

POS-PHY Level 4 MegaCore Optimization for the Intel® IXP2800 Network Processor

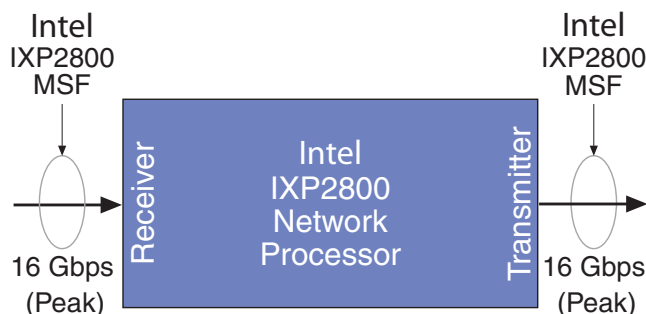
This white paper describes an enhanced configuration of the Altera® POS-PHY Level 4 MegaCore® function that is optimized to ensure efficient buffer usage for the Intel® IXP2800 network processor.

Intel® IXP2800 Network Processor Overview

The Intel® IXP2800 network processor enables rapid deployment of intelligent network services by providing maximum programming flexibility, code reuse, and high performance packet processing. The IXP2800 network processor supports a variety of wide area network (WAN) and local area network (LAN) applications, with speeds ranging from optical carrier (OC) line rates of OC-3 to OC-192 and Ethernet rates of up to 10 gigabits per second (Gbps).

The IXP2800 network processor media and switch fabric (MSF) interface is used to connect the IXP2800 network processor to a physical (PHY) layer device and/or to a switch fabric. The MSF consists of separate, and unidirectional receive and transmit interfaces, see Figure 1. Each receive and transmit interface can be independently configured as a System Packet Interface Level 4 Phase 2 (SPI-4.2) interface for a PHY device, or as an LVDS physical interface supporting the Common Switch Interface Layer 1 (CSIX-L1) protocol. This white paper focuses on the IXP2800 network processor MSF in SPI-4.2 mode. The MSF interface is designed to receive and transmit packets at a peak rate of 16 Gbps.

Figure 1. Intel IXP2800 Network Processor Media & Switch Fabric Interface



Altera POS-PHY Level 4 MegaCore Function Overview

The Altera POS-PHY Level 4 MegaCore function is a highly configurable core used to interface high-speed cell and packet transfers between PHY and link-layer devices. The POS-PHY Level 4 core supports SONET/SDH OC-192, and 10 Gbps Ethernet traffic.

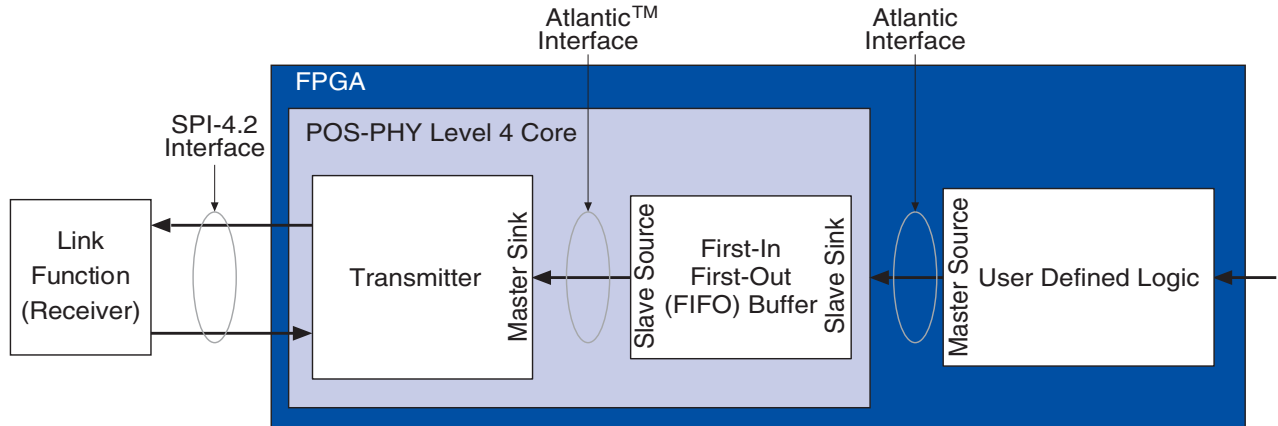
The POS-PHY Level 4 core uses an SPI-4.2 interface data width of 16 bits (True-LVDS™ solution), and can support up to 1 Gbps on each LVDS channel, with integrated dynamic phase alignment (DPA) in the Stratix™ GX device family.

