  - FTP

Server Loopback IP = Load Balancer VIP

Layer-2 Switch
Load Balancer VIP = 192.1.1.1

Clients
IP Network
Load Balancer High Availability

- Two modes of high availability
  - Active-Standby (one load balancer as hot standby)
  - Active-Active (load sharing between load balancers)
- Stateful session failover maintains active sessions after failover and improves client performance
- Fully transparent to applications and clients
- GSLB offers site level protection
Agenda

• Need for Load Balancing
• Load Balancing Technology
• Load Balancing Applications
• Benefits of Load Balancing
• Foundry Layer 4-7 Overview
Global Server Load Balancing (GSLB)

- Return IP of best site in DNS responses to clients
- Gather site load information using GSLB protocol
- Deploy scalable GSLB within existing DNS infrastructure
ISP Link Load Balancing (LLB)

- Utilize all available ISP links simultaneously
- Intelligently balance traffic to achieve optimal utilization
- Gain leverage against ISPs for price and service
- Aggregate low-capacity links to create “fat” virtual links
Firewall Load Balancing (FWLB)

- Load balance among firewalls to scale and improve performance
- Transparently failover during firewall failure
- Protect network and servers during firewall outage
Server Farm Security from DoS Attack

- Protect against TCP SYN attacks on server farm
- Conserve resources for complete and valid connections
- Avoid using resources for pending connections
- Deny service to select user-configured source IPs
Agenda

- Need for Load Balancing
- Load Balancing Technology
- Load Balancing Applications
- Benefits of Load Balancing
- Foundry Layer 4-7 Overview
Improve Return on Server Investment

- Use servers of varying capacity and performance
- Distribute load based on server “weight”
- Leverage not replace installed servers
- Optimize cost by using diverse vendors
Dynamically Adjust Server Capacity

- Scale server capacity on demand
- Add and remove servers transparently
- Use slow-start to avoid overwhelming new server
- Perform server maintenance without impacting application performance
- Remove server after active connections serviced
Maximize Server Utilization with Multiple Applications

- Deploy multiple applications on each server for maximum utilization
- Select “best” server based on performance of each application
- Customize performance and scalability per application
Always-On Applications
Customizable Application Level Health Checks, High Availability and Security

- Server or application failure simply results in lower capacity
- Health checks for individual applications
- Select servers with best health and performance for each application
- Individual applications - not servers - out of service upon failed health checks
Differentiated Client Services

• Segment clients and offer differentiated service
• Customize performance and response time to meet different customers’ needs
• Differentiate clients using
  - Cookies and other layer 7 content
  - Source IP based access lists
Agenda

- Need for Load Balancing
- Load Balancing Technology
- Load Balancing Applications
- Benefits of Load Balancing
- Foundry Layer 4-7 Overview
Layer 4-7 Product Approaches

• Purpose-built Layer 4-7 devices
  ➢ Pro: Performance, Functionality and Simplicity
  ➢ Con: A new product to install

• Plug-in Layer 4-7 blade for Layer 2-3 switches
  ➢ Pro: Leverage an existing system
  ➢ Con: Complex to understand flows, bottleneck in performance, and lag in functionality
Meeting a Range of Customer Needs

**Entry Level**
Single-purpose device supporting one application in a given deployment

**Mid Level**
Fixed configuration supporting multiple applications on a single device without high cost

**High Performance**
High-density high-performance applications with high scalability

Relative Performance

- 12X
- 6X
- 2X
- 1X

Relative Price

- ServerIron XL
- ServerIron 100
- ServerIron 400
- ServerIron 800
Full Suite of Layer 4-7 Applications

- Server Load Balancing
- Global Server Load Balancing
- Firewall Load Balancing
- ISP Link Load Balancing
- Transparent Cache Switching
- IronShield™ Server and Application Security
Thank You

Don’t Plan for Yesterday.