Streaming Interface (NPSI) Implementation Agreement

October 18, 2002

Revision 1.0

Chris Bergen
Streaming Interface Technical Editor

ZettaCom
2055 Laurelwood Road
Santa Clara, CA 95054, USA
Phone: +1 408-869-7002
Email: chris@zettacom.com

Jeffrey Lynch
Streaming Interface Task Group Chair & Asst Editor

IBM
P.O. Box 12195
Research Triangle Park, NC 27709, USA
Phone: +1 919-254-4454
Email: jjlynch@us.ibm.com

For additional information contact:

The Network Processing Forum, 39355 California Street,
Suite 307, Fremont, CA 94538
+1 510 608-5990 phone ✦ info@npforum.org

Table of Contents

Document revision history

NPF2001.121.23	August 13, 2002 Version 1.0 of the NPSI (Network Processing Forum Streaming Interface) implementation agreement.
NPF2001.121.24	September 4, 2002 Version 1.0 of the NPSI (Network Processing Forum Streaming Interface) implementation agreement. (Corrected typo.)
NPF2001.121.25	October 17, 2002 Version 1.0 of the NPSI (Network Processing Forum Streaming Interface) implementation agreement. (Comments in npf2002.540.00 after final ballot.)

1. Scope and Purpose

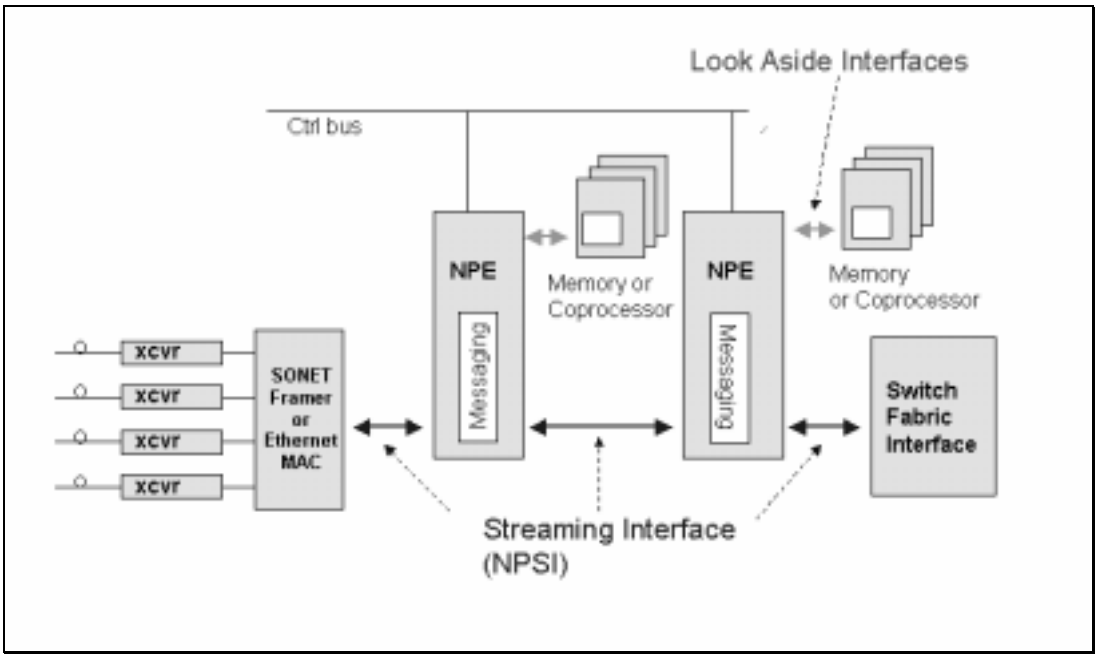
This specification defines the Network Processing[1] Forum (NPF) Streaming Interface (SI) implementation agreement (also, NPSI). The interface supports the transfer of network traffic between a pair of network processing devices, specifically including physical layer (PHY) devices (e.g., SONET framers/mappers or Ethernet MAC's), network processors, network coprocessors, and switch fabrics. It is targeted primarily, though not exclusively, at passing packets between a network processor and an adjacent physical layer, coprocessor, or switch fabric device at OC-192 line rates.

The streaming interface addresses processing in the data path and is complementary to the Look-Aside interface[2].

1.1 System Block Diagram

Figure 1 is a reference diagram used to illustrate the separate efforts within the NPF.

Figure 1. NPF System Block Diagram



1.2 Objectives and Non-objectives for this specification

The objectives for this specification are as follows:

- Provide a standard interface for connecting network processing devices.
  - Network processor to network processor
  - Network processor to coprocessor
  - Network processor to switch fabric
  - Coprocessor to switch fabric
  - Network processor to framer
  - Coprocessor to framer

[1] Note that the name is "Network Processing" Forum (and not "Network Processor" Forum).

[2] The Look-Aside Interface is a separate task group effort within the NPF.

- Provide a logical protocol that is scalable from 10 Gbps to 40 Gbps.
- Allow for the feasible implementation of a device that converts between NPSI and CSIX-L1 compliant devices.

The following are not objectives for this specification:

- The interface is not required nor specified to drive across a backplane.

2. Normative References

The following documents contain provisions, which through reference in this text constitute provisions of this specification. At the time of publication, the editions indicated were valid. All referenced documents are subject to revision, and parties implementing to agreements based on this specification are encouraged to investigate the possibility of applying the most recent editions of the standards indicated below.

[1] System Packet Interface Level 4 (SPI-4) Phase 2: OC-192 System Interface for Physical and Link Layer Devices, Optical Internetworking Forum Implementation Agreement, January 2001[3].

[2] OIF-SPI5-01.0, System Packet Interface Level 5 (SPI-5): OC-768 System Interface for Physical and Link Layer Devices, Optical Internetworking Forum Implementation Agreement, November 2001[4].

[3] CSIX-L1: Common Switch Interface Specification-L1, Network Processing Forum Implementation Agreement, August 5, 2000

[4] ANSI/TIA/EIA-644-A-2001, “Electrical Characteristics of Low Voltage Differential Signaling (LVDS) Interface Circuits”, Published February 1, 2001.

Portions of the OI Forum SPI-4 Phase 2 [1] and SPI-5 [2] specifications were used to create this document.

[3] At the time of publication of this specification, the SPI-4 Phase 2 specification may be found on the Internet at the following location: http://www.oiforum.com/public/documents/OIF-SPI4-02.0.pdf

[4] At the time of publication of this specification, the SPI-5 specification may be found on the Internet at the following location: http://www.oiforum.com/public/documents/OIF-SPI5-01.0.pdf

3. Conventions in this Specification

This specification follows the following protocol of terminology:

SHALL indicates that the item is a requirement for conformance to the NPSI specification.

MAY indicates that the item is optional.

SHOULD indicates that the item is not required by this specification, but is offered as implementation guidance.

Footnotes used in this specification are informative only.

A reserved field or bit shall be transmitted as zero and ignored on reception.

If there is a conflict between a state machine diagram and the text of this document (including a table entry), the state machine takes precedence over the text.

Figures (except state diagrams) are informative only.

3.1 Definitions

The following terms are used throughout this specification:

Class: A field (up to 8 bits) that is used to discriminate and manage traffic flows. The mechanism for differentiation of services among multiple classes is implementation-specific and beyond the scope of this specification.

Egress: From the switch fabric.

Egress Port: An addressable endpoint at an egress NPE-Fabric interface.

Fabric: A switch fabric.

Flow: A sequence of packets conveyed via the NPF Streaming Interface.

Flow Identifier: A set of values designating a flow at specific reference points in the NPF Streaming Interface architecture.

Ingress: Toward the switch fabric.

Ingress Port: An addressable endpoint at an ingress NPE-Fabric interface.

Logical Port: A port that may share a physical instantiation of the NPSI with other logical ports.

Multicast: The fabric service that delivers a packet to one or more egress ports.

Multicast Bitmap:
A bitmap enumerating a set of egress port identifiers.

Multicast Flow: A flow that uses the multicast service of a fabric.

Multicast Identifier:
A label designating a set of egress ports of a fabric.

Network Processing Element (NPE):
Any device that uses the NPF Streaming Interface as a data path interface to communicate with either a switch fabric, a framer, or another NPE.

Packet: A service data unit that is conveyed via the NPSI.

Physical Port: A port that is a unique physical instantiation of the NPSI.

Pool: An aggregation of ports for flow control.

Port: An addressable endpoint of the NPSI protocol.

Port Identifier: An address designating a single ingress or egress port of a fabric or a channel of the NPE-NPE interface.

Segment: The payload of a protocol data unit that is transferred across a port.

Sub-port: A logical port within a given physical port.

Sub-port Identifier: An address designating a sub-port (without the physical port information).

Switch Fabric: A facility that accepts packets at multiple ingress ports and delivers them to their designated egress ports.

Unicast: The fabric service that delivers a packet to exactly one egress port.

Unicast Flow: A flow that uses the unicast service of a fabric.

Word: A 16-bit value.

3.2 Acronyms and Abbreviations

The following acronyms and abbreviations are used in this specification

DIP	Diagonal Interleaved Parity.
EOP	End Of Packet.
FIFO	First-In First-Out (queue).
HSTL	High Speed Transceiver Logic.
IEEE	Institute of Electrical and Electronics Engineers.
I/O	Input / Output (signal).
LSB	Least Significant Bit.
LVDS	Low-Voltage Differential Signaling.
LVTTL	Low-Voltage Transistor to Transistor Logic.
MSB	Most Significant Bit.
NPE	Network Processing Element.
NPF	Network Processing Forum.
NPSI	Network Processing Forum Streaming Interface.
PHY	Physical Layer Device.
SONET	Synchronous Optical Network.
SOP	Start Of Packet.

4. Introduction

This document specifies the Network Processing Forum's data path interface, the NPF “Streaming Interface” (NPSI), for the interconnection of network processing devices. Included in the definition of network processing devices are framers, switch fabrics, network processors, and network coprocessors (e.g., classification or encryption coprocessors).

The interface supports the transfer of data for nominally 10 Gbps (OC192) aggregate bandwidth applications between two adjacent devices. It is a point-to-point interface with support for addressing and flow control for multiple framer channels, switch fabric and/or coprocessor destinations and classes, as well as multicast traffic.

The interface defines the link-level requirements, including data framing and packet delineation, flow control, address formats, and error detection.

The interface is applicable to multiple applications, and as such is a combination of multiple interface types. Because of this, the interface requirements are explained in multiple sections as three different NPSI modes, one addressing the requirements for a switch fabric interface, one addressing the requirements for a framer interface, and one addressing the requirements between network processors and coprocessors. A device may choose to implement one or more of these NPSI modes and, as such, should clearly state which NPSI mode or modes it supports.

5. Architectural Overview

The NPSI provides for the transfer of data traffic and control information between two adjacent network processing devices. There are three NPSI modes that are described individually due to specific requirements of the application. These three modes are described in the following NPF Streaming Interface (NPSI) Reference Model section.

