**Scale system infrastructure to support High TCP on RE : 16K+ (kernel)**

**Software Functional Specification**

**Ananda Kumar <anandmr@juniper.net>**

|  |  |  |  |
| --- | --- | --- | --- |
| Rev # | Date | Revised by | Comments |
| 1.0 | 5th Dec 2015 | Ananda Kumar M.R | Initial version |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

TEMPLATE REVISION HISTORY AND GUIDELINES:

The revision history for this template can be found at the end of this document.

Table of Contents

1. Introduction 6

1.1 Reference 6

1.2 RLI List 6

1.3 Feature Parity Traceability 6

2. Functionality 7

2.1 Goals 7

2.2 Exceptions 7

2.3 Assumptions 8

2.4 Functional competitive data 8

2.5 APIs/Messages 8

2.6 Manageability 9

2.7 Examples or Interaction Descriptions 9

2.8 Supportability (Serviceability, Diagnose-ability and Fault Handling) 9

2.9 Dependencies and Interactions with other Components in the System 10

2.10 Legal Considerations 10

3. Other Requirements 10

4. System Resource Estimation 13

4.1 Performance Related Resources 13

5. Scaling and Performance 14

5.1 Target Scaling 14

5.2 Target Performance 14

6. Compatibility Issues 15

7. Security Considerations 16

8. Platforms Supported 17

9. Common Infrastructure 18

9.1 High Availability (HA) 18

9.2 Aggregated Ethernet/ SONET/ IRB Support 18

9.3 Services/JSF (JUNOS Services Framework) Impact 18

9.4 Multi-Chassis Support 18

9.5 64-Bit Support 18

9.6 IPv6 Support 18

9.7 Logical System Support 18

10. SDK Impact 19

10.1 SDK Customer Usage 19

11. JUNOS Ready Software considerations 20

12. Notes 21

13. Glossary 22

14. Design Specification exception 23

15. Functional Specification Approver Checklist 24

Functional Specification Document Checklist

|  |  |  |  |
| --- | --- | --- | --- |
| Section | Section Title | Recommendation | Compliance |
| 1 | Introduction | MANDATORY |  |
| 1.1 | Reference | MANDATORY |  |
| 1.2 | RLI List | MANDATORY |  |
| 1.3 | Feature Parity Traceability | OPTIONAL |  |
| 2 | Functionality | MANDATORY |  |
| 2.1 | Goals | MANDATORY |  |
| 2.2 | Exceptions | MANDATORY |  |
| 2.2.1 | Caveats | MANDATORY |  |
| 2.2.2 | Limitations | MANDATORY |  |
| 2.2.3 | Non Goals | MANDATORY |  |
| 2.3 | Assumptions | MANDATORY |  |
| 2.4 | Functional Competitive Data | OPTIONAL |  |
| 2.5 | API/Messages | MANDATORY |  |
| 2.6 | Manageability | MANDATORY |  |
| 2.6.1 | CLI Config | MANDATORY |  |
| 2.6.2 | CLI Commands | MANDATORY |  |
| 2.6.3 | SNMP | MANDATORY |  |
| 2.6.4 | Syslog- ERRMSG | MANDATORY |  |
| 2.6.5 | Software Feature Licensing | MANDATORY |  |
| 2.6.6 | Software Packaging | MANDATORY |  |
| 2.6.7 | J-Web Quick Configuration and Monitor Screen | MANDATORY |  |
| 2.6.8 | Intergration with Management Systems | MANDATORY |  |
| 2.7 | Examples or Interaction Descriptions | OPTIONAL |  |
| 2.8 | Serviceability and diagnose-ability | MANDATORY |  |
| 2.9 | Dependencies and Interactions with other Components in the System | MANDATORY |  |
| 2.10 | Legal Considerations | MANDATORY |  |
| 3 | Other Requirements | OPTIONAL |  |
| 4 | System Resource Estimation | MANDATORY |  |
| 4.1 | Performance Related Resources | MANDATORY |  |
| 5 | Scaling and Performance | MANDATORY |  |
| 5.1 | Target Scaling | MANDATORY |  |
| 5.2 | Target Performance | MANDATORY |  |
| 6 | Compatibility Issues | MANDATORY |  |
| 7 | Security Considerations | MANDATORY |  |
| 8 | Platforms Supported | MANDATORY |  |
| 9 | Common Infrastructure |  |  |
| 9.1 | High Availability (HA) | MANDATORY |  |
| 9.1.1 | Graceful RE Switchover (GRES), ISSU and NSSU Impact | MANDATORY |  |
| 9.1.2 | NSR Impact | MANDATORY |  |
| 9.2 | Aggregated Ethernet/ SONET/ IRB Support | MANDATORY |  |
| 9.3 | Services/JSF (JUNOS Services Framework) Impact | MANDATORY |  |
| 9.4 | Multi-Chassis Support | MANDATORY |  |
| 9.5 | 64-Bit Support | MANDATORY |  |
| 9.6 | IPv6 Support | MANDATORY |  |
| 9.7 | Logical System Support | MANDATORY |  |
| 10 | SDK Impact | MANDATORY |  |
| 10.1 | SDK Customer Usage | MANDATORY |  |
| 11 | JUNOS Ready Software considerations | MANDATORY |  |
| 12 | Notes | OPTIONAL |  |
| 13 | Glossary | OPTIONAL |  |
| 14 | Design Specification exception | OPTIONAL |  |

