Issue Report: Stabilization of Load Balancing Tier for Libero Home Page Application

Reported By: Peter Gale, Joyent Support

Report Date: 05-Mar-12

Abstract Summary:

The following report summarizes the production problems experienced by Libero for the Home Page application running on the Joyent SDC-based Libero Cloud, specifically with the load-balancer tier which is using a clustered-implementation of the Riverbed Stingray product (formerly Zeus traffic manager.) Both organizations placed significant focus and effort on identifying the problems, and ultimately resolving the issue with the eventual cooperation of Riverbed. Although the period did result in a number of questions and issues raised, specifically about the Riverbed load-balancer, through close collaboration and teamwork, Joyent and Libero/ITNet were able to minimize any impact to the production service, and ultimately drive to a low-level code analysis that identified the root cause of the problem. This report summarizes the efforts taken, the problem identified, the chronology of events, and a summary of the related support tickets.

Beginning in mid-2011, Libero started testing the implementation of Riverbed Stingray for the purposes of using Stingray as a software-based load balancer for the Libero Home Page application. Beginning in November 2011, Libero went live on the Joyent SDC-based Libero Cloud, using Stingray as the load balancing solution for the Home Page application. Shortly after the application was moved into production, Libero experienced several problems with the Stingray product including listen drops and complete hangs of the HP application. The Stingray problems were compounded by a handful of kernel panics seen in the compute nodes on which the Stingray product was running.

In the first week of December 2011, both teams (Joyent and Libero/ITNet) enacted a war room protocol to track the issues relating to the Stingray implementation and Home Page stabilization. Over the course of the next 3 months, the team tracked progress through daily issue reporting and daily (eventually 3x week) conference calls.

Significant resources were brought to bare to address the issue, including:

- Set-up of separate test environment on BORA test pod with load testing
- Set-up of test environment for larger scale load testing on dedicated compute nodes within Levante production pod
- Implementation of tracing, cloud analytics and other diagnostics to help triage the problem and narrow the issue to specific process failures
- Multiple on-site visits for focused collaborative troubleshooting
- Code changes and product upgrade
- Three-way engagement with Riverbed Support, and ultimate on-site engagement with Riverbed, Joyent and Libero/ITNet which led to final resolution of the problem.

For an extended period, both Joyent and Libero/ItNet had significant challenges getting engagement from Riverbed Support. There was very little traction on the tickets that were submitted, and overall, we saw limited response and engagement. Following several escalations, and a final escalation from Libero

CTO, Riverbed became engaged, and appointed a named support engineer, and ultimately provided onsite engineering in support of the problem.

The problem was identified as a cache that was getting too large and slowing down IP processes. The team identified a workaround to reboot the nodes to clear cache and a code change in the SmartOS platform image was identified to change the way the cache is handled.

The problem was long, and challenging; however, both sides ultimately represented great team work, and a high level of engagement, and significantly advanced our capabilities for additional problem solving and proactive work in the future.

Problem Description

Symptoms:

- Gradual increasing load average
- Unexpected fail-over of a single Zeus node (diagnostics give no indication why fail-over should occur)
- Fail-over hangs during un-plumbing of interface
- Traffic backs up causing time-out.

Problem description:
The IP layer in the kernel has a cache that steadily grows as new IP addresses are connected to the page. Entries in the cache are only deleted when the system is under memory pressure, (which doesn't occur on these systems.) When a new connection request comes in at the NIC, the interrupt handler searches this cache. The search was taking a long time (2-3 msecs per search) The result is that the NIC driver is not responding quickly enough to handle the incoming requests, which would time out and drop, and the processor handling interrupts from the NIC was running 80+% in system mode causing normal processing to queue up behind it (causing high load averages.)

Resolution:
The agreed upon workaround is to reboot the Stingray nodes, staggered over several days, which causes the cache to be emptied. We have also identified a code change that changes the way the cache is handled.