5.1 NPSI Reference Model

The NPSI supports 3 modes of operation:

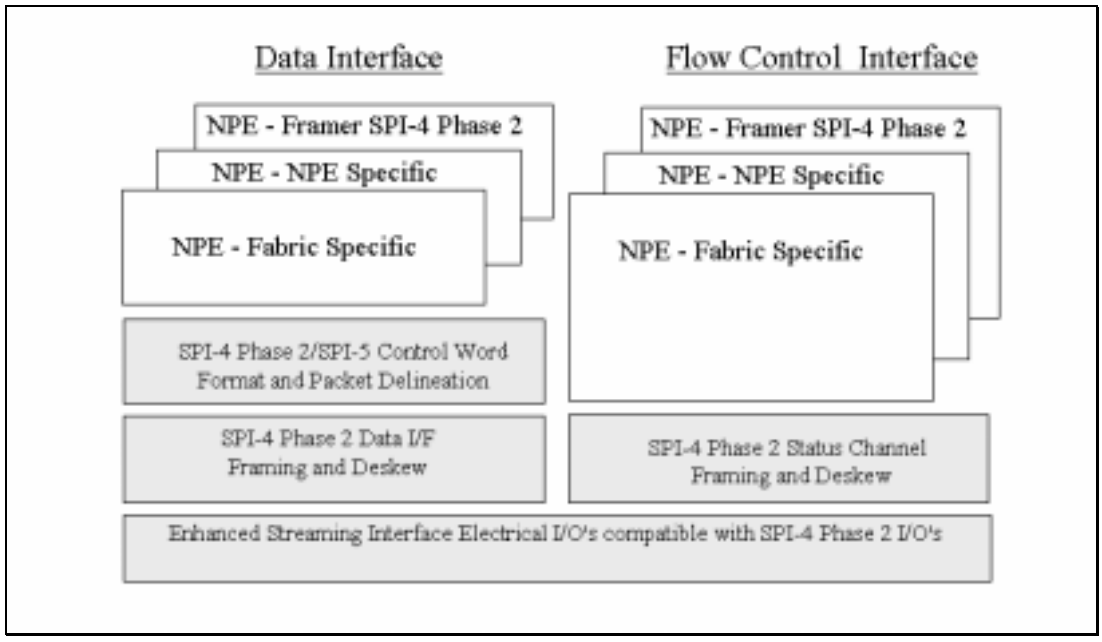
- Between an NPE and a Framer (NPE-Framer)
- Between an NPE and a Switch Fabric (NPE-Fabric)
- Between an NPE and an adjacent NPE (NPE-NPE)

Each mode of operation is independent from the other modes of operation. As such, a device that claims conformance to the NPSI should indicate whether the specified interface is conformant to the NPE-Framer mode, NPE-NPE mode, or the NPE-Fabric mode (or more than one of these modes).

The interface is based on SPI-4 Phase 2 and SPI-5, with protocol concepts from CSIX-L1. The interface has a data path and an out-of-band (reverse channel) flow control path, thereby supporting simplex operation. (Two NPSI's are required, one in each direction, for a full-duplex interface. In that case, however, the two NPSI's are independent interfaces, although the two may be implemented on a single device.)

The following reference model identifies that the interface shares many SPI-4 Phase 2 functions, but each mode of operation has operational differences tailored to its specific environment.

Figure 2. Streaming Interface Reference Model



5.2 General Features of the NPSI

5.2.1 Common Functions

The NPSI provides a data path with in-band control and framing information, and an out-of-band status path for flow control information, allowing for the unidirectional transfer ("streaming") of data between two devices.

The NPSI has the following general characteristics:

- Point-to-point connectivity between two adjacent devices
- Data Path:
  - 16-bit wide data path
  - Source-synchronous, double-edge clocking at 311 MHz (622 Mbps[5] per bit lane[6]) minimum
  - LVDS electrically compatible with SPI-4 Phase 2, with support for rates up to 1.3 Gbps and above per bit lane
  - In-band data multiplexing and flow control context, packet delineation, and error control coding
- Status (Flow Control) Path:
  - 2-bit wide status path
  - Source-synchronous, double-edge clocking
  - In-band framing and error control coding
- Support for packet and cell-based protocols
- Support for board-level connections of 8 inches FR-4 plus one connector.

5.2.2 NPE-Framer Mode

The NPE-Framer mode of operation of the NPSI is the OI Forum SPI-4 Phase 2 specification [1]. The NPSI makes no changes to the SPI-4 Phase 2 interface.

5.2.3 NPE-Fabric Mode

- Support for up to 4096 egress ports (destinations) with up to 256 classes
- Support for multicast and unicast traffic
- Data Path:
  - Support for multiplexing on port and sub-port basis
- Status (Flow Control) Path:
  - LVDS electrically compatible with SPI-4 Phase 2, with support for rates up to 1.3 Gbps and above per bit lane
  - Signals operate at the full rate of the data path signals
  - Link-level flow control (Data Ready indication)
  - Directed status messages for flow control
  - Support for flow control on port and sub-port basis
  - Optional flow control of directed status messages
  - Required 2-bit status bus optionally extended to 4 bits

[5] Data rates are specified in bps, not Hz. Because the clock is double-edged, stating the data rate is less ambiguous than stating the clock rate.

[6] A bit lane is a signal carrying one data bit (i.e., one LVDS differential pair).

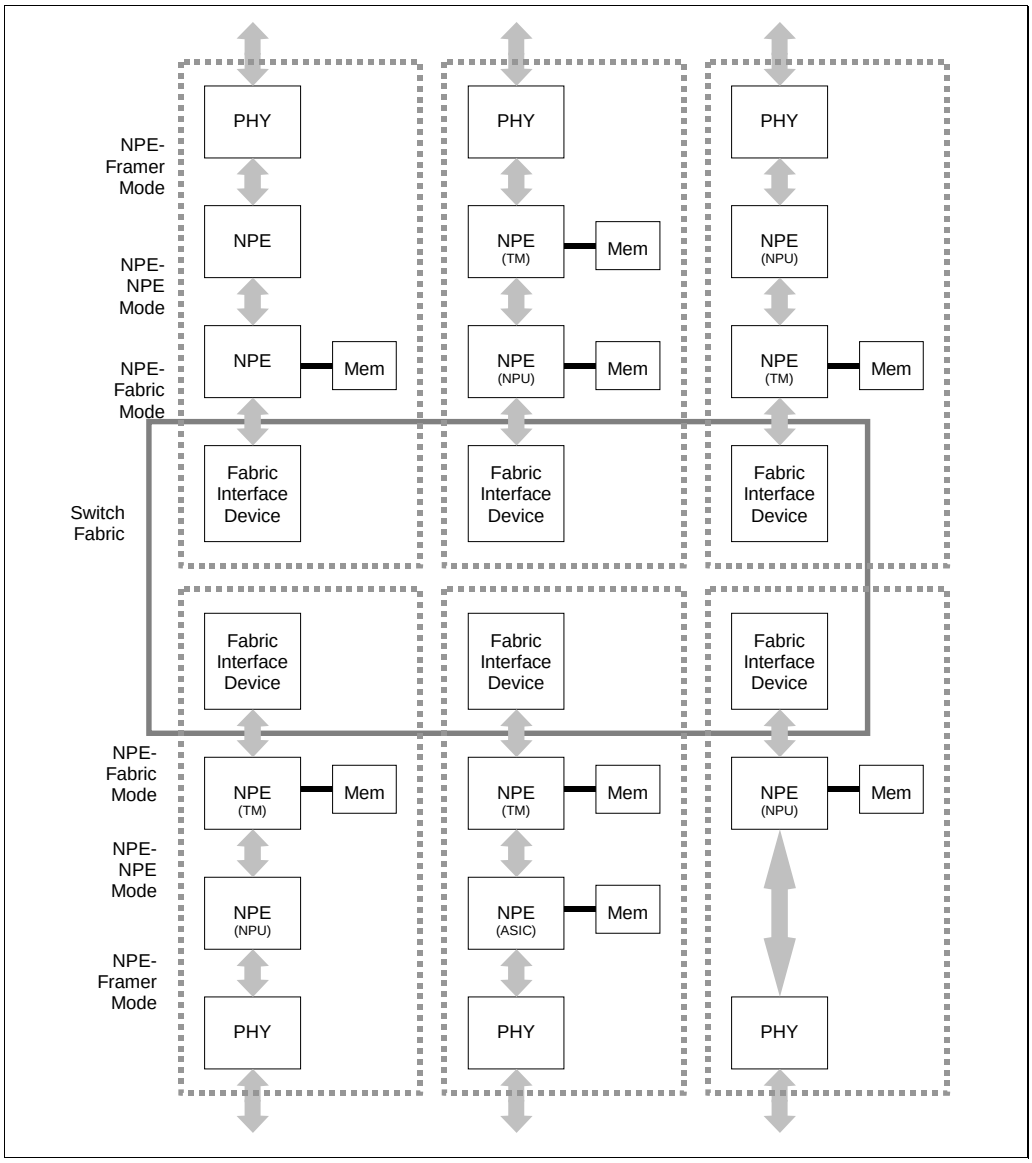
5.2.4 NPE-NPE Mode

- Support for up to 256 ports (channels) with the ability to scale to a much larger number of ports
- Status (Flow Control) Path:
  - LVDS electrically compatible with SPI-4 Phase 2, with support for rates up to 1.3 Gbps and above per bit lane
  - Signals operate at the full rate of the data path signals
  - Support for the SPI-5 Pool Status Mechanism [2], which allows the flow control granularity to be different from the full data multiplexing capability

5.3 Implementation Examples using NPSI

Figure 3, following, shows the NPSI at various points in an implementation.

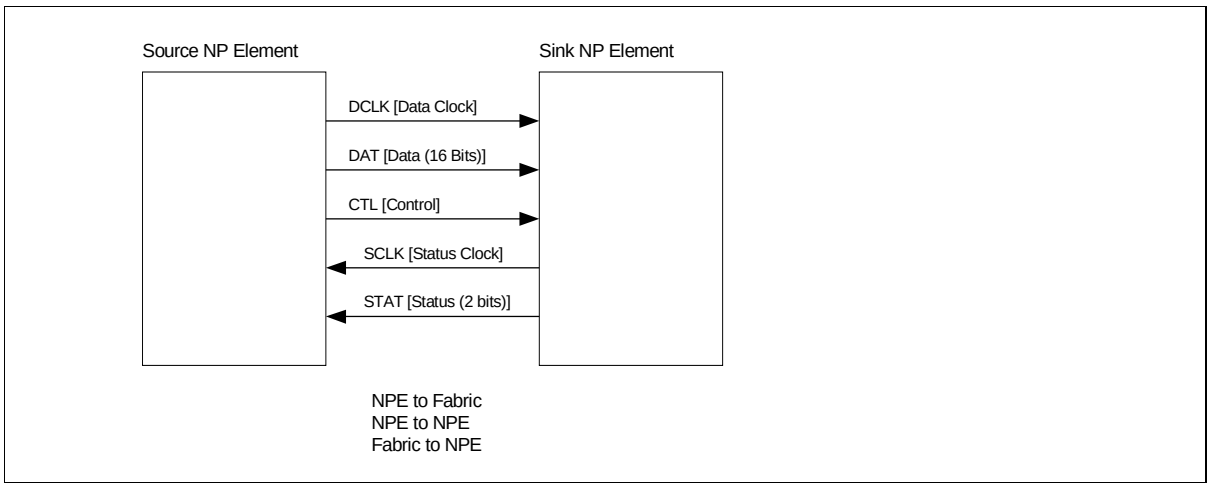
Figure 3. Example Uses of the Multiple Modes of the NPSI



5.4 Interface Signals

Figure 4 is a block diagram of the interface signals of an NPSI.

