Progress®
Kemp® LoadMaster®

Technical Note Resource Utilization

Generated on: 21 April 2026

Copyright

Visit the following page online to see Progress Software Corporation's current Product Documentation Copyright Notice/Trademark Legend: Product Documentation Copyright Notice & Trademarks | Progress

Table of Contents

Chapter 1: Introduction 5

Intended Audience 5

Chapter 2: Resources 6

CPU 6

RAM 6

Network Load 7

Cavium TPS 7

Chapter 3: Resource Consumption 8

Force Layer 7 Services 9

Idle Connection Timeout 9

Native Layer 7 Services 9

Layer 7 Persistence 9

Caching 9

Compression 10

Intrusion Prevention System (IPS) 10

Content Switching and Header Modification 10

SSL Transactions Per Second (TPS) 10

SSL Bulk Encryption 10

Inter-HA L4 Connection Updates 11

Inter-HA L7 Persistency Updates 11

Drop on Real Server Failure 11

Chapter 4: References 12

1

Introduction

The LoadMaster contains a wealth of features which utilize various hardware resources. While the LoadMaster is tested in laboratory conditions to obtain hardware capabilities, features are tested in isolation and the interactions between features is not reported. This document is intended to give users a better understanding of what resources each of these major features utilize and what interactions can be expected when utilizing multiple features.

Contents

- Intended Audience

Intended Audience

This document is intended to be read by anyone who is interested in learning about LoadMaster resource utilization.

2

Resources

The sections below outline the resources that the LoadMaster uses.

Contents

- CPU
- RAM
- Network Load
- Cavium TPS

CPU

CPU indicates the processing power of an appliance. CPU is measured in cycles. These cycles represent the time utilized to process a piece of data. Depending on the model of LoadMaster, the processor may have multiple cores. This allows multiple processes to be performed simultaneously. However, this does not automatically double the processing power of a unit. Processes must be specifically designed to utilize multiple cores. Certain processes in the LoadMaster are multi-threaded while others are not. This can put constraints on certain processes within the LoadMaster.

RAM

RAM indicates the amount of memory available to a LoadMaster. This is primarily used to keep track of connections and persistence records. A certain amount is also consumed by the operating system. RAM can

also be used for caching operations.

Network Load

Network load indicates the level of activity both on a particular interface and as a global value. All current LoadMaster models contain 10/100/1000 interfaces which can pass a maximum amount of 1Gbps. The other constraint on network load is the global throughput value. This value represents the maximum amount of network traffic which can be processed at a given time. Increases in network load have a consequence of driving up CPU utilization as each packet requires processing in order to determine what needs to be done with it.

Cavium TPS

On hardware LoadMaster models containing a Cavium SSL ASIC card, there is an additional metric to consider called Cavium TPS. An ASIC is an Application-Specific Integrated Circuit; this provides acceleration for SSL handshakes. This allows SSL offloading to be dramatically accelerated while also reducing CPU load versus comparable models without the ASIC. Each model has a limited SSL TPS which can be accomplished with the card. This will limit the amount of new SSL connections per second regardless of other metrics.

3

Resource Consumption

Before discussing individual features and their impact on performance, it is important to understand how resources are utilized by a Virtual Service and by Real Servers. Each Virtual Service consumes a small amount of memory to save state information about the server. Each Real Server requires a small amount of additional CPU to perform health checks as well as 512KB of RAM for IP mapping tables.

Contents

- Force Layer 7 Services
- Idle Connection Timeout
- Native Layer 7 Services
- Layer 7 Persistence
- Caching
- Compression
- Intrusion Prevention System (IPS)
- Content Switching and Header Modification
- SSL Transactions Per Second (TPS)
- SSL Bulk Encryption
- Inter-HA L4 Connection Updates
- Inter-HA L7 Persistency Updates
- Drop on Real Server Failure

Force Layer 7 Services

Forced Layer 7 services use slightly more RAM per connection than a Layer 4 counterpart due to the additional pieces of data that need to be recorded. Additionally, there is a slight increase in CPU utilization due to the fact that the LoadMaster is acting as a full proxy and as a result needs to process more data about the connection.

Idle Connection Timeout

The Idle Connection Timeout is a way of garbage collecting discarded connections. Since Layer 7 services act as full proxies, if a client discards its connection without informing the LoadMaster the connection will not be closed. As a result, it continues to take up RAM on the LoadMaster. Increasing the Idle Connection Timeout may have the result of increasing these dead connections and leading to increased RAM utilization. Lowering this will ensure that these connections are closed in a reasonable amount of time.

It is recommended that the Idle Connection Timeout be set as low as possible and as allowed by both client and server. Please consult application-specific documentation or a Customer Support Engineer. As a general rule of thumb, web services work well with the LoadMaster default value of 660 seconds.