Configurations of the POS-PHY Level 4 core can be obtained by selecting parameters, some examples include:

- Burst mode (BRSTMODE) which optimizes the transmitter core for use with network processors, such as the IXP2800 network processor
- Burst size (BRSTSIZE) which specifies the minimum number of payload bytes (i.e. 64, or 128 bytes) transmitted after a payload control word, and before a non-end of packet (EOP) control word is inserted

The POS-PHY Level 4 configuration, discussed in this white paper, is a single-PHY transmitter with support for a 64- or 128-byte contiguous burst mode. See Figure 2.

Figure 2. POS-PHY Level 4 Single-PHY Mode



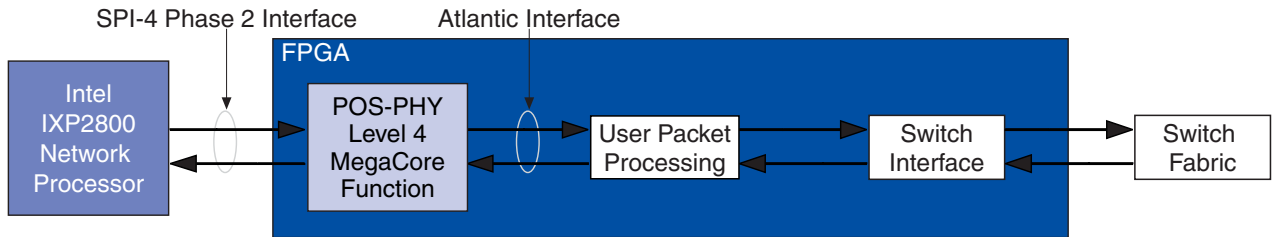
SPI-4.2 Interface

The System Packet Interface Level 4 Phase 2 (SPI-4.2) interface is an external interface protocol developed by the Optical Internetworking Forum (OIF). The SPI-4.2 interface features a high-speed portion and a FIFO buffer status portion. The high-speed portion comprises a 16-bit data bus, a 1-bit control line, and a double data rate (DDR) clock. The FIFO buffer status portion comprises a 2-bit status channel and a clock. This interface supports a data width of 16 bits, and can be a PHY-link, link-link, link-PHY, or PHY-PHY connection.

The SPI-4.2 interface supports up to 256 port addresses, with independent flow control for each. For data received by the PHY and passed to the link layer device, flow control is optional. The flow control mechanism is based upon independent pools of credits, corresponding to 16-byte blocks, for each port.

The POS-PHY Level 4 core uses the SPI-4.2 interface to pass data and control words in both the transmitter (source) and receiver (sink) directions. For the purposes of this white paper, the SPI-4.2 interface is a link-PHY connection, where the IXP2800 network processor is the link function and the POS-PHY Level 4 core is the PHY function. See Figure 3.

Figure 3. SPI-4.2 Interface Application



When the IXP2800 network processor MSF interface is configured in SPI-4.2 mode, each port has a data path and a FIFO buffer status path. The data path consists of 16-bit data signals, a clock, and a control signal; all of which use LVDS (differential) signaling, and are sampled on both edges of the clock (i.e. DDR). The data path clock can run up to 500 MHz. The FIFO buffer status path of each port consists of a clock signal and two data signals, that support either LVTTTL signaling clocked at up to 125 MHz or DDR LVDS signaling clocked at up to 500 MHz. The type of signaling can be configured by setting the appropriate registers in the software interface. When configured for DDR LVDS signaling, the dynamic training sequence of the SPI-4.2 interface—used to de-skew the signals—is implemented for both the data and the FIFO buffer status paths.