Figure 4. Diagram of NPSI Signals



The data path includes clock, data, and control/framing (DCLK, DAT[15:0], and CTL).

The flow control status channel includes clock and flow control status (SCLK and STAT[1:0][7]).

Note that the interface consists of a unidirectional data path, with an out-of-band flow control interface corresponding to (and in the opposite direction of) the data path. This allows for a device implementation that processes traffic in a single direction without the need for a companion device (in the opposite direction) for managing flow control information.

Table 1. NPSI Signal Summary

SIGNAL	Signal Name	DESCRIPTION
DCLK	Data Clock	Clock associated with data and control (DAT and CTL).
DAT[15:0]	Data	Data payload path with in-band control words for packet multiplexing and delineation. These signals are driven off the rising and falling edges of DCLK.
CTL	Control	CTL is high when a control word is present on DAT[15:0]. It is low otherwise. This signal is driven off the rising and falling edges of DCLK.
SCLK	Status Clock	Clock associated with status (STAT).
STAT[1:0]	Status	Status (flow control) information, along with associated framing and error control coding. These signals are driven off the rising and falling edges of SCLK.

[7] This signal is optionally 4 bits in the NPE-Fabric Mode.

6. NPE-Framer Mode

The NPSI framer interface is the OI Forum SPI-4 Phase 2 interface specification [1]. The NPSI makes no changes to that specification.

7. Common Functions for the NPE-NPE and NPE-Fabric Mode

The NPE-NPE and NPE-Fabric modes of the NPSI (data path and status path) shall use the LVDS I/O electrical specifications in Section 10, “Physical Layer” for all interface signals.

7.1 Common Data Path Operation

The NPSI data path provides delineation of data on a packet basis and the ability to multiplex data from multiple packets on the data bus. Packets are delivered over the NPSI data bus in bursts of data that have a provisionable, constant length, with the exception of bursts that terminate with an EOP (End Of Packet); EOP bursts may be shorter. A data burst consists of one or more words of payload data, with a control sequence (as described below) immediately preceding it and following it. The control sequence differs for the various modes of operation. All data transfers include a payload section.

The data contained in the payload of a single data burst is called a segment. A packet is transferred as a sequence of one or more segments. Control words are inserted between segments. All segments of a packet, except the last segment (EOP), must be MAX_SEGMENT_SIZE (bytes)[8]. The last segment (EOP) may be shorter if the total packet length is not a multiple of MAX_SEGMENT_SIZE (bytes). Once a transfer has begun, the control sequence and data words of a segment are sent uninterrupted until the end of the packet or the MAX_SEGMENT_SIZE has been reached.

An NPSI device shall support one or more values of MAX_SEGMENT_SIZE that are integer multiples of 16 bytes. An NPSI device may also support one or more values of MAX_SEGMENT_SIZE of any size. If an NPSI device supports more than one value of MAX_SEGMENT_SIZE, it shall provide a mechanism to configure the value.

The transfer of data requires the use of control words and data words. Control words and data words are differentiated by the CTL signal (as described in Section 5.4, “Interface Signals”).

There are four types of control words:

- Address Control Words (ACW),
- Payload Control Words (PCW),
- Idle Control Words (ICW), and
- Training Control Words (TCW).

Note that Training Control Words may only occur during a training sequence, and are never included in the control sequence of a data transfer.

There are three types of data words:

- Address Data Words (ADW),
- Payload Data Words (PDW), and
- Training Data Words (TDW).

Address Data Words may be included in a control sequence. Training Data Words may only occur during a training sequence.

[8] Note that the framer interface, SPI-4 Phase 2, requires bursts (other than the EOP burst) to be multiples of 16 bytes and allows bursts of up to the maximum configured payload data transfer size, which is required to be a multiple of 16 bytes. In contrast to the SPI-4 Phase 2 specification, which permits non-EOP data bursts to be less than the maximum configured payload data transfer size, the NPSI requires all bursts (other than the EOP burst) to be segments of MAX_SEGMENT_SIZE in length (and MAX_SEGMENT_SIZE is an integer multiple of 16 bytes).

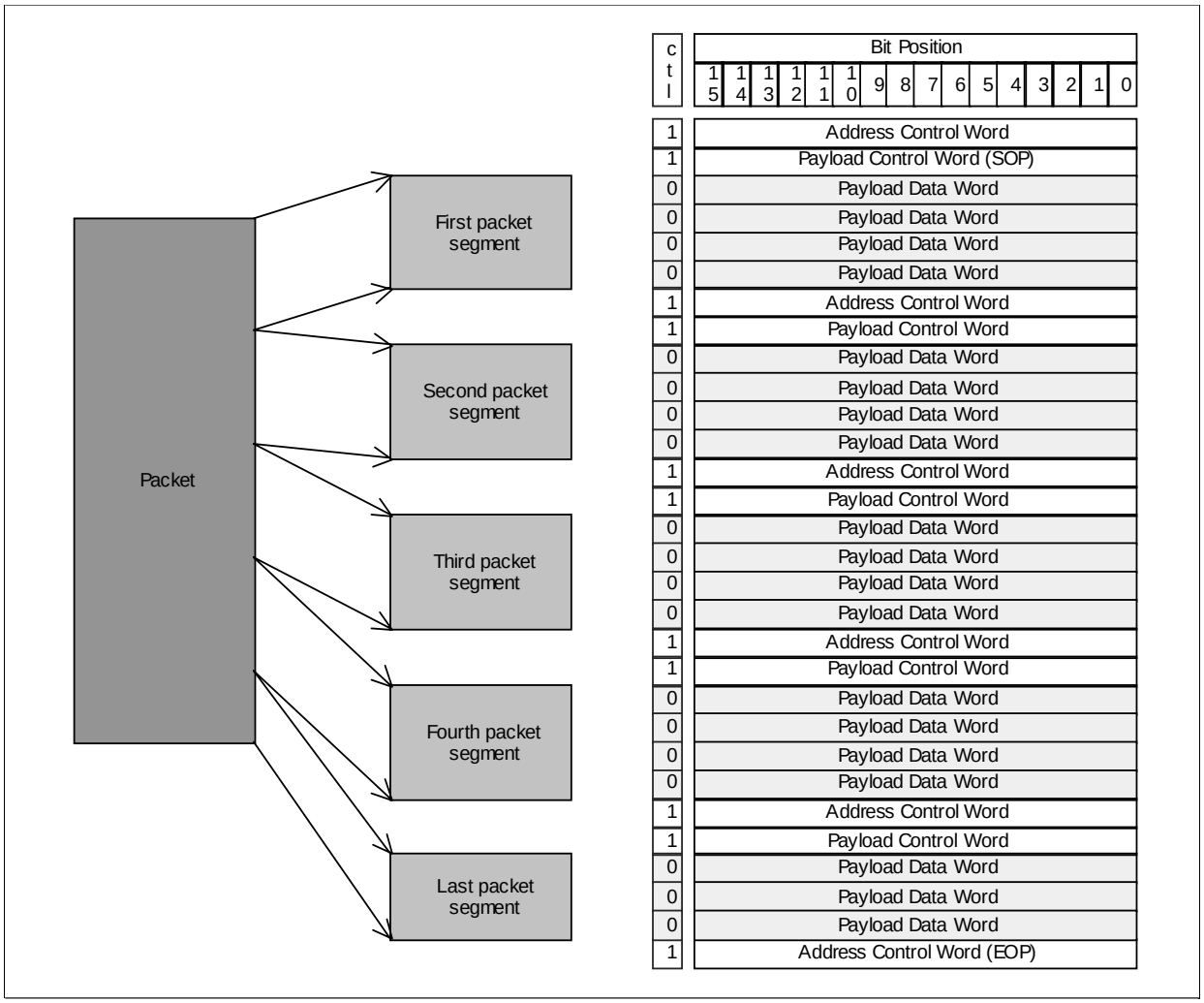
Data is always transferred with a control sequence occurring before and after each segment. A control sequence consists of zero or one ACW, followed by zero or more ADW's, followed by one PCW. An ACW or a PCW can indicate the start of a control sequence (and therefore the start of a transfer), depending on the interface mode. (Typically, an ACW starts the control sequence in NPE-Fabric mode and a PCW starts the control sequence in NPE-NPE mode.) Control words contain the SOP and EOP delineation of the packet.

A field in the control word immediately preceding the segment indicates if the segment is the SOP. The start of a packet (SOP) is always aligned to the beginning of a segment. Data from disparate packets do not share a segment.

A field in the control word immediately following the segment indicates if the segment was the end of the packet (EOP). The interval between the last payload word of a given transfer and the start of the next control sequence (marking the start of another transfer) consists of zero or more idle control words and zero or more training patterns (used for data path de-skew). The transfer of a segment may not be interrupted for any reason (e.g., insertion of idle words or training patterns).

The following figure (Figure 5) is an example of how a single packet is broken into segments and sent across the NPSI.

Figure 5. Example NPSI Data Path Segment Multiplexing



A control word carries the Start-of-Packet or End-of-Packet status. The last packet segment may need to be padded (by one byte) to the 2-byte (16-bit) width of the interface.

When inserted in the data path, a control word is aligned such that the MSB of the control word is sent on the MSB (bit 15) of the data lines. A control word or sequence of control words (control sequence) that separates two adjacent transfers may contain status information pertaining to the previous transfer, the following transfer, or both transfers.

During idle periods, when no data is being delivered, Idle Control words are sent.

Control words are described further in the following section.

7.1.1 Data Framing Formats

This section describes the general operation of the NPSI data framing. In certain modes of operation, some framing formats or sequences may not apply.

The control sequence must contain a Payload Control Word (PCW). The control sequence may require an Address Control Word (ACW) and one or more Address Data Words (ADW). The required and optional formats are specific to the NPE-NPE Mode or NPE-Fabric Mode of operation, and are explained in Section 8, “NPE-Fabric Mode,” and Section 9, “NPE-NPE Mode.”

There are three possible data burst formats. Each has a defined control sequence:

- A Basic Data Burst control sequence contains a single Payload Control Word.
- An Extended Address Data Burst without ADW control sequence contains a single Address Control Word followed by a single Payload Control Word.
- An Extended Address Data Burst with ADW control sequence contains a single Address Control Word followed by one or more (up to eight) Address Data Words, followed by a single Payload Control Word.

A segment contains one or more Payload Data Words (PDW) and it must be preceded by one of the three valid control sequences, depending on the mode of operation.

The end of the transfer is indicated by a Control Word, which may be an Idle Control Word (ICW), a Payload Control Word (PCW), or an Address Control Word (ACW), and contains the end of packet (EOP) status. For back-to-back transfers (i.e., when no idle periods or training sequences need to be inserted between data transfers), a single Control Word (an ACW or PCW) terminates the previous transfer and initiates the next one. During idle periods, Idle Control Words are sent. During training, Training Control Words and Training Data Words are sent, with an Idle Control Word immediately preceding the training sequence.

