Application Note

CPE

Buffer Pool Manager

V4.10L.XX

CPE-AN1000-R

5300 California Avenue • Irvine, CA 92617 • Phone: 949-926-5000 • Fax: 949-926-5203

March 14, 2014

Revision History

Revision	*Date*	*Change Description*
CPE-AN1000-R	03/14/14	Initial release

Broadcom Corporation
5300 California Avenue
Irvine, CA 92617

Table of Contents

List of Tables

About This Document

Purpose and Audience

This document provides a high level description of the Buffer Pool Manager (BPM) feature, including CLI, parameters, and its usage.

This document is intended for software and system engineers.

Acronyms and Abbreviations

In most cases, acronyms and abbreviations are defined on first use.

Acronym	*Definition*
BPM	Buffer Pool Manager
CMF	Classify, Modify and Forward H/W
CPE	Customer Premises Equipment
IP	Internet Protocol version 4/6
QoS	Quality of Service
RXBD	RX buffer descriptor
TCP	Transmission Control Protocol
ToS	Type of Service
WAN	Wide Area Network

For a comprehensive list of acronyms and other terms used in Broadcom documents, go to: http://www.broadcom.com/press/glossary.php.

Document Conventions

The following conventions may be used in this document:

Convention	*Description*
Bold	User input and actions: for example, type **exit**, click **OK,** press **Alt+C**
`Monospace`	Code: `#include <iostream>` HTML: `<td rowspan = 3>` Command line commands and parameters: `wl [-l] <command>`
< >	Placeholders for *required* elements: enter your <username> or `wl <command>`
[ ]	Indicates *optional* command-line parameters: `wl [-l]` Indicates bit and byte ranges (inclusive): [0:3] or [7:0]

Technical Support

Broadcom provides customer access to a wide range of information, including technical documentation, schematic diagrams, product bill of materials, PCB layout information, and software updates through its customer support portal (https://support.broadcom.com). For a CSP account, contact your Sales or Engineering support representative.

In addition, Broadcom provides other product support through its Downloads and Support site (http://www.broadcom.com/support/).

Overview

The buffer pool manager (BPM) entity manages a globally shared pool of packet buffers. Upon request, the BPM gives a shared buffer to an RX interface. When a buffer is recycled and the interface RX ring is full, this excess buffer goes back to BPM. Prior to the addition of the BPM, the number of buffers at each RX interface was fixed and not able to scale for varying traffic scenarios. With BPM, there are additional buffers available and, based on demand, they are allocated from a centralized buffer pool to RX interfaces using configurable threshold levels.

Build

By default, the BPM feature is enabled in all profiles starting in release 4.10L.01.

Note: BPM requires a larger number of buffers and more memory than when the feature is disabled. This feature may need to be disabled along with Ingress QoS on small memory footprint designs (32 MB or smaller).

Enable BPM

To enable the BPM feature, use the *make menuconfig* command on Linux command prompt before build:

```
$ make menuconfig
```

then navigate to menuconfig→Buffer Pool Manager and Ingress QoS →<> Buffer Pool Manager (BPM). Use the space bar to select the BPM.

Note: The BPM feature can be built as a module (M) or be compiled out.

DISable BPM

To disable the BPM feature, use the *make menuconfig* command on Linux command prompt before build:

```
$ make menuconfig
```

then navigate to menuconfig→Buffer Pool Manager and Ingress QoS →<M> Buffer Pool Manager (BPM). Use the space bar to unselect the BPM feature.

Note: The BPM feature can be built as a module (M) or be compiled out.

Buffer Pool Manager

This document uses the term “ring” to indicate an RX interface DMA ring unless it is explicitly mentioned as something else. Similarly, BPM indicates a system/global buffer pool manager.

