Quality of service on KeyStone™ II architecture

Jason Reeder
Applications Engineer
Texas Instruments

Introduction

In networking, quality of service (QoS) is the act of prioritizing traffic and providing differing levels of service based on priority at a node in the network. Audio, voice and video streams are typically prioritized higher than other network traffic due to the real-time constraints of these streams. As the demand on the existing Ethernet infrastructure grows, QoS will become increasingly important.

This paper discusses the hardware blocks and coprocessors used to offload QoS operations from the host processors in the KeyStone™ II architecture from Texas Instruments. A full eight priority QoS networking solution can be implemented with little to no intervention from the host cores.

KeyStone II quality of service overview

Multiple blocks within the architecture contribute to the overall Quality of Service (QoS) system that can be implemented.

- KeyStone II devices contain a 5- or 9-port Gigabit Ethernet switch and may also contain a 3-port 10 Gigabit Ethernet switch depending on the device variant. Each switch has support for:
  - Multiple priority level QoS functionality at the MAC layer (802.1p)
  - Priority-based flow control (802.1Qbb) (available on K2E and K2L devices)
  - Priority-based rate limiting (802.1Qav) (available on K2E and K2L devices)
- Packet accelerator subsystem
  - Capable of packet classification and routing based on VLAN header PCP bits (802.1Q)
  - Also capable of packet routing based on IPv4/IPv6 header DSCP bits
  - This subsystem may also be used to drop packets if certain data flows are unwanted in the QoS system
- Queue manager subsystem (Multicore Navigator)
  - QoS firmware may be loaded into the queue manager subsystem that provides:
    - The ability to create a hierarchy of round-robin, weighted round-robin or strict priority scheduling blocks
    - Drop schedulers that can implement tail drop or random early detect (RED) drop in order to control latency within the QoS hierarchy

The 66AK2E05 SoC block diagram is shown in Figure 1 on the following page as an example of the KeyStone II architecture.

Quality of service in the switch subsystems

Transmit FIFO queue priority

KeyStone II devices have a 5- or 9-port GbE switch subsystem and may possibly have a 3-port 10 GbE switch subsystem. Each port in these switches, including the internal host facing port, has a