# Introduction

The purpose of this effort is to identify and address all the bottlenecks in JUNOS kernel infra that limits the scaling of TCP/JSR connections on RE to the maximum possible value.

Key motivation behind this effort: In case of BGP based on information from the field competitors are claiming support for 8K BGP sessions. At this point Juniper scale is 4K BGP sessions (IPv4 and IPv6) with NSR. This is primarily due to the limitation on numbers file/socket that can be opened in legacy JUNOS. This limitation has been overcome in JUNOS through OCCAM project by updating the FreeBSD to latest version.

## Reference

Tracking PR: 1054561

## RLI List

|  |  |
| --- | --- |
| **RLI No** | **Description** |
| 23560 | Scale system infrastructure to support High TCP on RE : 16K+ (kernel) |

## Feature Parity Traceability

Not Applicable.

# Functionality

At this point of time we are still in the process of identifying the limitations in JUNOS kernel infra that is blocking in scaling the number of TCP/JSR connections on RE. Since RPD/BGP is the major consumer of this effort, below is the approach adopted to achieve this goal.

* Create more number of BGP sessions (4k to start with) between two routers.
* Configure NSR and perform couple of iterations of RE switchover and monitor the BGP sessions with various BGP HT values.
* If the NSR switchover is successful without BGP session flaps, Increase the number of BGP session (by 4k) and repeat above steps.

Based on the scale reached so far, there were no issues found in kernel TCP/JSR infra but there are limitations/issues observed in the other modules and sub-systems on RE at various scale. More details on this are discussed in section 5 of this document.

1. **Proposed implementation in detail:**

As mentioned above, as of now, this is a test only effort focused towards identifying the limitations in kernel TCP/JSR infra that blocks in scaling the TCP/NSR connections to maximum supported value. This section will be populated as and when the scaling related issues are found in TCP/JSR infra and root caused.

1. **Debugging ability**

Same as above.

## Goals

* To identify all the bottlenecks in JUNOS TCP/JSR kernel infra that limits the scaling of TCP/JSR connections on RE to the maximum possible value.
* If any issues are identified in TCP/JSR infra that impacts the scaling, update those limitations in this document, root cause them and depending on the complexity and priority of the issue, address them as part of this RLI or another RLI/PR in the subsequent release.

## Exceptions

### Caveats

### Limitations

* May not be possible to address/fix all the limitations in JUNOS TCP/JSR kernel infra (in case found at the later stage of testing with higher scale values) with in the timelines of this RLI. If any such issue are found, they will be scoped separately and fixed either as PR or RLI based on the complexity and the effort required to address those issues.