The global buffer pool manager is described below:

- It is a hybrid of completely partitioned and completely shared buffer schemes.
- All buffers are of the same size.
- Initially, at boot time, all buffers are owned by the BPM and are later made available for allocation by all RX interfaces in the system.
- Each driver maintains its own private ring(s) and initializes the ring(s) with buffers from the BPM. The number of buffers allocated is equal to the size of the ring.
- The remaining buffers after the initialization of ring(s) constitute the shared dynamic buffers. When the total number of shared dynamic buffers in the BPM is more than the low buffer level, all the TX queues start using a high threshold to decide whether to queue or drop the incoming packet.
- An interface, upon accepting a valid packet, passes the packet to the upper layer and requests that the BPM allocate a new buffer from the system global buffer pool to replenish its ring. There are two possible outcomes:
  - The BPM allocates a buffer.
  - The BPM is out of buffers, and the interface has to try again.
- After a packet has been transmitted (successfully or failed), the TX interface calls the recycle function of the RX interface. The recycle function checks whether the ring has any used RXBDs (pending buffer allocations). If it does, the recycled buffer is directly replenished to the ring. If it does not, the recycled buffer is freed to the BPM.

The above BPM features are achieved by buffer allocation and recycle mechanisms at the RX interface and TX queue thresholds.

Buffer Allocation and Freeing

The main objectives of the BPM buffer allocation and freeing are:

- The BPM makes the buffers available to any RX interface that needs more buffers.
- The BPM minimizes the number of buffer allocation/freeing requests to BPM to as few as possible. The number of buffer requests to the BPM is minimized in order to reduce contention on the shared resource, the centralized buffer manager. The BPM must use a lock when accessing the buffer rings, and the goal is to minimize the overhead of acquiring and releasing the lock, which consumes CPU cycles.

Buffer Allocation

- RX interface requests a buffer allocation to replenish its ring.
  - Each RX interface, on receiving a buffer allocation request, increments a local count to indicate the number of pending buffer allocation requests (used buffers). The RX interface requests the buffers from the BPM only when the used buffers from the ring go above a specified threshold.

  - The buffer allocation requests to the BPM are delayed and grouped together. By delaying the buffer allocation requests, there is an increased probability that the request will be successfully fulfilled using recycled buffers without having to go to the BPM. By grouping the requests, we achieve the objective of going to the BPM as few times as possible, while at the same time we make it more efficient by using the cache-aligned accesses.

Buffer Free/Recycle

- A buffer may be recycled back to the RX interface at various points in its path by calling the recycle function. Some of the possible scenarios where a packet may be dropped:
  - At the RX interfaces when the packet validation fails.
  - At the RX interfaces when there is congestion (set by Ingress QoS feature), and flow cache drops a packet belonging to a low-priority flow.
  - At the TX interface when the queue depth is greater than the current queue threshold for the TX queue.
- The recycle function first checks whether there are any pending buffer allocations (used buffers).
  - If yes, the buffer is directly recycled to the buffer ring and the number of pending buffer allocations (used buffers) is decremented. This is the most common scenario.
  - If no, the buffer is freed to the BPM.

Buffer Allocation and Free in FAP

Buffer allocation and free logic in FAP is very similar to the non-FAP case, except for the following differences:

- There are four instances of the local free buffer cache. The additional level of free buffer cache is required because interaction between FAP and the BPM is message-based, which is comparatively slow and could cause packets to drop if every request had to go to the BPM on the host MIPS.
- Buffer allocation change:
  - Before requesting buffers from the BPM after the allocation trigger threshold is crossed, FAP first checks whether there is a free buffer available in its own free buffer cache. If yes, the request is fulfilled from the free buffer cache. If no, the request is sent to the BPM.
- Buffer Free change:
  - The freed buffer does not get freed directly to the BPM, rather the buffer first goes to the FAP's free buffer cache.
  - Free buffer cache also has the free trigger threshold, and when the threshold is crossed, all the buffers of the free buffer cache are freed to the BPM.

TX Queue Thresholds

All TX queues on a slower interface (XTM, MoCA, and WLAN) are configured with high and low thresholds. The TX Q thresholds for each interface type can be independently configured.

The number of available dynamic buffers in the BPM is used to make the decision to use the TX queue low or high threshold for deciding to queue or drop a packet. If the number of dynamic buffers is less than the configured low buffer level threshold, the TX queue uses low threshold; otherwise high threshold is used.

The queue thresholds are checked to find out whether a packet should be queued or dropped. If the current queue depth is less than the current queue threshold, the packet is queued; otherwise it is dropped.