The SPI-4.2 interface performs two types of transfers, depending on the status of the receive control (RCTL) signal:

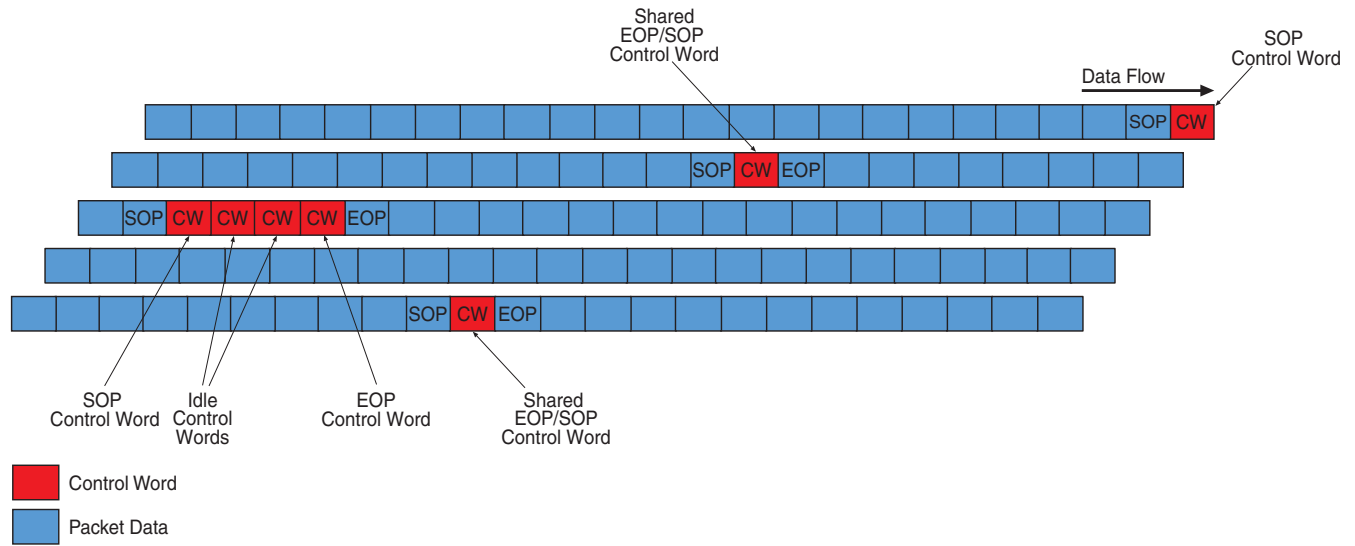
- When RCTL is asserted, control words are transferred
- When RCTL is deasserted, data bursts are transferred

Control words are categorized as follows:

- Payload control word
- Idle control word
- Training control word

Control words are transferred between data burst transfers, and data burst transfers immediately follow payload control words, as shown in Figure 4. For further information on SPI-4.2 operation, refer to the “Implementation Agreement: OIF-SPI4-02.0” document.

Figure 4. Embedded Control Words



Receive Buffer

The IXP2800 network processor MSF uses the receive buffer (RBUF), an 8 kilobyte (KB) RAM, to store received SPI-4.2 bursts. The RBUF is structured into elements. The element size can be configured, via register settings, to be either 64, 128, or 256 bytes. Therefore, RBUF can be configured to be either 128 64-byte elements, 64 128-byte elements, or 32 256-byte elements.

At chip reset, all elements are available. When an SPI-4.2 control word is received (i.e. when RCTL is asserted), the type field determines the action to be taken:

- If type is idle or training, the control word is discarded
- If type is not idle or training, an available RBUF element is allocated

Once a data burst occurs, two conditions (listed below) determine when an element is considered to be full, and another element is required. See Figure 9 for an example.

- When a control word is received
- When the element is filled

SPI-4.2 Data Burst Size

As stated in the previous section, the element size can be configured to be either 64, 128, or 256 bytes. Every time a payload control word is received, the IXP2800 network processor receive control logic allocates an available RBUF element to store the data of the following data burst. A transmitter can transmit a large data packet in multiples of short data bursts, such as 16 or 32 bytes, until the EOP is reached.

To ensure optimal use of the IXP2800 network processor RBUF, the SPI-4.2 data burst size should equal the RBUF element size, except for the EOP which can be smaller than the element size. The designer is responsible for implementing an appropriate transmission scheme.

When the data burst size is smaller than the element size, the IXP2800 network processor RBUF is used inefficiently. For example, when the RBUF element size is set to 64 bytes and a transmitter is designed to send multiples of short data bursts of 32 bytes with idle control words and payload control words between each data burst, every payload control word allocates one available 64-byte element for each 32-byte data burst. Thus, the RBUF is only utilized at 50%. This is also the case when the RBUF element size is set to 128 bytes and the data burst is less than 128 bytes.