Native Layer 7 Services

Native Layer 7 services are an extension of Force Layer 7 services with the difference being that some Layer 7 functionality has been enabled. Regardless of whether Layer 7 persistence, caching, IPS or Content Switching is enabled, the LoadMaster must be observing the client-server interactions in order to follow the conversation and be an effective Layer 7 device. This naturally increases both CPU and RAM usage due to the processing and tracking components of this role.

Layer 7 Persistence

Layer 7 Persistence is any persistence type other than Source IP. These require looking into application traffic and identifying persistence markers in this data. The parsing and checking for such persistence increases CPU utilization. In addition, persistence tables for Layer 7 data are generally larger than a Layer 4 persistence table meaning that RAM utilization will also rise.

Caching

Caching allows the LoadMaster to store server content locally for fast retrieval and to reduce server load. Since cached content is stored in RAM, it will naturally drive up RAM utilization. Caching is allowed to utilize up to 20% of the total RAM available to LoadMaster. Utilization of this can then be controlled on a per-service basis as a percentage of the global value.

Caching does increase CPU utilization, but not significantly. It requires each request and response to be evaluated for whether the request should be pulled from cache and the response stored in cache.

Compression

Compression allows the LoadMaster to compress uncompressed responses from servers. The main resource utilized is CPU. This can be a non-negligible amount, depending on the situation.

It is recommended that Compression and Caching be used in conjunction as this will cause the compressed response to be cached. This reduces RAM utilization to store the response and reduces CPU utilization by requiring that the response only be compressed once.

Intrusion Prevention System (IPS)

The LoadMaster IPS system operates by comparing the URL requested with a list of known attack vectors. This requires predominantly CPU utilization. Of course the SNORT rule set must also be loaded into RAM beforehand, so there is an amount of RAM that will also be utilized.

Content Switching and Header Modification

Content Switching and Header Modification both examine the HTTP stream and make decisions based on the content therein. The result can be directing to specific servers or modifying the stream based on rules. These consume some more memory. The main drain on resources is CPU load. Depending on the quantity and complexity of rules, this can dramatically impact performance. If more than 5 rules are being applied to a service it is strongly recommended that you contact support to check if things can be simplified in some way.

SSL Transactions Per Second (TPS)

SSL TPS can impact one of two metrics depending on the hardware present in a LoadMaster. If a LoadMaster contains a Cavium SSL ASIC card, SSL TPS will impact the Cavium TPS count. If this value is at 100%, there may be a backlog of incoming SSL connections to be processed. This can lead to timeouts and rejected connections under sustained loads. If the LoadMaster does not contain an ASIC, SSL TPS will impact CPU resources. As TPS climb towards the rated TPS for the model, CPU utilization will approach 100%. In addition to creating a backlog of SSL connections and eventually timeouts, this also impacts any other services which are vying for CPU resources.

SSL Bulk Encryption

SSL bulk encryption is everything which is encrypted within an SSL connection after the initial handshake. Regardless of the presence of an ASIC, bulk encryption is carried out on the main CPU. Depending on the profile of the HTTPS connections, it is possible to heavily utilize the CPU in this manner without approaching the rated SSL TPS for a particular model. This can happen if large amounts of data are being transferred over a relatively small number of connections.

Inter-HA L4 Connection Updates

Inter-HA L4 connection updates increase both CPU and network load. CPU is utilized to process the data and send updates to the partner device. Network load is naturally consumed in the process of sending the updates. This can be mitigated at the interface level by migrating the multicast interface to a direct connection between the HA pair. This is always the recommended approach when enabling this setting. Failing to implement this can quickly saturate an interface since increases in production traffic will increase the number of updates needed. This can be an insidious problem as during periods of low traffic there will be no indication of this problem. Please consult with a Customer Support Engineer for further information and guidance if you are interested in implementing this feature.

Inter-HA L7 Persistency Updates

Inter-HA L7 persistence updates increase both CPU and network load. CPU is utilized to process the data and send updates to the partner device. Network load is naturally consumed in the process of sending the updates. This can be mitigated at the interface level by migrating the multicast interface to a direct connection between the HA pair. This is always the recommended approach when enabling this setting. Failing to implement this can quickly saturate an interface since increases in production traffic will increase the number of updates needed. This can be an insidious problem as during periods of low traffic there will be no indication of this problem. Please consult with a Customer Support Engineer for further information and guidance if you are interested in implementing this feature.

Drop on Real Server Failure

Drop on Real Server Failure allows connections to be immediately closed upon failure of the Real Server. This does introduce a fair amount of CPU utilization at the time of the event. Typically, this is non-intrusive to other services as it is a one-time event. However, if servers are rapidly coming in and out of service, this can introduce enough CPU load that it becomes an issue affecting all load balanced services.

We hope this guide has been helpful and informative. If you have any specific questions or if you have any feedback or suggestions for this document, please send them to a Customer Support Engineer who will direct your feedback to the appropriate teams.

4

References

For further details on the WUI fields, refer to the following document:

Web User Interface (WUI) Configuration Guide