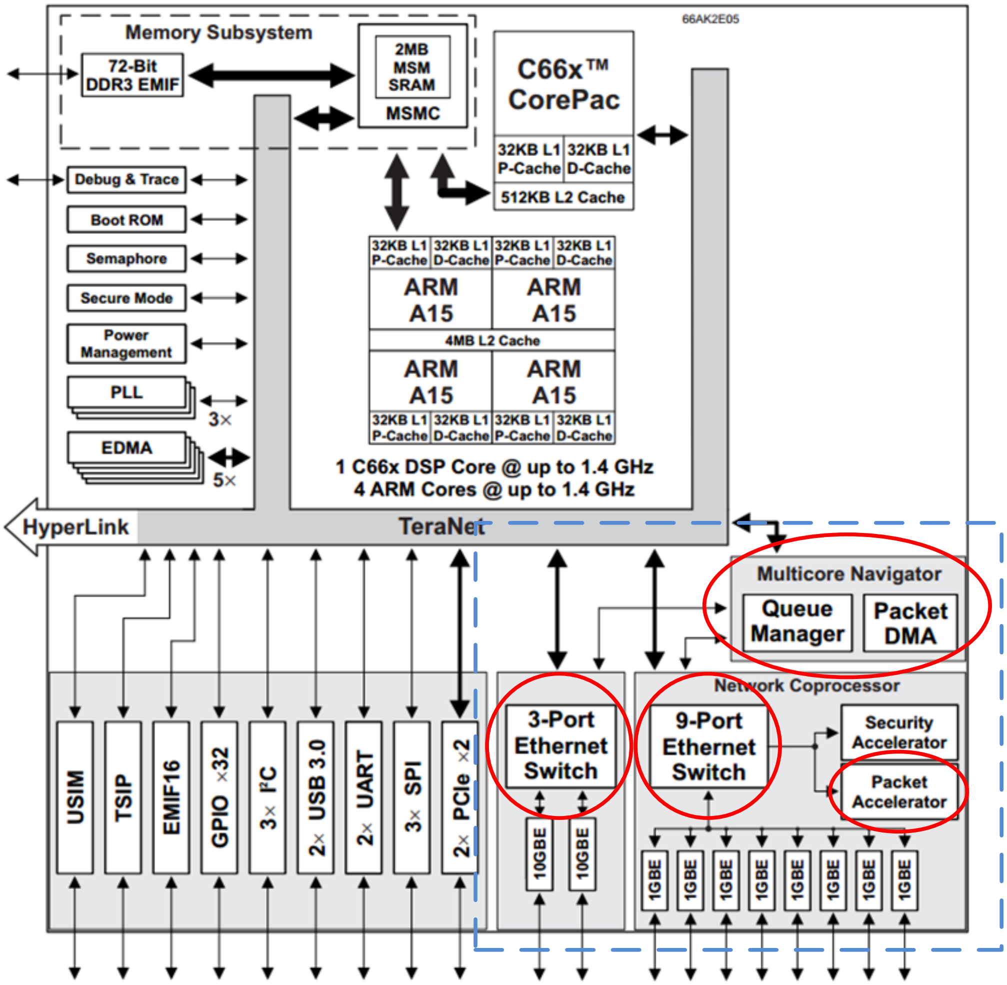

Figure 1: 66AK2E05 SoC block diagram

transmit buffer that is divided into multiple FIFOs, each assigned a different priority. Every packet that enters the switch is assigned a switch priority that determines which of the FIFOs that packet will use while awaiting switch egress. The switch priority (and in turn the TX FIFO priority) is user configurable through a set of mapping registers and is based on:

- The VLAN PCP bits of the VLAN header if the packet has one, or
- The DSCP bits if the packet is an IPv4 or IPv6 packet (available on K2E and K2L GbE Switch), or
- The priority assigned to the port that the packet entered the switch through

Priority flow control (802.1Qbb) (not available on K2H devices)

Flow control allows Layer 2 networking devices to provide feedback to each other without the need of upper layer protocols or management. Flow control was first introduced in IEEE 802.3x PAUSE control frames to allow overloaded Layer 2 receivers to inform a sender to stop sending packets for a period of time. Without flow control and PAUSE frames, an overloaded receiver would silently discard packets and the sender would not know that the data was getting dropped. Upper layer protocols had to be used to detect the dropped packets and request that the packets be resent.

A drawback of using IEEE 802.3x PAUSE control frames is that when a PAUSE frame arrives at a sender, it halts all traffic, regardless of priority, to the receiver. This drawback is remedied in the KeyStone II switch subsystems by using priority flow control (802.1Qbb). A priority flow control PAUSE frame includes the priority of the traffic flows that should be halted. The intent of priority flow control is to shutdown lower priorities as traffic congestion increases so that higher priorities can continue to operate. Once configured, priority flow control is handled entirely by the switch hardware and does not require any upper layer protocols or host processor cycles.

Packet accelerator quality of service

The packet accelerator is made up of several 32-bit coprocessors called packed data structure processors (PDSPs). These PDSPs each perform different packet classification and routing functions. Texas Instruments provides the firmware images for these coprocessors in order to allow them to complete their assigned functions.

At least two of those dedicated functions can be used in the overall QoS data path for packets coming into the device from the external Ethernet ports.

- The first PA QoS function is the ability to route packets to different queues based on the priority, or the differentiated services code point (DSCP), of the packet. The packet priority can be discerned from the VLAN PCP bits if the packet is VLAN or priority tagged, or the priority can be found in the DSCP bits of untagged IPv4 and IPv6 packets. This means that you can configure up to 72 (8 VLAN priorities + 64 DSCP priorities) queues and have the PA automatically route each incoming packet into the desired queue based on the priority, all without any host processor intervention.
- The other function of the PA that can be useful in QoS systems is the ability to drop unwanted traffic. Rules can be provided to the PA that match certain packets in the traffic flow based on one, or multiple, header parameters (e.g., MAC address, IP address, UDP port, etc.), and then drop those packets instead of forwarding them on to the host processor. This saves processing cycles on the host processor by not wasting time on unwanted packets.

Queue manager subsystem quality of service firmware

Within the queue manager subsystem (QMSS) of the KeyStone II architecture there are eight additional PDSPs that can be loaded with TI-provided firmware images. One firmware image that users can choose to load into the QMSS PDSPs is a QoS firmware that provides the ability to create a QoS tree made up of round-robin, weighted round-robin and strict priority schedulers as well as tail drop and random early detect (RED) drop schedulers.

Priority schedulers

The logical blocks that provide the priority scheduling functions are called lite ports. Each PDSP that is loaded with the QoS firmware provides 20 lite ports that are capable of performing round-robin, weighted round-robin or strict priority scheduling on up to four inputs queues in order to feed one output queue. Figure 2 on the following page is a block diagram representation of a lite port.