Depending on the application, setting the SPI-4.2 data burst size to equal the IXP2800 network processor RBUF element size also has the following advantage: the first data burst—following the payload control word—to be marked by a start of packet (SOP) contains the data packet header information, or even multiple headers for encapsulated data packets. An RBUF element size of 64 or 128 bytes is large enough to store multiple headers, thus enabling more efficient packet header processing. For a short data burst, such as 16-bytes, a long header of over 16 bytes is split into two or more data bursts, thus complicating header processing.

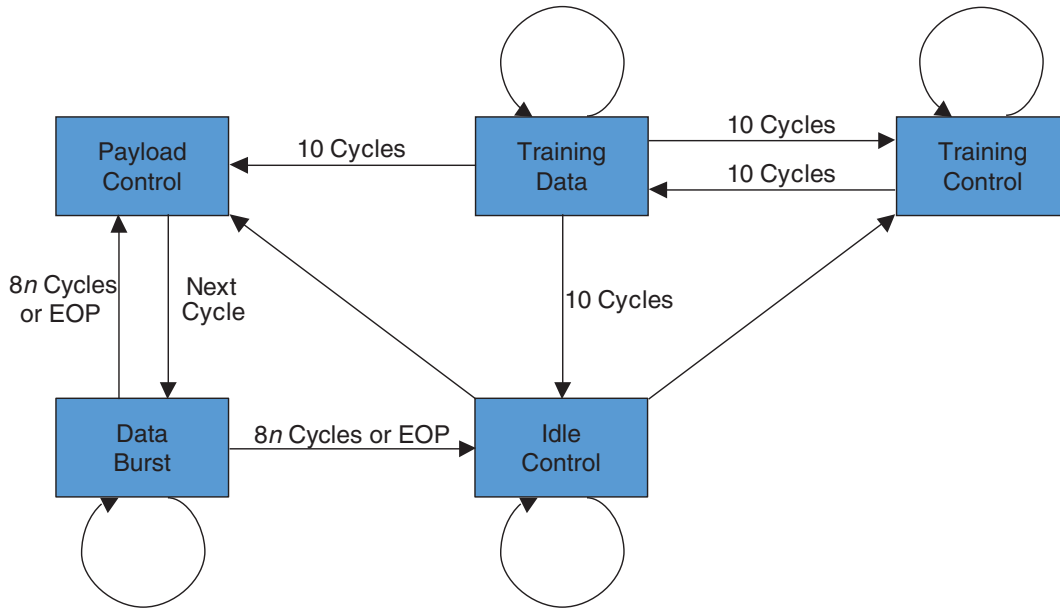
POS-PHY Level 4 MegaCore Optimizations

The Altera POS-PHY Level 4 MegaCore function has been optimized to efficiently use the IXP2800 network processor RBUF elements. The optimizations consist of two parts:

- The core uses a special burst mode scheduler to fill an entire IXP2800 network processor element
- The core uses a modified FIFO buffer to ensure complete element transfers

During data burst transfers across a POS-PHY Level 4 link, control words are inserted to carry information, such as training patterns, idles, continues, SOP markers, and EOP markers, to the connecting interface. Figure 5 shows the standard SPI-4.2 data path state machine—reproduced from the OIF-SPI4-2.0 specification—showing control words being inserted at specific intervals of $8n$ cycles, or following an EOP.

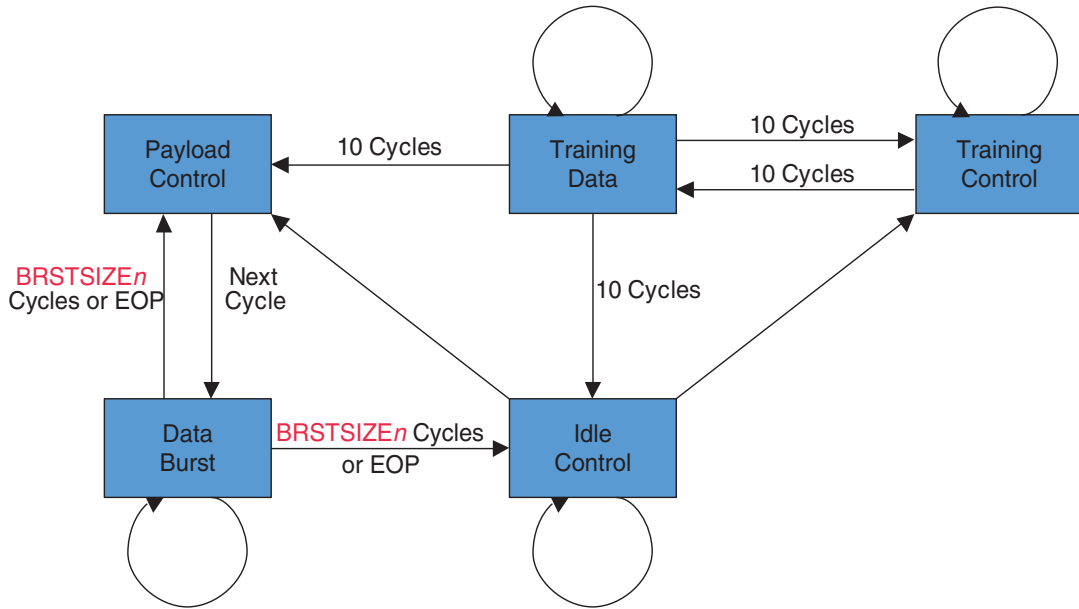
Figure 5. SPI-4.2 Data Path State Diagram