### Non Goals

* Any Non TCP/JSR infra issues found during testing will not be fixed by TCP/JSR team. PR’s will be created and assigned to respective team/group to fix any such issues.

For ex: If BGP is not able to scale to the extent that kernel can support due to limitations in RPD infra or if system misbehaves after NSR switchover due to some limitations in other modules like GRES or chassisd infra, such issues will be assigned to respective team/group.

* Performance related issues if observed during scale testing may not be addressed as a part of this RLI.

For ex: If BGP sessions flaps after NSR switchover with lower HT values at higher scale due to some limitations in kernel TCP/JSR infra. Such issues may not be addressed as a part of this RLI. Any such issues found during subsequent testing will be updated in this document and has to be addressed separately.

## Assumptions

None.

## Functional competitive data

None.

## APIs/Messages

## Manageability

### CLI Config

Not Applicable.

#### CLI Config Details

Not Applicable.

### CLI Commands

Not Applicable.

#### CLI Command Details

Not Applicable.

#### JUNOScript

Not Applicable.

### SNMP

Not Applicable.

### Syslog – ERRMSG

None.

### Software Feature Licensing

Not Applicable.

### Software Packaging

Not Applicable.

### J-Web Quick Configuration and Monitor Screen

Not Applicable.

### Integration with Management Systems

## Examples or Interaction Descriptions

## Supportability (Serviceability, Diagnose-ability and Fault Handling)

### Serviceability and Diagnose-ability

This section will be populated as and when limitations in kernel TCP/JSR infra that blocks in scaling the TCP/NSR connections are identified and root caused.

### Fault Handling

Same as above.

## Dependencies and Interactions with other Components in the System

## Legal Considerations

### Third-Party Materials:

Are there any **Third-Party (non-Juniper) Materials** incorporated into the SW? “**Third-Party Materials**” include (1) Open Source Software, technology, and materials (“OSS”), and (2) Third-Party commercial technology, materials, and code (including code embedded in an ASIC).

**No**

**If Yes,** submit a [**Third-Party Tracking Chart**](http://www-int.jnpr.net/legal/ip/docs/third_party_tracking_chart.xls) created by your [**Codemaster**](http://www-int.jnpr.net/legal/ip/docs/list_of_codemasters.pdf)listing all such Third-Party Materials to your [BU IP Technology Transaction Attorney](http://www-int.jnpr.net/legal/contacts/byFunction.html?IP). Your “**Codemaster”** is the team member assigned to track all Third-Party Materials for a Functional Specification.

**All Third-Party Materials must be approved prior to Functional Specification approval**, as follows:

* All OSS must be approved in accordance with [**Juniper’s Inbound OSS Policy**](http://www-int.jnpr.net/legal/ip/docs/inbound_oss_policy.pdf); and
* All Third-Party commercial technology must be approved by Legal in accordance with Juniper’s [**Third-Party Commercial Technology Policy**](http://www-int.jnpr.net/legal/ip/docs/third_party_commercial_technology_policy.pdf).

### Export Control Matters

### Encryption

All encryption features/content must be reviewed and all US/International legal requirements met prior to HW/SW release. Such review/approval may take **up to 50 days**. “HW/SW” includes a system, card, IC device, software OS or application or module embedded therein, or revision of any of the foregoing.

|  |  |  |
| --- | --- | --- |
| **#** | **Encryption Questions** | **Yes/No** |
| a | Is there cryptography incorporated in, controlled by, or linked (dynamically or otherwise) to HW/SW? | No |
| b | Is cryptography used for key exchange, authentication, or message integrity? | No |
| c | Does cryptography operate in the HW/SW to decrypt already-encrypted content? | No |
| d | Is cryptography used to encrypt or decrypt network management, monitoring or administrative traffic? | No |
| e | Is “publicly available” or OSS-derived cryptography used? | No |
| f | Is the HW/SW designed to use encryption embedded in another product? | No |
| g | Is there cryptography in or used by the HW/SW that was furnished, or developed, by a Third-Party? | No |
| h | Does any hardware component incorporated in the HW/SW implement (or accelerate computations for) any cryptographic algorithm or security protocol? | No |

If you answered “Yes” to any of the above questions, then there is encryption and you must complete the [Encryption Checklist](https://www.t-baccess.com/juniper/engineering/default.aspx) at <https://www.t-baccess.com/juniper/engineering/default.aspx>.