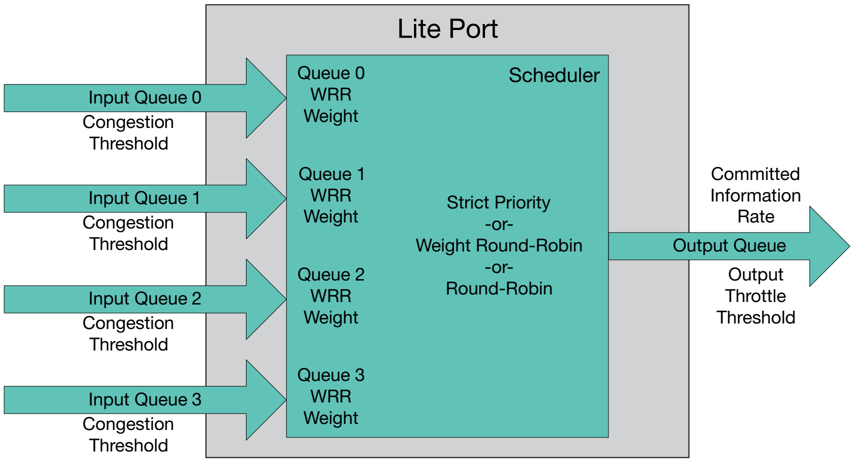

Figure 2: Lite port block diagram

Lite ports are the building blocks of the QoS trees that are made possible with this firmware. As shown in the block diagram the lite ports are extremely configurable:

- Different scheduling algorithms available for each lite port
  - Strict priority
  - Weighted round-robin
  - Round-robin
- Committed information rate for the port output allows you to limit the output data rate at each lite port
- Output throttling may be used to check for a configurable congestion threshold at the output queue in order to halt packet forwarding at that lite port, even if the committed information rate is not yet met
- Input congestion thresholds may be set at each input queue that drops packets once a selected threshold is reached
- Weights for the weighted round-robin algorithm are configured for each input queue (only applicable when the queue is selected for weighted round-robin scheduling)
- The output queue can be selected to go to any queue in the device. E.g.:
  - The input to another lite port to create a scheduling hierarchy
  - A general-purpose queue that the host processor can monitor
  - A transmit queue for the network coprocessor or the 10 gigabit Ethernet switch

Drop schedulers

There are 80 drop schedulers that are provided with the QMSS QoS firmware. They are intended to be placed at the input of a QoS hierarchy to enforce an upper bound on latency as well as ensure that memory resources are not exhausted. The drop schedulers have a single input queue and a single output queue (which is almost always an input queue to a lite port). Drop schedulers continuously monitor the depth at their output queue and compare it to a configurable threshold to determine whether to drop or forward packets that are being placed on the input queue. Figure 3 on the following page is a block diagram representation of a drop scheduler followed by its available configurations.

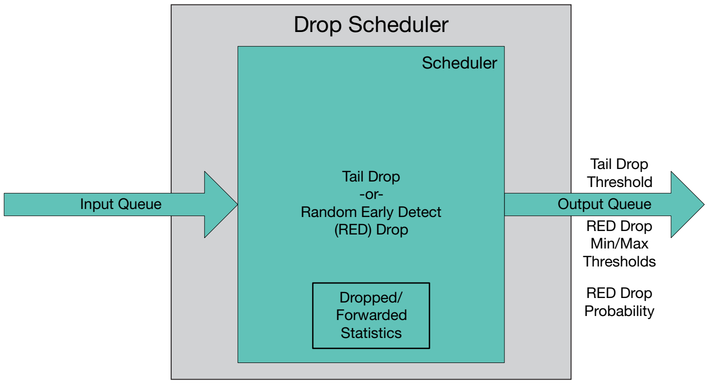

Figure 3: Drop scheduler block diagram

- Two different dropping strategies may be selected for a drop scheduler
  - Tail drop is used to drop all packets that arrive at the input queue when the output queue is above a user-configured threshold
  - Random Early Detect (RED) drop allows you to specify a minimum threshold and a maximum threshold for the output queue as well as a drop probability. If the output queue depth is:
    - below the minimum threshold, then no packets are dropped from the input queue
    - above the maximum threshold, then all packets are dropped from the input queue
    - between the minimum and maximum thresholds, then packets are dropped from the input queue based on a user configurable probability
- The output queue that is being monitored (and that packets are being forwarded to) is user configurable
- Statistics block can be configured to keep up with bytes dropped, bytes forwarded, packets dropped and packets forwarded at each output queue of the drop schedulers

Example quality of service/shaper tree

Using lite ports and drop schedulers as building blocks, there are a vast number of QoS hierarchies/trees that can be created. Figure 4 is one example.