However, a number of other instances or events not linked to burst size transfers can trigger the insertion of control words, hence the need for the optimized POS-PHY Level 4 core.

The Altera POS-PHY Level 4 core enables the transmission of fixed bursts in lengths of 64 or 128 bytes. The first optimization, a burst mode scheduler ensures that burst size transfers occur without control word interruptions. When the BRSTMODE parameter is enabled, the POS-PHY Level 4 core uses an internal state machine (a modified version of the SPI-4.2 state machine) to synchronize on a SOP, and to delay all control word insertions until the end of the burst size transfer. Since the delay is equal to $BRSTSIZE \cdot n$, control words are only inserted once the burst transfer is complete, or after an EOP. Also, since the burst size can be set to match the IXP2800 network processor element size of 64 or 128 bytes, the initial SOP transfers and following transfers of packet data must lock to the burst size granularity, except on an EOP where the next packet data is placed in a new element. Figure 6 shows the modified state machine, optimized for network processor units (NPUs) such as the IXP2800 network processor.

Figure 6. NPU Data Path State Diagram Note (1)



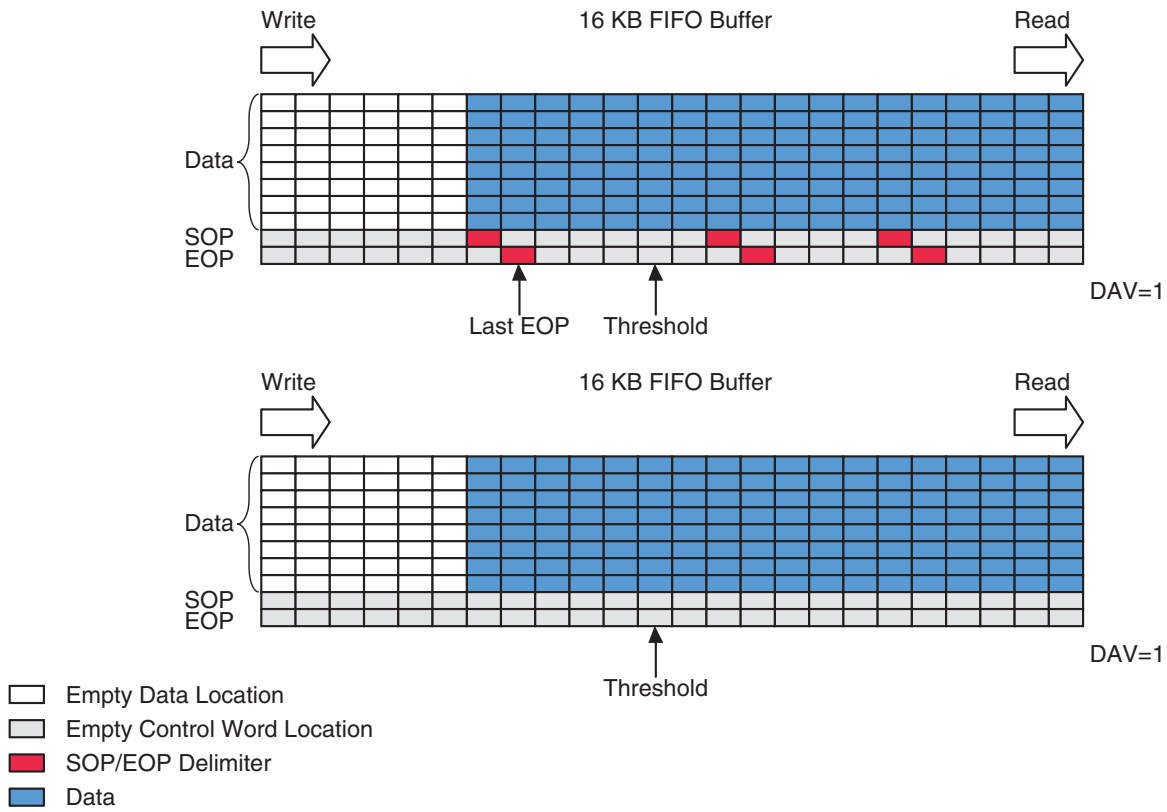
Note:

(1) BRSTSIZE is equal to the element size (64 or 128 bytes).

The second POS-PHY Level 4 core optimization is a modified FIFO buffer that ensures complete element transfers. To avoid inserting idle control words and to guarantee complete header transfers into a single element, the FIFO buffer takes in an entire packet before initiating a transfer. The *dav* control signal, implemented by the Atlantic interface, is asserted when the POS-PHY Level 4 FIFO buffer contains at least one complete packet. A complete packet is defined by a successfully received EOP. The *dav* signal remains asserted until the last remaining EOP in the buffer is read out. See Figure 7.

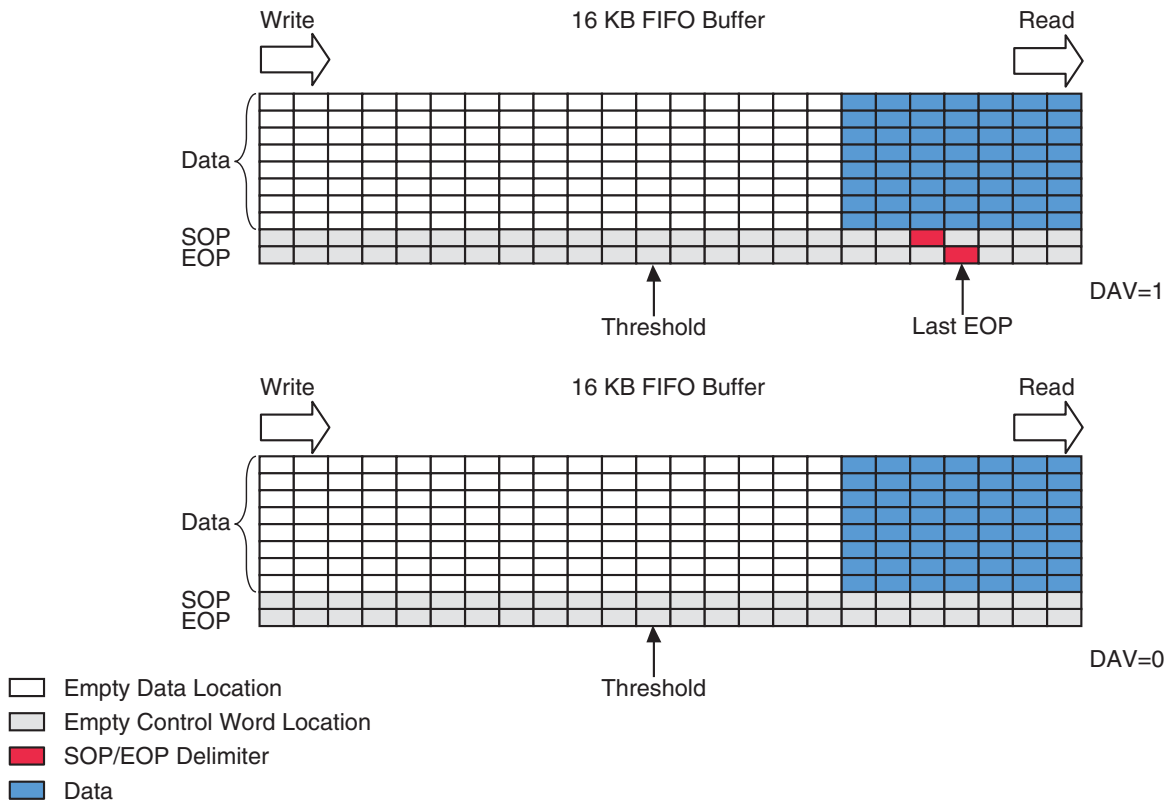
To accommodate large packets, the FIFO buffer supports an override mode with programmable high and low thresholds. The *dav* control signal is asserted when the pointer reaches a high threshold, and there are no EOPs contained in the FIFO buffer. See Figure 7.