### High-Speed IO Technology

Is the product, subsystem, component or feature designed to, or does it operate to enable the system in which it is incorporated to support >120Gbps per channel I/O? **No**

If Yes, then email [kniven@juniper.net](mailto:kniven@juniper.net) and cc [compliance\_helpdesk@juniper.net](mailto:compliance_helpdesk@juniper.net) with “URGENT–High Speed Interconnect Issue” as the Subject Line. In the email, identify the project code name/feature description and a brief description of the product or feature.

For questions about encryption or the [Encryption Checklist](https://www.t-baccess.com/juniper/engineering/default.aspx), email [new\_product\_export\_classification@juniper.net](mailto:new_product_export_classification@juniper.net). For all other questions, email [legal-ip checklist@juniper.net](mailto:legal-ip%20checklist@juniper.net).

# Other Requirements

None as of now.

# System Resource Estimation

As mentioned above, as of now, this is a test only effort focused towards identifying the limitations in kernel TCP/JSR infra that blocks in scaling the TCP/NSR connections to maximum supported value. This section will be populated as and when the scaling related issues are found in TCP/JSR infra and root caused.

.

## Performance Related Resources

Performance related issues will not be addressed as a part of this RLI.

# Scaling and Performance

The aim of the entire effort was to investigate the challenges faced by kernel TCP/IP and JSR sub-system while scaling bgp sessions. The study was aimed to study the threshold limits imposed by TCP sub-system on the number of open TCP sockets with socket replication. Along with that, we wanted to check the performance of the system under when we reach the threshold limits.

When socket replication is functional, for each TCP socket opened on the primary, we have a pair of secondary sockets opened on the backup. When we try to scale BGP sessions with NSR configured, the memory and the processing requirements on the backup is just the double of that on the master. As we go up scaling BGP sessions, not only the memory requirements increase but we need more processing power as each socket first replicates data/packets to the backup and then processes it further down/up the stack. System performance was also monitored closely with increased scale. Any kernel issue observed during scaling that imposes challenge, need to be addressed.

BGP was configured minimal to avoid restrictions imposed by it’s own heaviness so that we could purely gauge kernel based scaling and performance.

To start with, we configured 4k BGP sessions with 360 secs Holt Time with no routes configured per VPN. BGP sessions was planned to be scaled by 4k with the iteration. Initial few hurdles came from console logging which retarded the system performance on it’s own. In this way, we could scale upto 16k. At 16k bgp scale, very occasionally, few BGP session flap was observed after NSR switchover. Reason for those BGP session flaps were:

a) RPD was slow in reading the data from the receive socket buffer.

b) Also noticed that RPD on DUT was sending few EOF message to BGP peer

To reach the scale of 16k BGP sessions, there weren’t any direct limitations from kernel TCP/JSR infra.

Also, tested these two cases at 16k scale (after all 16k sessions are replicated on backup)

a) Crashing RPD on master: In this case, all sessions were flapped. This was be due to ‘switchover-on-routing-crash’ feature was not working on OCCAM build. There is open PR 1003484 to address this issue.

b) Restarting master RE: In this case, some of the BGP session’s failed immediately when master RE was rebooted and that was again due to RPD on DUT sending EOF message to peer. And some of the sessions flapped few seconds after the RE switchover and they were mainly due to packet drops.

The next step was to increase the scale to 20k BGP sessions. With 20k scale following are the observation:

1. 20k BGP sessions established successfully and 20k sessions got replicated onto backup successfully (though it took more time to replicate them)
2. Upon NSR switchover (manual) system was sometime stable sometimes unstable.
3. Lot of sessions were flapping after switchover and the reasons for flaps were
   * RPD not reading the data from the receive socket buffer on some of the BGP sessions
   * Peer not getting TCP ACKs due to packet drops
   * RPD on DUT closing connections (by sending Cease code 6 to peer)
   * Ksyncd core was noticed during switchover and ksyncd was continuously logging kernel back trace and filling the hard disk. PR1056859 is created to address this.