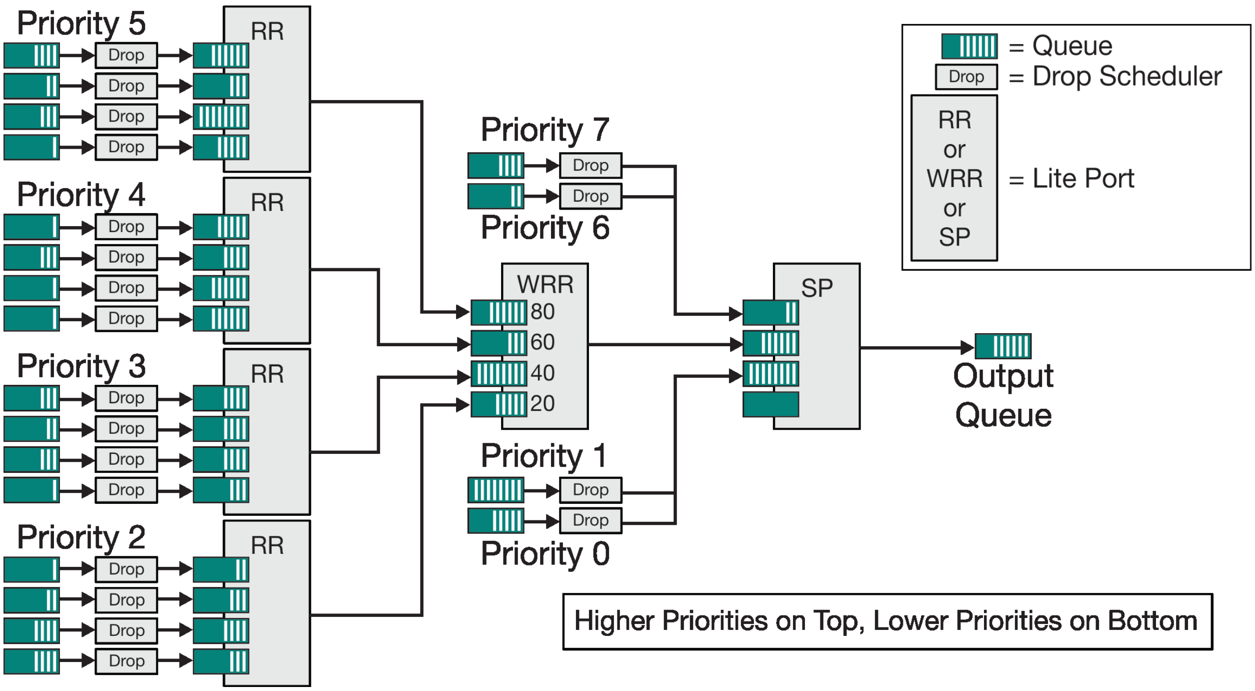

Figure 4: Example QoS hierarchy

Putting it all together

The data path of each packet entering the SoC can be configured to pass through a QoS hierarchy before ever reaching the host processor. This means that as the host processor becomes overloaded, higher priority packets will be able to reach the host faster while lower priority packets are delayed or dropped. It also allows for unwanted data flows or traffic patterns to be dropped completely in the hardware during ingress so the host processor does not waste any cycles.

Since the output queue of each lite port is completely configurable, the SoC egress data path can also benefit from these hardware blocks. For instance, a QoS hierarchy could be created using lite ports and drop schedulers that has its output queue pointed at the network coprocessor peripheral. If the network coprocessor begins to get overwhelmed with data being generated by the host processor, then the QoS hierarchy will ensure that higher priority traffic makes it out of the device and lower priority traffic gets dropped or delayed.

Up to four of the eight PDSPs in the queue manager subsystem can be loaded with the QoS firmware. Each PDSP loaded with the firmware provides 80 drop schedulers and 20 lite ports as building blocks. This allows for multiple QoS hierarchies, shaping traffic both into and out of the device, to exist simultaneously without using a single cycle of the host processor.

Software enablement

All of the functionality presented in this paper can be enabled with any host operating system using software provided by Texas Instruments. Using Linux™, for example, the QoS hierarchy mentioned above can be specified and configured through the Linux device tree. As the device boots, Linux will load the QoS firmware into the PDSPs and configure the QoS tree. For the GbE switch and packet accelerator functionality, a low-level-driver software library is provided that includes APIs which are accessible from Linux user space. These low-level drivers can also be used with a real-time operating system or in bare metal applications.