Figure 7. FIFO Buffer Example—Data Exceeds Threshold Requirements



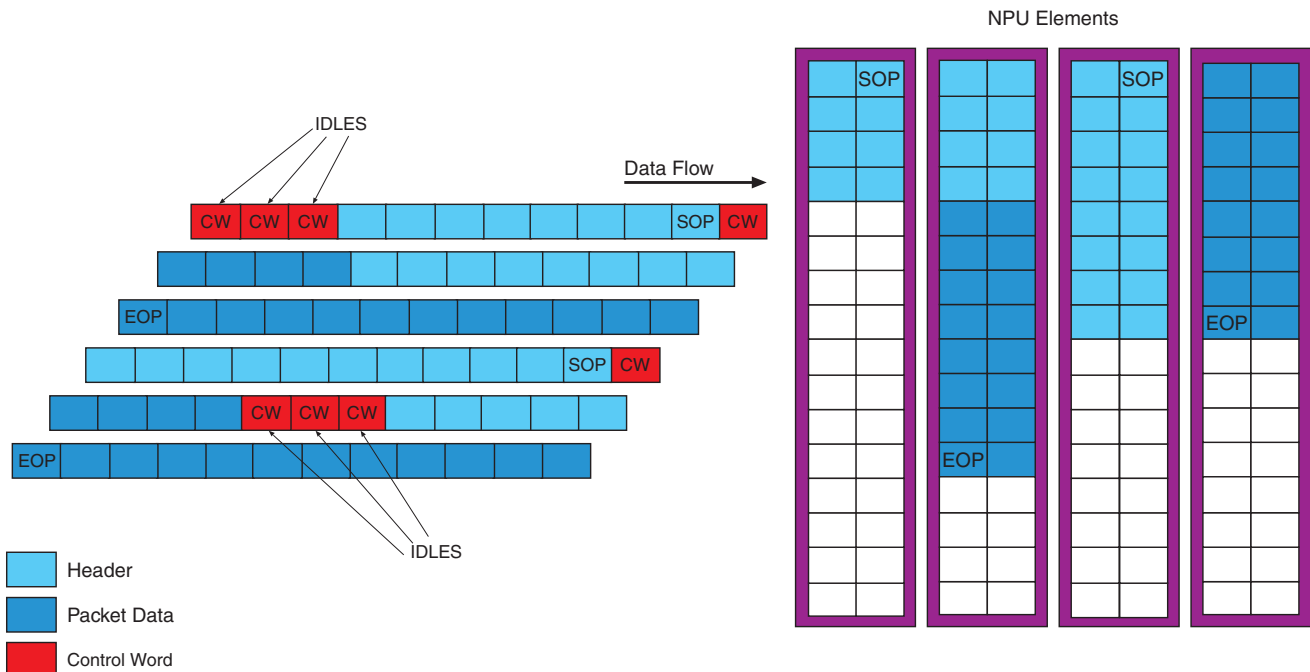
When the pointer falls below the low threshold, the *dav* control signal is no longer asserted, unless an EOP is present. See Figure 8. The low threshold should be set to at least the equivalent of the chosen *BRSTSIZE*, to ensure that idles are not inserted. If the low threshold is set to zero, the FIFO buffer should empty completely, which may require that the last transfer be padded with idle control words.