## Target Scaling

Based on the testing conducted so far there were no issues found in kernel TCP/JSR infra in scaling up to 16k BGP sessions. Target is to continue testing by increasing the scale and find the maximum limit that kernel infra can support within the given physical resource limit (CPU & RAM).

## Target Performance

As mentioned above performance related issues will not be addressed as a part of this RLI.

# Compatibility Issues

At this point of time, this is a test only effort focused towards identifying the limitations in kernel TCP/JSR infra that blocks in scaling the TCP/NSR connections to maximum supported value. This section will be populated as and when the scaling related issues are found in TCP/JSR infra and root caused.

# Security Considerations

At this point of time, this is a test only effort in identifying the limitations in kernel TCP/JSR infra that blocks in scaling the TCP/NSR connections to maximum supported value. This section will be populated as and when the scaling related issues are found in TCP/JSR infra and root caused.

# Platforms Supported

Since the scaling tests are conducted on OCCAM image, Only MX platform is targeted as part of this RLI.

# Common Infrastructure

## High Availability (HA)

### Graceful RE Switchover (GRES), ISSU and NSSU Impact

If there are any issues/limitations observed in GRES infra while trying to achieve the scaling goals discussed above, same will be shared with the GRES infra team. ISSU/NSSU tests will not be focused as part of this effort.

### NSR Impact

NSR scaling impacts due to kernel TCP/JSR infra will be identified and addressed as discussed above.

## Aggregated Ethernet/ SONET/ IRB Support

Not Applicable.

## Services/JSF (JUNOS Services Framework) Impact

Not Applicable.

## Multi-Chassis Support

Not Applicable.

## 64-Bit Support

Scaling tests will be conducted on 64 bit platforms also as discussed above.

## IPv6 Support

Not Applicable.

## Logical System Support

Not Applicable.

# SDK Impact

Not Applicable.

## SDK Customer Usage

Not Applicable.

# JUNOS Ready Software considerations

No potential value for JRS applications.

# Notes

None.

# Glossary

TCP – Transmission Control Protocol

NSR – Non Stop Routing

JSR – JUNOS socket replication

RPD – Routing protocol daemon

BGP – Border gateway protocol

HT – Hold timer

# Design Specification exception

Not applicable.

# Functional Specification Approver Checklist

|  |  |  |
| --- | --- | --- |
| **Org.** | **Role** | **Approver’s checklist**  (If you agree with the statements in the cell next to your role, please sign and date in the Approved by column) |
| PLM | Approver | The FS has detailed requirements for the feature and is in-line with the customer/product requirements. |
| SW Eng. | Approver | The FS has detailed functional and system scale & performance requirements and correctly captures the overall system impact |
| Sys Test | Approver | The FS has detailed functional and system scale & performance requirements for use to create the feature test plan |
| Information Experience | Approver | The FS has detailed functional and system scale & performance requirements for use as input to create the documentation plan |
| Architecture Reviewer | Approver  (optional) | The FS has detailed functional and system scale & performance requirements and is consistent with overall system architecture |
| Hardware | Approver  (Optional) | The FS has detailed functional and system scale & performance requirements and is consistent with the H/w design |
| Regression Test | Approver | The FS has detailed requirements that provide inputs to identify the regression test plan |
| Program Mgmt | Approver  (Optional) | The FS has detailed requirements and correctly captures the overall requirements for the feature |
| Customer Service | Approver  (Optional) | The FS captures the overall requirements for the feature and provide inputs required to create any Customer support plan document |
| Advanced Customer Engineering | Approver  (Optional) | The FS has detailed supportability requirements in Section 2.8 and captures supportability, diagnosability, debug requirements needs for the feature |
| Manufacturing | Approver  (Optional) | The FS correctly identifies any impact to Manufacturing process |