Conclusion

Functionality embedded into the switch subsystems, the packet accelerator and the QoS firmware for the queue manager subsystem allows for a robust and configurable QoS implementation in the KeyStone II architecture from Texas Instruments. Using these resources frees up valuable cycles on the host processors allowing for further differentiation in the overall end application.

IMPORTANT NOTICE

Texas Instruments Incorporated and its subsidiaries (TI) reserve the right to make corrections, enhancements, improvements and other changes to its semiconductor products and services per JESD46, latest issue, and to discontinue any product or service per JESD48, latest issue. Buyers should obtain the latest relevant information before placing orders and should verify that such information is current and complete. All semiconductor products (also referred to herein as “components”) are sold subject to TI’s terms and conditions of sale supplied at the time of order acknowledgment.

TI warrants performance of its components to the specifications applicable at the time of sale, in accordance with the warranty in TI’s terms and conditions of sale of semiconductor products. Testing and other quality control techniques are used to the extent TI deems necessary to support this warranty. Except where mandated by applicable law, testing of all parameters of each component is not necessarily performed.

TI assumes no liability for applications assistance or the design of Buyers’ products. Buyers are responsible for their products and applications using TI components. To minimize the risks associated with Buyers’ products and applications, Buyers should provide adequate design and operating safeguards.

TI does not warrant or represent that any license, either express or implied, is granted under any patent right, copyright, mask work right, or other intellectual property right relating to any combination, machine, or process in which TI components or services are used. Information published by TI regarding third-party products or services does not constitute a license to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property of the third party, or a license from TI under the patents or other intellectual property of TI.

Reproduction of significant portions of TI information in TI data books or data sheets is permissible only if reproduction is without alteration and is accompanied by all associated warranties, conditions, limitations, and notices. TI is not responsible or liable for such altered documentation. Information of third parties may be subject to additional restrictions.

Resale of TI components or services with statements different from or beyond the parameters stated by TI for that component or service voids all express and any implied warranties for the associated TI component or service and is an unfair and deceptive business practice. TI is not responsible or liable for any such statements.

Buyer acknowledges and agrees that it is solely responsible for compliance with all legal, regulatory and safety-related requirements concerning its products, and any use of TI components in its applications, notwithstanding any applications-related information or support that may be provided by TI. Buyer represents and agrees that it has all the necessary expertise to create and implement safeguards which anticipate dangerous consequences of failures, monitor failures and their consequences, lessen the likelihood of failures that might cause harm and take appropriate remedial actions. Buyer will fully indemnify TI and its representatives against any damages arising out of the use of any TI components in safety-critical applications.

In some cases, TI components may be promoted specifically to facilitate safety-related applications. With such components, TI’s goal is to help enable customers to design and create their own end-product solutions that meet applicable functional safety standards and requirements. Nonetheless, such components are subject to these terms.

No TI components are authorized for use in FDA Class III (or similar life-critical medical equipment) unless authorized officers of the parties have executed a special agreement specifically governing such use.

Only those TI components which TI has specifically designated as military grade or “enhanced plastic” are designed and intended for use in military/aerospace applications or environments. Buyer acknowledges and agrees that any military or aerospace use of TI components which have ***not*** been so designated is solely at the Buyer's risk, and that Buyer is solely responsible for compliance with all legal and regulatory requirements in connection with such use.

TI has specifically designated certain components as meeting ISO/TS16949 requirements, mainly for automotive use. In any case of use of non-designated products, TI will not be responsible for any failure to meet ISO/TS16949.

Products		**Applications**	
Audio	www.ti.com/audio	Automotive and Transportation	www.ti.com/automotive
Amplifiers	amplifier.ti.com	Communications and Telecom	www.ti.com/communications
Data Converters	dataconverter.ti.com	Computers and Peripherals	www.ti.com/computers
DLP® Products	www.dlp.com	Consumer Electronics	www.ti.com/consumer-apps
DSP	dsp.ti.com	Energy and Lighting	www.ti.com/energy
Clocks and Timers	www.ti.com/clocks	Industrial	www.ti.com/industrial
Interface	interface.ti.com	Medical	www.ti.com/medical
Logic	logic.ti.com	Security	www.ti.com/security
Power Mgmt	power.ti.com	Space, Avionics and Defense	www.ti.com/space-avionics-defense
Microcontrollers	microcontroller.ti.com	Video and Imaging	www.ti.com/video
RFID	www.ti-rfid.com		
OMAP Applications Processors	www.ti.com/omap	**TI E2E Community**	e2e.ti.com
Wireless Connectivity	www.ti.com/wirelessconnectivity		

Mailing Address: Texas Instruments, Post Office Box 655303, Dallas, Texas 75265