Figure 8. FIFO Buffer Example—Data Does Not Meet Threshold Requirements



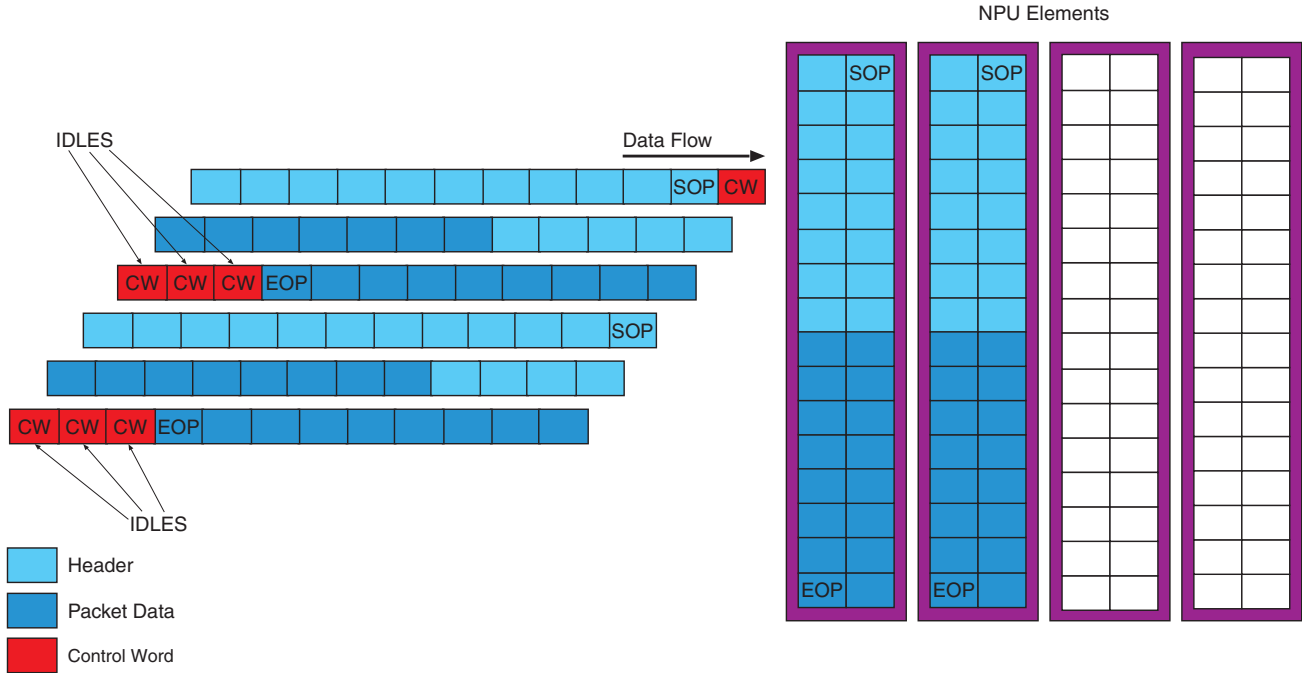
When a non-optimized data burst occurs, an element is perceived to be full once a control word or an EOP is received. This perception causes the IXP2800 network processor RBUF to switch to the next element. Figure 9 shows an example of non-optimized data burst transfers that lead to the inefficient use of four elements.

Figure 9. Non-Optimized Transfers



However, with the optimized POS-PHY Level 4 core, control words are inserted only after the BRSTSIZE_n (which is equal to the element size) or after an EOP, thus leading to the full use of the elements. Figure 10 shows an example of optimized data burst transfers and their efficient use of only two elements.

Figure 10. Optimized Transfers



References

For more information regarding the SPI-4.2 interface, or the Altera POS-PHY Level 4 MegaCore function, refer to the following documents:

- Optical Internetworking Forum (OIF), System Packet Interface Level 4 (SPI-4) Phase 2: OC-192 System Interface for Physical and Link Layer Devices, OIF-SPI4-02.0, January 2001
- Altera Corporation, POS-PHY Level 4 MegaCore Function User Guide

ALTERA[®]
 101 Innovation Drive
 San Jose, CA 95134
 (408) 544-7000
 www.altera.com

Copyright © 2003 Altera Corporation. All rights reserved. Altera, The Programmable Solutions Company, the stylized Altera logo, specific device designations, and all other words and logos that are identified as trademarks and/or service marks are, unless noted otherwise, the trademarks and service marks of Altera Corporation in the U.S. and other countries. Intel is a registered trademark of the Intel Corporation or its subsidiaries in the United States and other countries. All other product or service names are the property of their respective holders. Altera products are protected under numerous U.S. and foreign patents and pending applications, mask work rights, and copyrights. Altera warrants performance of its semiconductor products to current specifications in accordance with Altera's standard warranty, but reserves the right to make changes to any products and services at any time without notice. Altera assumes no responsibility or liability arising out of the application or use of any information, product, or service described herein except as expressly agreed to in writing by Altera Corporation. Altera customers are advised to obtain the latest version of device specifications before relying on any published information and before placing orders for products or services.