Chronology

Period	Events
Oct-Nov	• Libero begins pre-production testing of Stingray for home page application • Libero goes live with Home Page application on Libero Cloud using 2-node (48 Gig) Stingray cluster • Shortly after go-live Libero begins to experience problems with Stingray cluster (listen drops) and complete hangs of the Stingray compute nodes.
1H December	• As a result of production issues and instability of the load-balancer tier, Joyent and Libero/ITNet enacted a war room protocol to put

	high priority on tracking the problem and the progress being made, including daily calls and management reporting. ● Established test environment on BORA - was able to reproduce the problem. ● Max was able to pinpoint exactly what was happening with the problem, however, not able to identify *why* the problem was occuring. ○ Zeus cluster fails-over ○ Failover hangs during unplumbing of interface ○ Traffic bacup causing time-outs ● Recommendation to increase the number of Stingray nodes (from 2 to ultimately 8) ● Max B onsite to implement tracing within the problem and to try and reproduce the problem on Levante. ● Teams addressed "limited bandwidth" license issue, by securing an unlimited license from Riverbed. ● Riverbed Stingray cluster configuration reviewed by Joyent expert ● The Libero Cloud (Bora and Levante) were upgrade to 6.5.1 OS Update 1 to address the fixes that were identified for the kernel panics that we saw during this period.
2H December	● Joyent recommendation to balance traffic evenly across cluster. Initially an IP address problem that we opened with Riverbed prevented us from doing this until early January. ● Testing environment established on Levante to reproduce the problem by directing more production-like load on new test environment. ● Prior to holiday break, LIbero/ITNet split the load, with static content being served off of Libero Cloud, and dynamic content being served from Milan datacenter. This reduced the load on Libero Cloud reducing risk of production impact during holiday period.
1H January	● Max B onsite again during week of January 9th to examine problem on test environment on Levante ● Additional diagnostics, and tracing and analysis from Cloud Analytics to look at problem. ● Full workload (dyamic and statis) was returned to the Libero Cloud, with traffic balanced across all cluster nodes. ● Very little engagement from Riverbed despite a number of tickets opened during this period.
2H January	● Following repeated escalations, and an ultimate escalation from Libero CTO, Riverbed support became fully engaged, appointing a designated senior support engineer.
1H February	● Number of actions following Riverbed enagement, including upgrade to 8.1, and implementation of a number of configuration changes. ● Riverbed onsite with Max and Libero/ITNet team to collaborative examine the problem. ● Max discovers a cache in the IP code in the kernel getting very large causing a gradual slow down in IP processes. ● This resulted in ultimate identification of the problem, current workaround, and ultimate fix.

2H February	• Libero begins to bring additional services online, and Libero Cloud (Bora and Levante) successfully upgraded to 6.5.3 • Recommendation to reduce memory sizing on all Stingray nodes.

Primary Related Support Tickets

ZD Ticket #	Request Date	Status	Description
1464	10-Nov-2011	Closed	Kernel panic in SRVJ-18 (Fix identified and implemented)
1508	22-Nov-2011	Closed	Kernel panic in SRVJ-30 (No useful dump)
1509	22-Nov-2011	Closed	Kernel panic in SRVJ-15 (Problem related to a known issue that was already addressed; fix had not yet been applied)
1556	02-Dec-2011	Merged with 1589	Loss of service on HP production site for a few minutes (Zeuses needed to be restarted)
1589	12-Dec-2011	Closed	This was the primary tracking ticket for HP Stabilization actions and analysis
1620	19-Dec-2011	Closed	Tracking ticket for Riverbed Ticket 195986 (Zeus failover issues)
1627	20-Dec-2011	Closed	Configuring Zeus SMs with more than 19 Ips (configuration changes/charac limit addressed)
1753	20-Jan-2012	Pending	Tracking ticket for recommendations on appropriate Zeus node sizing.
1757	23-Jan-2012	Pending	Tracking ticket for Riverbed Ticket 202436 (Zes config review)
1778	27-Jan-2012	Closed	Tracking ticket for Riverbed Ticket 202443 (Benchmarking site, unexpected behavior)
1820	06-Feb-2012	Closed	Tracking ticket for Riverbed Ticket 206562 (Configuration change review/confirmation)