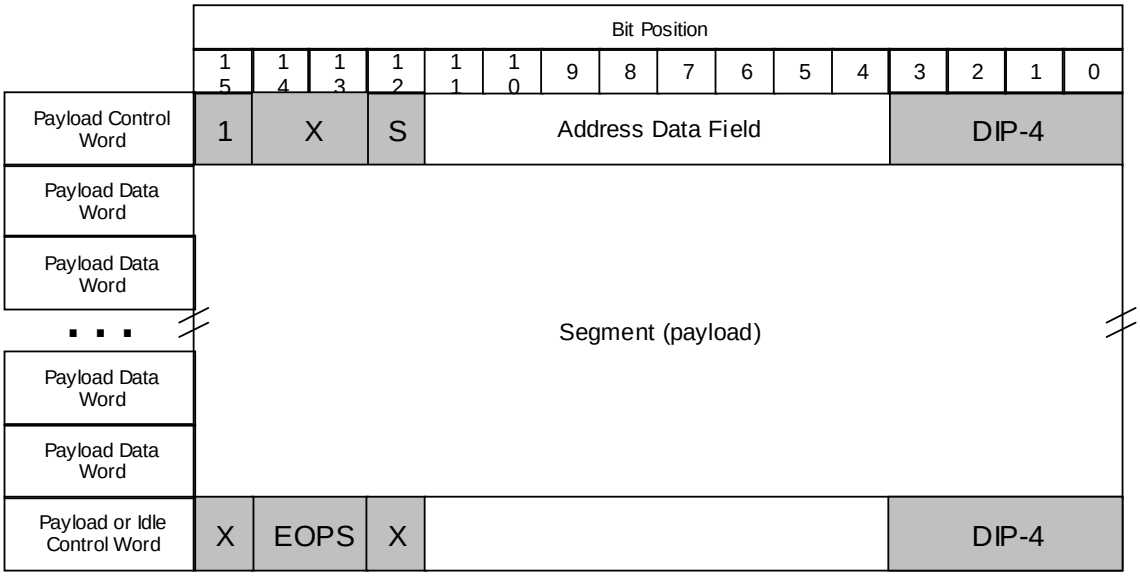
Control Words contain the packet delineation information: Start Of Packet (SOP), and End Of Packet Status (EOPS). Control Words also may contain a 4-bit Diagonal Interleaved Parity (DIP-4) as the error control code.

The control sequence for the Basic data burst consists of one control word: PCW. The PCW contains the SOP, EOPS, and DIP-4[9]. The ACW and ADW are not used in this format. The Basic data burst is only allowed (and is the required format) in the NPE-NPE mode.

Control words and data words are differentiated by the CTL signal (as described in Section 5.4, “Interface Signals”).

[9] This is consistent with SPI-4 Phase 2 [1].

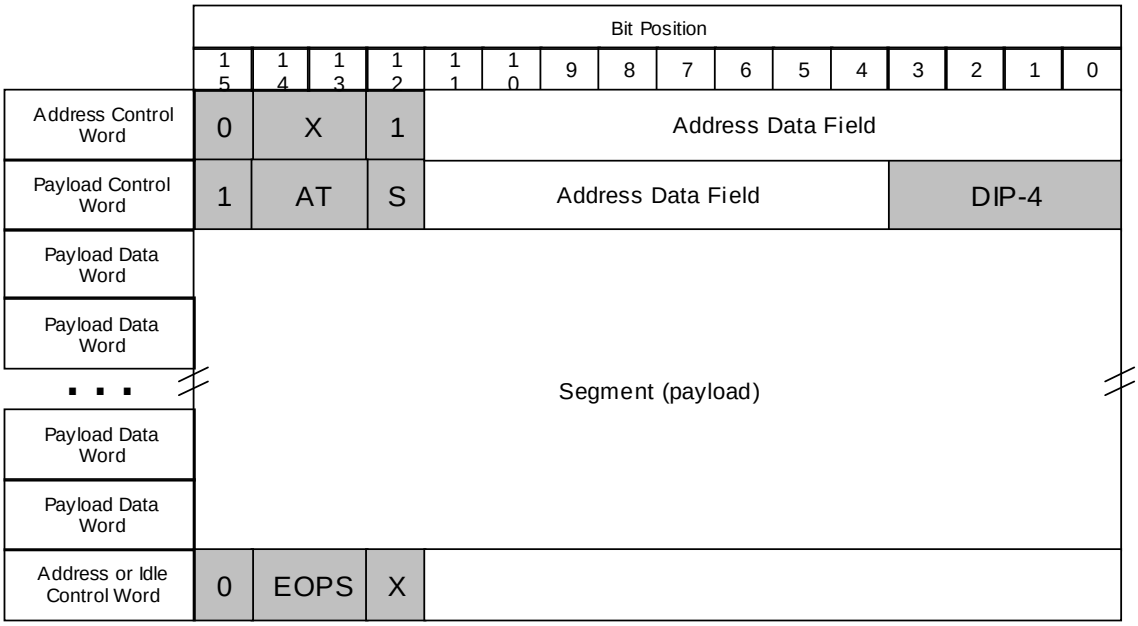
Figure 6. Basic Data Burst



There are two forms of the Extended Address data burst: with or without the ADW[10].

The control sequence for the Extended Address data burst without ADW consists of two control words: ACW and PCW.

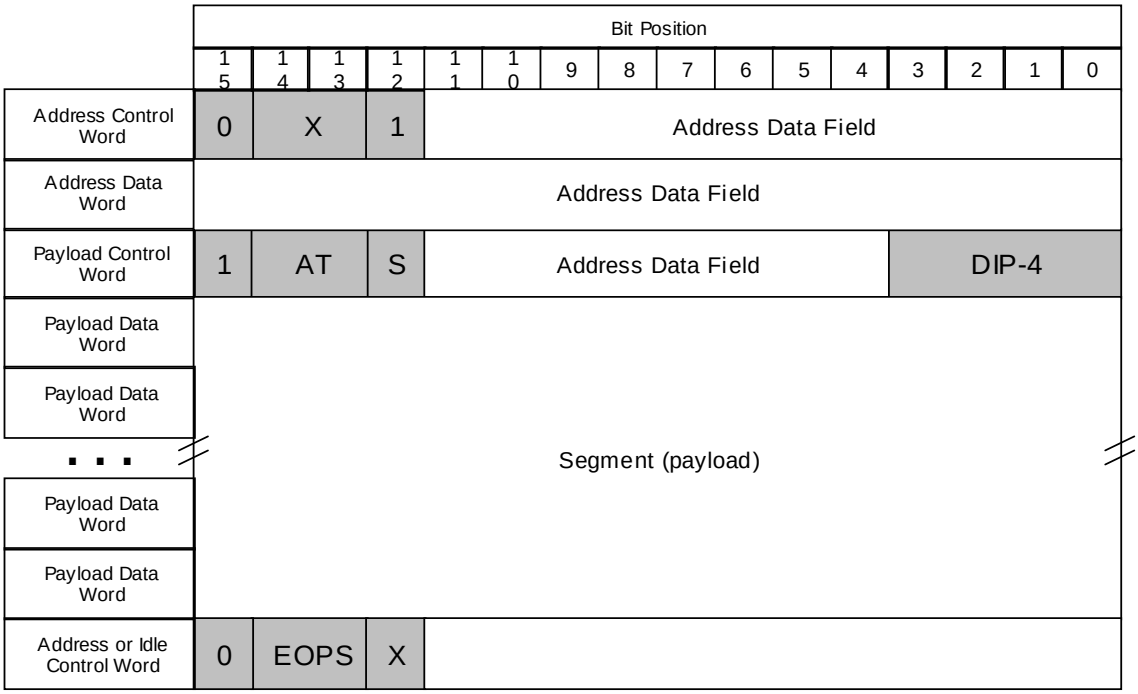
Figure 7. Extended Address Data Burst without ADW



[10] This is consistent with a mode of SPI-5 [2].

The control sequence for the Extended Address data burst with ADW consists of three (or more) control words: ACW and PCW, with up to eight ADWs in between.

Figure 8. Extended Address Data Burst with ADW



7.1.1.1 Control Word Field Definitions

Table 2 describes the fields used for ACW, PCW and ICW control words.

Table 2. Control Word Field Definitions

CW Bits	Mnemonic	Description
CW[15,12]	CWT	Control Word Type The Control Word Type CWT[1:0] identifies the type of the Control Word. CWT[1:0] = 0 0 : Idle Control Word (ICW) CWT[1:0] = 0 1 : Address Control Word (ACW) CWT[1:0] = 1 0 : Payload Control Word (PCW) and not Start-of-Packet (SOP) CWT[1:0] = 1 1 : Payload Control Word (PCW) and Start-of-Packet (SOP) CWT[1] is mapped to CW[15] while CWT[0] is mapped to CW[12]
CW[14:13]	EOPS/AT	End-of-Packet Status, or Address Type (in PCW, NPE-Fabric Mode only) The End-of-Packet Status Field EOPS[1:0] is present in Idle Control Words or Address Control Words (or in Payload Control Words in the Basic Data Burst format only) and reports the status of the payload transfer immediately preceding it. EOPS[1:0] is only valid in the first Control Word (ACW, ICW, or PCW) immediately following a PDW burst. EOPS[1:0] = 0 0 : Not EOP EOPS[1:0] = 0 1 : EOP with abnormal cell/packet termination (Abort) EOPS[1:0] = 1 0 : EOP with normal cell/packet termination, 2 bytes valid EOPS[1:0] = 1 1 : EOP with normal cell/packet termination, 1 byte valid EOPS[1:0] is mapped to CW[14:13]. The AT values are defined in Table 6, “Address Type Encoding for the NPE-Fabric Mode.”
CW[11:4] (in ICW or PCW) or CW[11:0] (in ACW)	ADF	Address Data Field The Address Data Field contains addressing information related to the data burst in PCW and ACW. ADF is all zeroes for ICW.
CW[3:0] (in ICW or PCW)	DIP4	4-bit Diagonal Interleaved Parity The 4-bit Diagonal Interleaved Parity (DIP4[3:0]) provides error control coding over the current Control Word and the preceding Data Words (and possibly Control Words). Odd parity is calculated diagonally over the covered words.

Training Control Words (in conjunction with Training Data Words) are used for data path de-skew. Their format and usage are described in Section 7.1.5, “Training Sequence for Data Path De-skew.”

7.1.1.2 Data Word Field Definitions

Control words and data words are differentiated by the CTL signal (as described in Section 5.4, “Interface Signals”).

Table 3 describes the fields used in ADW and PDW data words .

Table 3. Data Word Field Definitions

DW Bits	Mnemonic	Description
DW[15:0]	ADW	Address Data Word The Address Data Word contains 16 bits of address information. An ADW shall only appear according to the Extended Address Data Burst with ADW.
DW[15:0]	PDW	Payload Data Word The Payload Data Word contains 16 bits of payload (segment) data.