BPM Interaction with Packet Flow

The steps given below describe the BPM interaction with a packet:

1. Packet is received at an RX interface.
2. RX driver does basic packet validation checks.
   a. If the packet is not OK, drop the packet (see Buffer Free/Recycle).
   b. If the packet is OK, pass it to the next layer (flow cache, Linux stack, etc.).
3. Higher layer (flow cache) checks to see if the packet should be dropped (see Buffer Free/Recycle).
4. Higher layer processes the packet and invokes the *hard_xmit()* of the TX interface:
   a. Gets the current level of dynamic buffers available in the BPM.
   b. The dynamic buffer level is used to decide whether to use high or low threshold for determining to queue or drop the packet.
   c. If dynamic buffer level is high, then the high threshold becomes the current queue threshold; otherwise the low threshold of the TX queue becomes the current queue threshold.
   d. If the current queue depth is less than the current queue threshold of the TX queue, the packet is queued; otherwise the packet is dropped.
5. Buffer allocation is requested (see Buffer Allocation).
6. After one or more packets are transmitted or dropped, the buffer is recycled (see Buffer Free/Recycle).
7. The above steps are repeated for each packet.

CLI

To see the list of BPM CLI commands, type **bpm** at shell prompt.

```
# bpm
BPM Control Utility:

::: Usage:

:::::: BPM SW System:
       bpm status
       bpm enable
       bpm disable
       bpm thresh
```

BPM Status

```
# bpm status
bpm status
BPM status: enabled

------------------------ BPM Status -----------------------
 tot_buf     no_buf avail      alloc       free  head  tail
   19210          0 17594       1616          0  1616     0
-----------------------------------------------------------------------
     dev chnl      alloc used_b       free free_b rx_ring trig bulk
--- ---- ---- ---------- ------ ---------- ------ ------- ---- ----
    ENET    1          0      0          0      0     800  400  128
-----------------------------------------------------------------------
FAP FREE    -          -      -          0      0       0  256  256
FAP ENET    0          0      0          -      -     600  300  128
FAP XTM     0          0      0          -      -     200  100   64
FAP XTM     1          0      0          -      -      16    8   64
-----------------------------------------------------------------------
```

BPM status command prints information on the console, as detailed in Table 1.

Table 1: BPM Status—General

Field	*Description*
status	The current status of the BPM: enabled/disabled.
tot_buf	Total number of buffers allocated by the BPM at boot time.
no_buf	Number of times buffer allocation request failed because of non-availability of buffers.
avail	Number of buffers available.
alloc	Number of buffers allocated (including those were later freed) from the BPM.
free	Number of buffers freed to the BPM.
head	The head index for the buffer pool ring.
tail	The tail index for the buffer pool ring.

The fields described in Table 2 are per interface.

Table 2: BPM Status—per Interface

Field	*Description*
FAP	Indicates that the device or channel is used by FAP. If this field is blank, it means that it is a host device or channel.
dev	RX interface or device.
chnl	A channel on RX interface.
alloc	Number of buffers allocated from the BPM.
used_b	Number of RX BDs used (pending buffer allocation).
free	Number of buffers freed to the BPM.
free_b	Number of buffers available in local free cache. This field is only valid for FAP FREE.
rx_ring	RX ring size for the channel.
trig	The allocation/used buffer trigger threshold in number of buffers.
bulk	The number of buffers to allocate when allocation trigger threshold is crossed. For FAP FREE, it is the number of buffers to free to the BPM.

BPM Enable

```
# bpm enable
```

This command enables the BPM feature.

BPM Disable

```
# bpm disable
```

This command disables the BPM feature.

BPM Thresholds

```
# bpm thresh

-------------------- BPM Thresh --------------------
tot_buf tot_resv_buf max_dyn   avail dyn_buf_lo_thr
  19210         1616   17594   17594           8797

---------------------------------
port chnl rx_ring_buf alloc_trig
---- ---- ----------- ----------
ETH     0         600        300
ETH     1         800        400
XTM     0         200        100
XTM     1          16          8


-------------------------------------
     dev  txq loThr hiThr    dropped
--- ---- ---- ----- ----- ----------
FAP XTM     0    28    56          0
```

```
-----------------------------------
```

BPM thresh command prints information on the console, as detailed in Table 3.