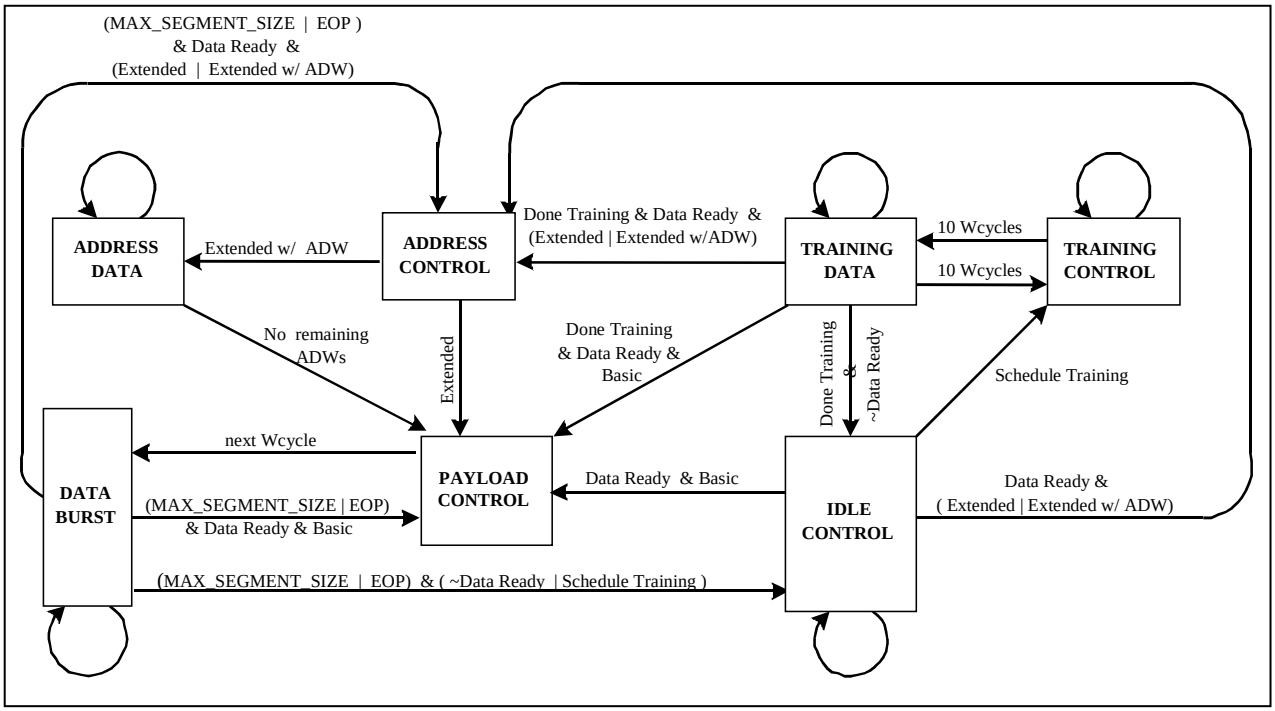
Training Data Words (in conjunction with Training Control Words) are used for data path de-skew. Their format and usage are described in Section 7.1.5, “Training Sequence for Data Path De-skew.”

7.1.2 Data Transfer Procedure

The proper sequence of Control and Data words on the data path constitutes a state machine. A state transition occurs for each 16-bit word cycle (Wcycle). Figure 9, below, shows the Wcycle-by-Wcycle

behavior of the data path state transitions. A transition is indicated by an arrow from one box to another.

Figure 9. Data Path State Diagram



For delivery of data over the NPSI, only states Idle Control, Address Control, Address Data, Payload Control, and Data Burst are involved. The remaining states, Training Control and Training Data, pertain to de-skew of the signals in the bus and will be described further in Section 7.1.5, “Training Sequence for Data Path .”Table 4 provides a description of the states.

Table 4. Data Path States

State	Description
Idle Control	This state handles the case where no data needs to be sent over the bus. An Idle Control word is sent in the current Wcycle.
Address Control	This state is entered at the beginning of each extended burst. An Address Control word is sent to deliver the first portion of a control sequence of addressing information. Additional addressing information may be sent in subsequent Address Data words and a Payload Control word.
Address Data	This state may be entered after the Address Control state to deliver the intermediate words of a control sequence. Transition into this state is only valid for Extended Address Data Bursts with ADW. Each Address Data word contains 2 bytes of the Address Data Field. Residence in this state is maintained until all but the final Payload Control word has been sent. The maximum duration in this state is 8 Wcycles. An Address Data word is sent for every Wcycle in this state.
Payload Control	This state is entered at the beginning of each basic burst. It is also entered from the Address Control or Address Data state for extended bursts. A Payload Control word is sent in this state.
Data Burst	The payload data (segment) of a data transfer is sent in this state. Residence is maintained until the first of the following two events occurs: MAX_SEGMENT_SIZE bytes have been sent or an end-of-packet is encountered. A Payload Data Word is sent for every Wcycle in this state.
Training Control	This state provides the Control word portion of the de-skew training pattern. A Training Control word is sent for every Wcycle in this state. Training is described in detail in Section 7.1.5.
Training Data	This state provides the Data word portion of the de-skew training pattern. A Training Data word is sent for every Wcycle in this state. Training is described in detail in Section 7.1.5.

7.1.3 Packet Delineation

EOP/SOP delineation applies to the multiplexing context in the Address Data Field (ADF), as explained in section 7.1.1.

No further multiplexing capability is defined in this specification.[11]

There is no minimum or maximum packet size assumed by this specification.

7.1.4 Error Detection

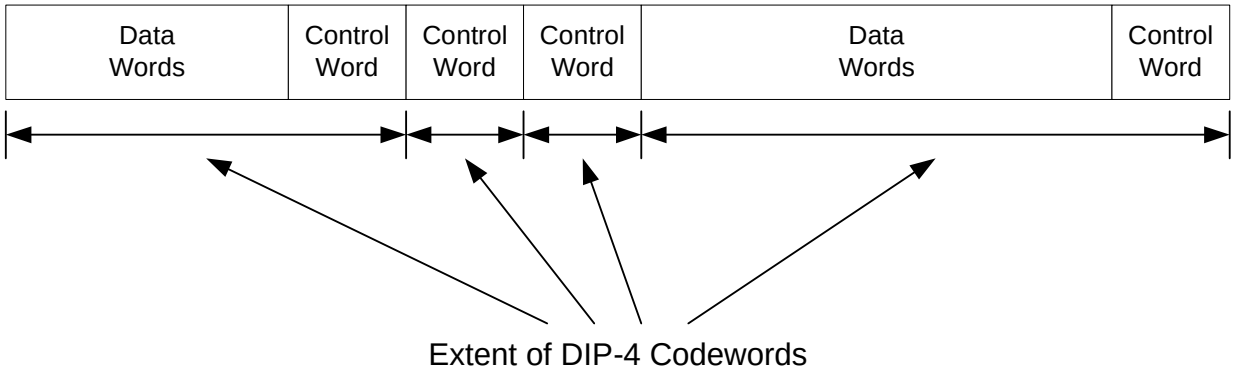
Error detection is provided using a 4-bit Diagonal Interleaved Parity (DIP-4)[12]. The Idle, Training, and Payload Control Words carry a DIP-4. The Address Control Word does not carry a DIP-4.

Unused bits in the ADF shall be transmitted as zero and included in the DIP-4 calculation. On reception, they shall be used only for the calculation of DIP-4 and otherwise ignored.

The range over which the DIP-4 parity bits are computed is from immediately after the last DIP-4 (in an ICW, PCW, or TCW) to the current control word (ICW, PCW, or TCW) containing the current DIP-4[13]. In the presence of random errors, DIP-4 offers the same error detection capability as a comparable BIP (Bit Interleaved Parity) code, but has an additional advantage of spreading single-column errors (as might occur in a single defective line) across the parity bits. Appendix A of the SPI-4 Phase 2 implementation agreement [1] discusses the error detection performance of this code.

Figure 10 and Figure 11 show the range over which the DIP-4 parity bits are computed.

Figure 10. Extent of DIP-4 Coverage for Basic Addressing Mode

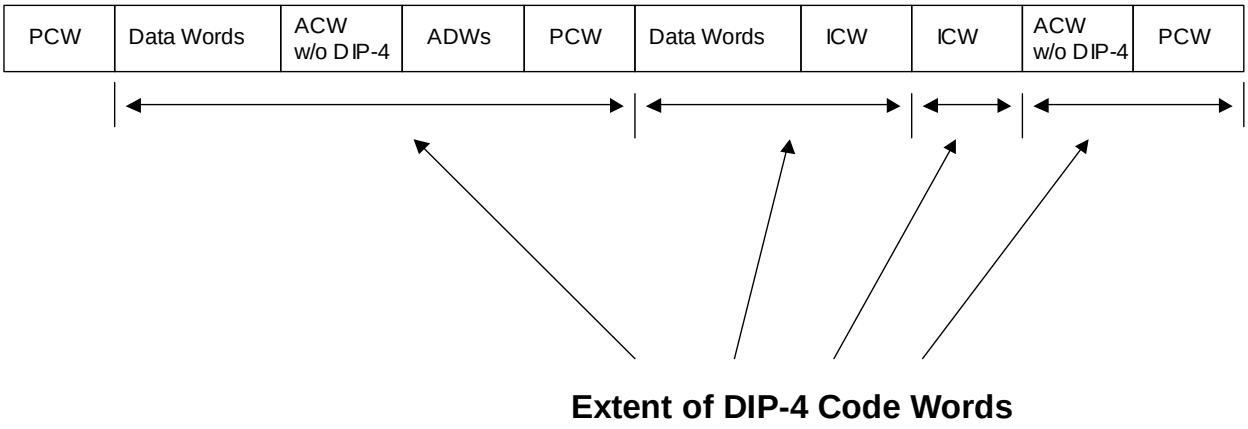


[11] If an additional level of multiplexing (multiple packets within a given ADF) is required, an additional tag embedded in the payload must be used. This is outside the scope of this specification.

[12] This is consistent with SPI-4 Phase 2 [1].

[13] Note that the DIP-4 is also calculated and checked from ICW to ICW, but this is not covering any packet data. From PCW to PCW, the DIP-4 covers a segment of data plus the control sequence of the next segment. From PCW to ICW, the DIP-4 covers a segment of data plus the ICW. From ICW to PCW, the DIP-4 covers the control sequence of the next segment. The DIP-4 in the TCW covers only the TCW. Note that, in some cases, a single DIP-4 error may cause two segments to be considered to be in error.

Figure 11. Extent of DIP-4 Coverage for Extended Addressing Mode (ACW & ADW)

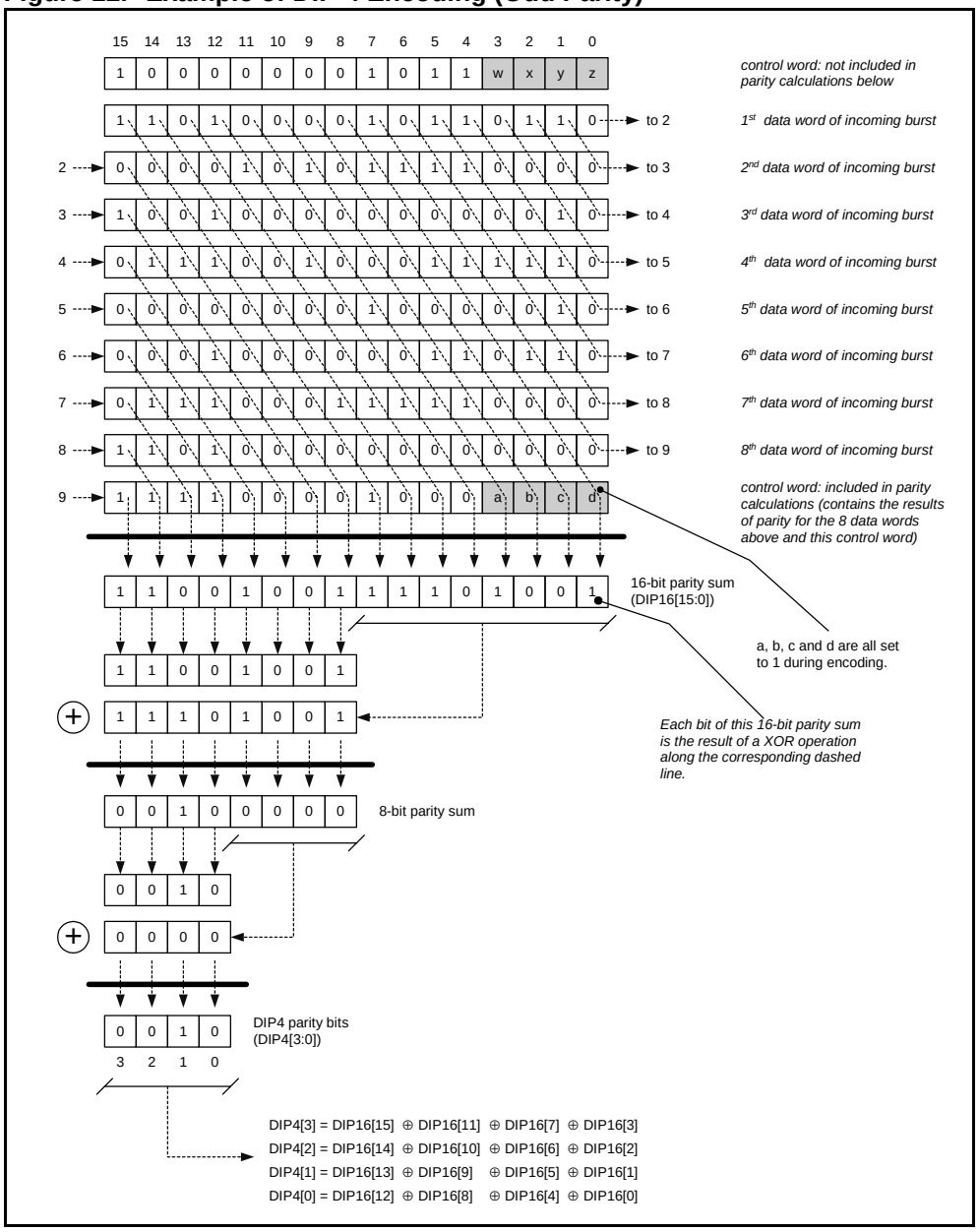


A functional description of calculating the DIP-4 code is given as follows. Assume that the stream of 16-bit data words are arranged in a vertical stack, MSB at the leftmost column, time moving downward. (The first word received is at the top; the last word is at the bottom.) The parity bits are generated by summing diagonally. (In the control word, the space occupied by the DIP-4 code (bits a, b, c, d) is set to all 1's during encoding.) The first 16-bit checksum is split into two bytes, which are combined (bitwise XOR) to produce an 8-bit checksum. The 8-bit checksum is then divided into two 4-bit nibbles, which are combined (bitwise XOR) to produce the final DIP-4 code[14]. The procedure described applies to either parity generation on data transmission or to check parity on data reception.

Note that the control signal, CTL, is not included in the DIP-4 calculation. An example is shown in Figure 12.

[14] This description is different from the description in the OIF SPI-4 Phase 2 and SPI-5 specifications ([1] and [2]) and is intended as a clarification. However, the process produces the same parity field values.

Figure 12. Example of DIP-4 Encoding (Odd Parity)



7.1.5 Training Sequence for Data Path De-skew

A training sequence is scheduled to be sent at least once every pre-configured bounded interval (DATA_MAX_T) on the data path. These training sequences may be used by the receiving end of each interface for de-skewing bit arrival times on the data (and corresponding control) lines. The sequence defined in this section is designed to allow the receiving end to correct for relative skew differences of up to +/- 1 bit time. The training sequence consists of one idle control word followed by one or more repetitions of a 20-word training pattern consisting of 10 (repeated) training control words and 10 (repeated) training data words. The initial idle control word removes dependencies of the DIP-4 in the training control words from preceding data words. Assuming a maximum of +/- 1 bit time in bit alignment jitter on each line, and a maximum of +/- 1 bit time relative skew between lines, there will be at least 8 bit times during which a receiver can detect a training control word prior to de-skew. The training data word is chosen to be to the binary complement of the training control word.

Table 5. Training Word Field Definitions

TW Bits	Mnemonic	Description
TW[15:0]	TCW	Training Control Word The TCW pattern is 0x0FFF. Bits [3:0] of the TCW contain the DIP-4 of the control word (by design always 0b1111).
TW[15:0]	TDW	Training Data Word The TDW pattern is 0xF000.

The sending (source) side of the data path must schedule the training sequence at least once every DATA_MAX_T interval, defined as DATA_MAX_T * 2^8 clock cycles, and repeat the training pattern DATA_ALPHA consecutive times, where DATA_MAX_T and DATA_ALPHA are configurable on start-up. Once the training sequence is scheduled, the source shall wait until the completion of the current data burst to start the transmission of the training sequence. The subsequent training sequence shall be no later than the DATA_MAX_T interval from the scheduled time, not transmission time, of the training sequence. Training sequences must not be inserted within a data transfer (i.e., not inserted between the first control word of a control sequence and any of the subsequent control or data words until the end of transfer). Setting DATA_MAX_T equal to zero shall disable the periodic transmission of the training sequence. The training sequence shall always be sent after reset before data is transmitted. A transmitter shall support all values of DATA_MAX_T up to 2^{32}-1 and shall provide a mechanism to configure the value. A transmitter shall support all values of DATA_ALPHA up to 255 and shall provide a mechanism to configure the value.

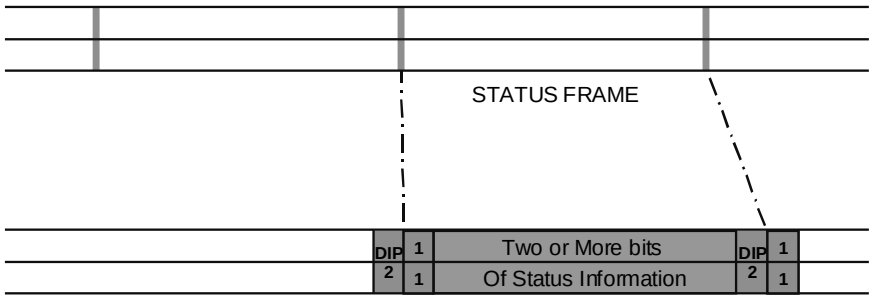
Note that the error-free reception of all 10 of the Training Data Words always results in a DIP-4 code of ‘0xF’ in a subsequent Idle Control Word. Likewise, if a Payload Control Word follows the training sequence, the DIP-4 result depends upon the content of the Payload Control Word. Also, if an Address Control Word (without a DIP-4 code) follows the training sequence, the DIP-4 result (in the Payload Control Word) depends upon the contents of the Address Control Word, any Address Data Words and the Payload Control Word.

7.2 Common Flow Control Path Operation

7.2.1 Flow Control Status Framing

Flow control (downstream queue status) information is sent from the data path sink to the data path source over the Status bus in a framing structure called a status frame. Each status frame contains one framing code, two or more bits of flow control (status) information, and one 2-bit diagonal interleaved parity (DIP-2) code.

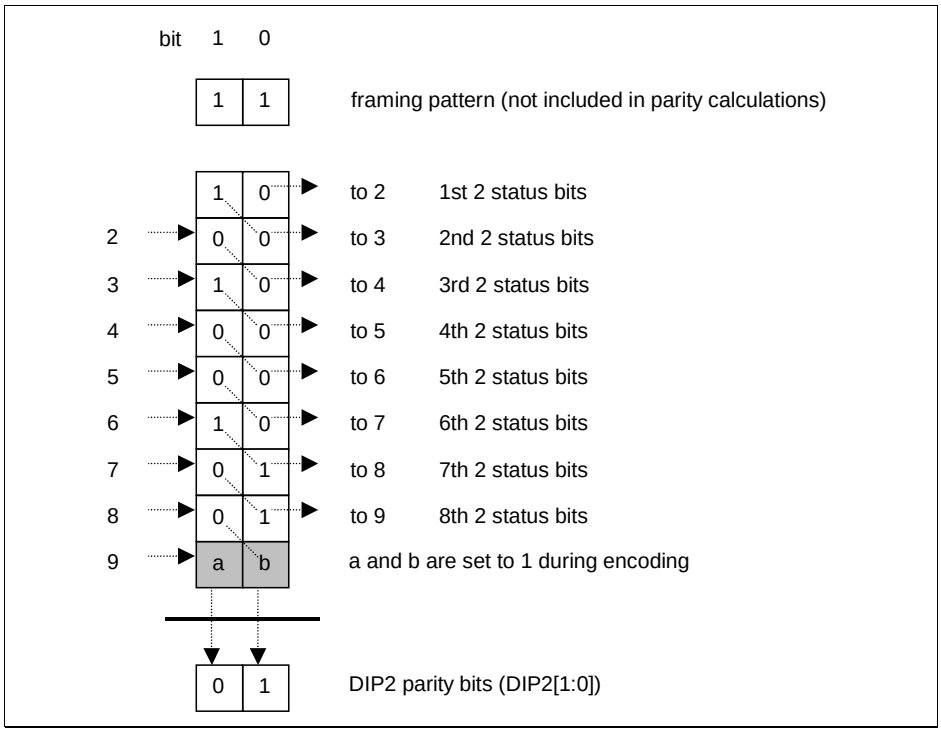
Figure 13. Common Status Channel Format



The status frame is sent as follows: The framing code, ‘0b11’, is sent, followed by the flow control (status) information, followed by a DIP-2 field. The status frame is then repeated, starting with the framing code. Three or more consecutive framing codes constitute a framing error. The DIP-2 field is an error check code that is computed diagonally over the contents of preceding status frame, not including the framing code, as shown below. The first 2-bit pair of the status frame immediately

following the framing code is at the top of the figure. The parity bits are computed by summing diagonally. Bits ‘a’ and ‘b’ in line 9 correspond to the space occupied by the DIP-2 parity bits, which are set to 1 during encoding. The procedure described applies to both parity generation and to parity checking functions.

Figure 14. DIP-2 Generation



Note that DIP-2 code and training patterns may both emulate the framing code (‘0b11’). The framer state machine must not confuse these occurrences with frame boundaries.