Table 3: BPM Threshold—General

Field	Description
tot_buf	Total number of buffers allocated by the BPM at boot time.
tot_resv_buf	Total number of buffers reserved by all the RX interface channels.
max_dyn	The maximum number of dynamic buffers available when the packet traffic is stopped.
avail	The current number of dynamic buffers available. This can be less than max_dyn because of queuing at TX Qs.
dyn_buf_lo_thr	When the current number of dynamic buffers available falls below this value, all TX Qs will start using the TX Q low threshold; otherwise TX Q high threshold is used.

The fields described in Table 4 are per interface.

Table 4: BPM Threshold—per Interface

Field	Description
FAP	Indicates that the device or channel is used by FAP. If this field is blank, it means that it is a host device or channel.
dev	TX interface or device.
chnl	A channel or Queue on TX interface.
loThr	TX queue low threshold.
hiThr	TX queue high threshold.
dropped	Number of packets dropped for the TX Queue because the queue depth is more than the threshold.

Tunable Parameters

The BPM makes use of many tunable parameters, which are defined in *CommEngine/bcmdrivers/broadcom/include/bcm963xx/bpm.h* .

Caution! Any change of the BPM tunable parameters from their default values may adversely affect the system performance. It is highly recommended **not to change** these tunable parameters from their default values.

For each of the parameters given below, a *default* value is specified. Additionally, the *type* field indicates whether the parameter is applicable to a system or a specific interface type.

Buffer Memory Size

On system initialization, the BPM reserves a percentage of total SDRAM memory exclusively for packet buffers. The percentage used is configured through the *make menuconfig* command.

Default: 15% of total memory

Type: system

BPM Auto Disable

On system initialization, if the BPM finds that the total number of RXBDs reserved by RX interfaces is greater than 80% of the total number of buffers allocated, the BPM feature is automatically disabled due to availability of very few dynamic buffers.

Default: 80%

Type: system

RX Ring Size

The BPM depends on the size of the ring (number of RX BDs). It is recommended to keep the RX ring size of the high speed interface large, especially for Gigabit interfaces. The percentage of used BDs for larger rings will remain low, and hence fewer buffer allocation requests will come to the BPM. For some platforms, RX ring size depends on the configured buffer memory size, while on other platforms, the RX ring size is fixed.

Allocation Trigger Threshold

Allocation trigger threshold is specified in terms of percentage of used buffers of the ring. A buffer allocation request to the BPM is made when the number of used buffers crosses the allocation trigger threshold. It is recommended to keep this threshold high so that fewer buffer allocation requests come to the BPM.

Default: 15% of ring size

Type: interface

Bulk Allocation Count

Bulk allocation count is the number of buffers to be allocated from the BPM when the buffer allocation trigger threshold is crossed. It is recommended to keep this count in a middle range, not too big but also not too small. A bigger value will cause fewer buffer allocation requests to the BPM, but may cause a temporary spike in CPU usage.

Default: 128 for Ethernet, 64 for XTM

Type: interface

Dynamic Buffer Low Threshold

Dynamic buffer low threshold is specified in terms of percentage of dynamic buffers available (when the packet traffic is stopped). When the dynamic buffer low threshold is crossed towards the down side, the system buffers are in a congested state, and all the TX queues switch to using TX Q low thresholds for deciding whether to accept or drop a packet. Similarly, when the dynamic buffer low threshold is crossed towards the up side, the system buffers are in a normal state, and all the TX queues switch to using TX Q high thresholds for deciding whether to accept or drop a packet.

Default: 50% of dynamic buffer available in packet traffic stopped state

Type: system

TX Q Thresholds

Each slow TX interface (XTM, MoCA, WLAN) has two TX queue thresholds: low and high.

When a packet is forwarded to a TX interface, the interface uses one of the low or high thresholds. The threshold to use depends on the current system buffer congestion level triggered by dynamic buffer low threshold.

TX Q Low Threshold

The TX Q low threshold is used when the system buffers are in a congested state.

Default: calculated based on the interface requirements (see *bpm.h* file)

Type: interface

TX Q High Threshold

The TX Q high threshold is used when the system buffers are in a normal state.

Default: calculated based on the interface requirements (see *bpm.h* file)

Type: interface

Broadcom Corporation
5300 California Avenue
Irvine, CA 92617

Phone: 949-926-5000
Fax: 949-926-5203
E-mail: info@broadcom.com
Web: www.broadcom.com