The signal data rate of the status lines is the same as the signal data rate of the data lines. A 4-bit status bus is optional (for the NPE-Fabric Mode only).

7.2.1.1 Training Sequence for Status Path De-skew

A Training Sequence is periodically inserted into the transmitted bit stream to allow the receiver to properly synchronize with the 2-bit (or 4-bit) signals of the status interface.

The training pattern consists of ten words of ‘0b00’ followed by ten words of ‘0b11’. (The training sequence was chosen so that it can be distinguished from a valid status channel message.) The training pattern can optionally be extended by repeating the 20-cycle pattern STAT_ALPHA times. The training sequence consists of the training pattern repeated STAT_ALPHA times.

A training sequence is scheduled to be sent at least once every pre-configured bounded interval (STAT_MAX_T) on the status interface. Once the training sequence is scheduled, the source shall wait until the completion of the current status frame to start the transmission of the training sequence. The subsequent training sequence shall be no later than the STAT_MAX_T interval, defined as STAT_MAX_T * 2^8 clock cycles, from the scheduled time, not transmission time, of the training sequence. Periodic status channel training can be disabled by setting the bus parameter STAT_MAX_T to ‘0’. The training sequence shall always be sent after reset and before normal operation begins. A transmitter shall support all values of STAT_MAX_T up to 2^{32}-1 and shall provide a mechanism to configure the value. A transmitter shall support all values of STAT_ALPHA up to 255 and shall provide a mechanism to configure the value.

These training sequences may be used by the receiving end of each interface for de-skewing bit arrival times on the status lines. The sequence defined in this section is designed to allow the receiving end to correct for relative skew differences of up to +/- 1 bit time.

The training sequence is inserted between two consecutive frames. Training sequences shall only be inserted between the DIP-2 checksum at the end of the preceding status frame and the '0b11' framing code of the following status frame.

7.2.1.2 Training Sequence for the 4-bit Status Path De-skew

The same de-skewing pattern and procedure is used for the 4-bit wide mode as for the 2-bit wide mode, except that the pattern is duplicated on both halves of the status bus interface; the training pattern consists of ten words of '0b0000' followed by ten words of '0b1111'.

The training pattern shall be run across the full width of the status interface. If the previous frame did not end on a 4-bit boundary, the transmitter shall insert an all zeroes pattern after the DIP-2 code to ensure that the training sequence runs across the full width of the status interface.

7.3 Loss of Synchronization

7.3.1 Loss of Data Path Synchronization (LODS)

An NPE-NPE and NPE-Fabric data path sink shall continuously monitor the data path for proper synchronization and error free operation. If the reaction to LODS is enabled, the following mechanism of detecting and reacting to LODS is required.

When a data sink detects multiple consecutive errors on the data path, it shall report a Loss of Data Synchronization (LODS) condition. This alarm condition is reported by sending a LODS alarm pattern back to the data path source on the status channel. Although no specific number of data path errors is specified, it is a requirement that any single error event shall not cause loss of synchronization.

The LODS alarm status pattern shall be sent immediately when a LODS condition is detected, without regard to the current position in the status frame. The LODS pattern shall be repeated continuously until the data path has regained synchronization, having received sufficient training patterns.[15]

When a data path source detects a LODS alarm status pattern on its corresponding status bus, resulting in status out-of-frame state (as described below), the data path source shall cancel all previously granted credits (NPE-NPE mode) or set its internal DR state to de-asserted (NPE-Fabric mode). Credits are only refreshed (or the internal DR state set to asserted) after the status channel returns to the in-frame state (as described below). While the status frame receiver is in out-of-frame state, the data path transmitter shall send a training sequence on the data path. The data path transmitter shall send at least four valid control words after completion of a training sequence before sending a subsequent training sequence due to the LODS alarm (status out-of-frame state).

The LODS alarm status pattern consists of the training pattern (as described in Section 7.2.1.1) sent at least (STAT_ALPHA+2) times. Receipt of no more than 12 repetitions of the LODS pattern shall be sufficient to force the status frame receiver to the out-of-frame state.

7.3.2 Loss of Status Path Synchronization (LOSS)

An NPE-NPE and NPE-Fabric status path sink shall continuously monitor the status path for proper synchronization and error free operation. If the reaction to Loss of Status Synchronization (LOSS) is enabled, the following mechanism of detecting and reacting to LOSS is required.

When a status sink detects multiple consecutive errors on the status path, it shall report a Loss of Status Synchronization (LOSS) condition. The status path sink shall cancel all previously granted credits (NPE-NPE mode) or set its internal DR state to de-asserted (NPE-Fabric mode). Credits are only refreshed (or the internal DR state set to asserted) after the status channel returns to the in-frame state. This alarm condition is reported by sending a LOSS alarm pattern back to the status path source on the data channel. Although no specific number of status path errors is specified, it is a requirement that any single error event shall not cause loss of synchronization.

[15] This behavior is consistent with the SPI-4 Phase 2 and SPI-5 behavior for LODS, since training patterns contain multiple control words.

The LOSS alarm status pattern shall be sent at the completion of the current data burst when a LOSS condition is detected. The LOSS pattern shall be repeated continuously until the status path has regained synchronization and has received multiple error-free status frames. Receipt of no more than 12 repetitions of the LOSS pattern shall be sufficient to force the data frame receiver into a LOSS alarm state. In LOSS alarm state the status path transmitter shall send a training sequence. The status path transmitter shall send at least four valid status frames after completion of each training sequence before sending a subsequent training sequence due to the LOSS alarm.

The status sink returns to the in-frame state when it observes multiple consecutive error-free status frames.

The LOSS alarm status pattern consists of the training pattern (as described in Section 7.1.5, “Training Sequence for Data Path De-skew,” sent at least (DATA_ALPHA+2) times.

8. NPE-Fabric Mode

8.1 Functional Description

The NPSI can be used to support connectivity between a switch fabric and a network processing element.[16]

The LVDS electrical specifications from Section 10, “Physical Layer” shall be used for all interface signals.

The switch fabric performs the physical layer transportation of user data elements from an ingress NPE to one or more egress NPE’s.

The interface supports up to 4096 switch fabric ports for connecting NPEs to the fabric. Ports are addressable endpoints of the NPSI, and multiple logical ports (sub-ports) can share a single physical port. The NPSI provides a label for communicating class of service levels to the switch fabric from the NPE. Classes may be used to isolate traffic, manage priority, or differentiate other QoS services provided by the fabric. The number of classes supported and the class differentiation provided by the fabric, as well as the use of sub-ports, are determined by the fabric and are vendor-specific.

8.2 Data Path Operation

The flows from an NPE (typically corresponding to fabric queues on the ingress interface) can be managed via NPE-based traffic shaping or via fabric-based flow control (or both).

Refer to Section 7.1, “Common Data Path Operation” for a description of the fundamentals of the data path operation.

A fabric shall support a MAX_SEGMENT_SIZE of 64 bytes.[17] Support for any other MAX_SEGMENT_SIZE is optional. A fabric may support multiple values of MAX_SEGMENT_SIZE; if a fabric supports more than one value of MAX_SEGMENT_SIZE, it shall provide a mechanism for configuring this value before normal operation of the interface begins.

An NPE sending data to or receiving data from a fabric shall support a MAX_SEGMENT_SIZE of 64 bytes. Support for any other MAX_SEGMENT_SIZE is optional. An NPE may support multiple values of MAX_SEGMENT_SIZE; if an NPE supports more than one value of MAX_SEGMENT_SIZE, it shall provide a mechanism for configuring this value before normal operation of the interface begins.

The NPE-Fabric data path uses the training operation defined in Section 7.1.5, “Training Sequence for Data Path .”

8.2.1 Data Framing

In the NPE-Fabric Mode, there are two possible data burst formats, depending on the control sequence, as described in Section 7.1, “Common Data Path Operation:”

- Extended Address Data Burst without ADW
- Extended Address Data Burst with ADW

An NPSI device operating in NPE-Fabric Mode shall support the “Extended Address Data Burst without ADW” format. If the device supports multicast, it may also support the “Extended Address Data Burst with ADW” format.

[16] In CSIX-L1, the network processing element is referred to as a Traffic Manager (TM).The network processing element (NPE) may also be a network processor or coprocessor.

[17] For those readers familiar with the CSIX-L1 specification, a segment is conceptually similar to a CFrame.

In the NPE-Fabric Mode, Payload Control Words also contain an indication of the address type (AT):

- Unicast,
- Multicast ID (ingress only),
- Multicast Bitmap (ingress only), or
- Egress Multicast (egress only).

Any AT may use the Extended address data burst without ADW. Only the multicast bitmap AT may use the extended address data burst with ADW.

Table 6, below, contains the AT encoding for the NPE-Fabric Mode.

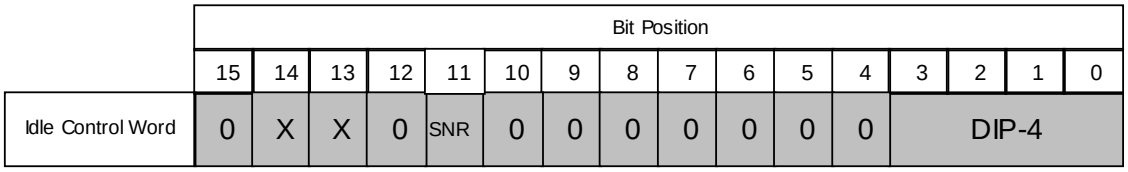
Table 6. Address Type Encoding for the NPE-Fabric Mode

CW Bits	Mnemonic	Description
CW[14:13]	EOPS/AT	End-of-Packet Status (in ICW or ACW) or Address Type (in PCW) The End-of-Packet Status Field EOPS[1:0] is present in Idle Control Words or Address Control Words and reports the status of the payload transfer immediately preceding it. The EOPS field is defined in Table 2 on page 17. The Address Type field AT[1:0] is present only in Payload Control Words and indicates the type of the data transfer. AT[1:0] = 0 0 : Unicast AT[1:0] = 0 1 : Egress Multicast AT[1:0] = 1 0 : Multicast Identifier AT[1:0] = 1 1 : Multicast Bitmap AT[1:0] is mapped to CW[14:13].

8.2.2 Status Not Ready Bit

The SNR bit is an optional Status Not Ready indication. The Idle Control Word can be used to convey the SNR bit. If an NPSI device supports this option, the following Idle Control Word format is used.

Figure 15. Idle Control Word with SNR Bit



When the SNR bit is not used, bit 11 should always have a value of 0. The SNR bit is described in detail in Section 8.4.9, “Flow Control of Directed Status.”

8.2.3 Data Transfer Procedure

Refer to Section 7.1.2 for the data transfer procedure. Note that only Extended bursts are allowed in the NPE-Fabric Mode.

8.3 Addressing

The Address Data Field (ADF) contains information relevant to the queueing, flow control, class of service treatment, multicasting, and multiplexing of data across the switch fabric. The ADF contains a Physical Port ID field and a Class field, and may contain additional Sub-port ID fields. If a fabric supports the sub-port formats, it shall support the same sub-port configuration for unicast and multicast. Support for a specific number of sub-ports, classes, multicast ID values, or multicast bitmap

sizes is not implied by the format; any field may have zero or more unused bits. If any bits of a given field are not used, the value shall be right-justified and the MSB's shall be unused (set to zero on transmission and ignored on reception).

The address field has multiple definitions. Required and optional address formats are described below.

There are two fundamental address types, unicast and multicast, that are explicitly supported. Unicast, ingress multicast ID, and egress multicast address types have a 20-bit ADF. Multicast bitmap address types may have an ADF of more than 20 bits.

8.3.1 Requirements of an NPSI Switch Fabric

The NPSI defines a physical interface between NPEs and switch fabrics. It specifies a data transfer format and an addressing mechanism (label format) for an NPE to indicate the forwarding and class of service treatment of the data to the switch fabric (within the addressing and CoS domain of the fabric).

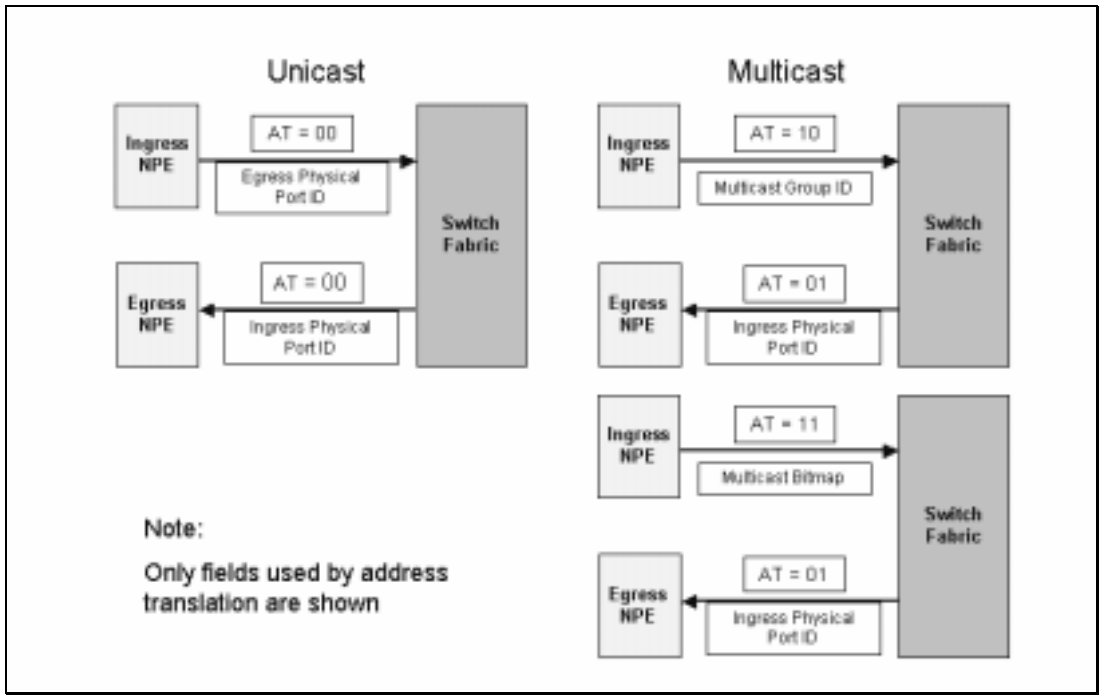
A switch fabric conforming to the NPSI shall ensure sequence integrity of data flow as described below.

A switch fabric shall provide at the egress of the fabric the address of the ingress port in the designated field[18] as described below.

8.3.1.1 Address Swapping Operation

The below reference diagram illustrates the address format changes that occur to a burst segment as it traverses from the NPE to Fabric (Fabric Ingress) port to the Fabric to NPE (Fabric Egress) port.

Figure 16. Address Swapping Operation



[18] Because the EOP/SOP delineation in the control word is used to identify packets multiplexed across the interface, the egress NPE can use the addressing provided to demultiplex packets based on the AT value, the ingress port ID, and the class field.

8.3.1.2 Sequence Integrity

A switch fabric conforming to the NPSI shall ensure that all data and packet boundaries within a flow at the fabric egress interface are in the same sequence as when the data was sent to the fabric on its ingress interface. There are no such guarantees with regard to data of different flows or from different ingress ports presented to the fabric for delivery to the same egress port.

Unicast and multicast flows are disjoint (i.e. a given flow is either unicast or multicast).

At any reference point where multiplexing can occur, the context for maintaining sequence integrity is the Flow Identifier. The Flow Identifier is a set of values designating a flow at specific reference points in the Streaming Interface architecture. The Flow Identifier sets are as follows.

Unicast:

- Ingress (NPE-to-Fabric) interface Unicast Flow Identifier:
  {Class, Egress Port ID[19], Ingress Sub-port ID, Unicast}
- Egress (Fabric-to-NPE) interface Unicast Flow Identifier:
  {Class, Egress Sub-port ID, Ingress Port ID, Unicast}

Multicast:

- Ingress (NPE-to-Fabric) interface Multicast ID Flow Identifier:
  {Class, Ingress Sub-port ID, Multicast}
- Ingress (NPE-to-Fabric) interface Multicast Bitmap Flow Identifier:
  {Class, Ingress Sub-port ID, Multicast}[20]
- Egress (Fabric-to-NPE) interface Multicast Flow Identifier:
  {Class, Egress Sub-port ID, Ingress Port ID, Multicast}

The use of certain fields, such as the Ingress Port ID (and Ingress Sub-Port ID) or Class, is optional. If the field is not used, the value does not appear as part of the flow identifier.

8.3.2 Summary of Address Formats

All NPSI devices operating in NPE-Fabric Mode shall support the ingress and egress unicast address formats. NPSI devices that support multicast shall support one of the multicast formats, and NPSI devices that support sub-ports shall support the corresponding sub-port format(s).

8.3.2.1 Unicast Address Formats

There are five fields defined in the unicast ADF:

- Egress Physical Port (ingress interface only)
- Ingress Physical Port (egress interface only)
- Class
- Ingress Sub-port
- Egress Sub-port

[19] In these definitions, the Port ID includes the Physical Port ID and the Sub-port ID, if applicable.

[20] Because of the requirement that order is preserved within a flow, the Multicast Flow Identifier definitions implicitly place a constraint on a switch fabric. That is, the fabric must preserve order on all multicast bitmap or multicast ID traffic of a given class (and ingress sub-port) on the ingress multicast queue, and can not rely on the bitmap or MC ID label as defining the flow for the purpose of preserving order within a flow.

There are three unicast address formats:

Table 7. Unicast Address Formats Summary

	Ingress Format				Egress Format			
	Egress Physical Port ID (bits)	Egress Sub-Port ID (bits)	Ingress Sub-Port ID (bits)	Class (bits)	Ingress Physical Port ID (bits)	Egress Sub-Port ID (bits)	Ingress Sub-Port ID (bits)	Class (bits)
Physical Port Addressing	12	n/a	n/a	8	12	n/a	n/a	8
Sub-port Addressing Option 1	8	4	4	4	8	4	4	4
Sub-port Addressing Option 2	10	2	2	6	10	2	2	6

An NPSI device operating in NPE-Fabric Mode shall support the Unicast Physical Port Addressing format and may support one or both of the optional Unicast Sub-port addressing formats. (Support for the sub-port addressing formats is optional.)

- A fabric may support one of the two sub-port addressing formats without regard to support for the other.
- An NPE shall support either none or both of the sub-port formats.

If an NPSI device supports more than one unicast addressing format, it shall provide a mechanism to configure the device to use one and only one unicast addressing format.

8.3.2.2 Multicast ID Address Formats

There are five fields defined in the Multicast ID ADF:

- Multicast ID (ingress interface only)
- Ingress Physical Port (egress interface only)
- Class
- Ingress Sub-port
- Egress Sub-port (egress interface only)

The multicast formats are independent of the unicast formats, except that if sub-ports and multicast are supported, then the same sub-port addressing option shall be supported for unicast and multicast.[21]

[21] Note that a fabric may support sub-port unicast addressing but only multicast to a physical port level. In this case, the same sup-port addressing option must be supported for unicast and multicast even if the fabric is only able to multicast to a physical port level. The ingress sub-port field contains information used for reassembly.

There are four ingress multicast ID address formats and three egress formats:

Table 8. Multicast ID Address Formats Summary

	Ingress Format				Egress Format			
	Multicast ID (bits)	Egress Sub-Port ID (bits)	Ingress Sub-Port ID (bits)	Class (bits)	Ingress Physical Port ID (bits)	Egress Sub-Port ID (bits)	Ingress Sub-Port ID (bits)	Class (bits)
Multicast Addressing Format 1	12	n/a	n/a	8	12	n/a	n/a	8
Multicast Addressing Format 2	16	n/a	n/a	4				
Sub-port Addressing Option 1	12	n/a	4	4	8	4	4	4
Sub-port Addressing Option 2	12	n/a	2	6	10	2	2	6

Support for the Multicast ID Address format is optional. An NPE device operating in NPE-Fabric Mode that supports the Multicast ID option shall support the Multicast Addressing Formats 1 and 2. An NPSI fabric device operating in NPE-Fabric Mode that supports the Multicast ID option shall support the Multicast Addressing Format 1, Format 2, or both formats.

Since sub-port addressing is also optional, an NPSI device operating in NPE-Fabric Mode that supports the Multicast ID option may also support one or both of the optional Sub-port addressing formats.

- A fabric may support one of the two sub-port addressing formats without regard to support for the other.
- An NPE shall support either none or both of the sub-port formats.

If an NPSI device supports more than one multicast ID addressing format, it shall provide a mechanism to configure the device to use one and only one multicast ID format.

8.3.2.3 Multicast Bitmap Address Formats

There are five fields defined in the Multicast Bitmap ADF:

- Multicast Bitmap (ingress interface only)
- Ingress Physical Port (egress interface only)
- Class
- Ingress Sub-port
- Egress Sub-port (egress interface only)

The multicast formats are independent of the unicast formats, except that if sub-ports and multicast are supported, then the same sub-port addressing option shall be supported for unicast and multicast.

There are four multicast bitmap address formats and three egress formats:

Table 9. Multicast Bitmap Address Formats

	Ingress Format				Egress Format			
	Multicast Bitmap (bits)	Egress Sub-Port ID (bits)	Ingress Sub-Port ID (bits)	Class (bits)	Ingress Physical Port ID (bits)	Egress Sub-Port ID (bits)	Ingress Sub-Port ID (bits)	Class (bits)
Multicast Addressing Format 1	12+16*n[22]	n/a	n/a	8	12	n/a	n/a	8
Multicast Addressing Format 2	16+16*n	n/a	n/a	4				
Sub-port Addressing Option 1	12+16*n	n/a	4	4	8	4	4	4
Sub-port Addressing Option 2	12+16*n	n/a	2	6	10	2	2	6

Support for the Multicast Bitmap Address format is optional. An NPE device operating in NPE-Fabric Mode that supports the Multicast Bitmap option shall support the Multicast Addressing Formats 1 and 2. An NPSI fabric device operating in NPE-Fabric Mode that supports the Multicast Bitmap option shall support the Multicast Addressing Format 1, Format 2, or both formats.

Since sub-port addressing is also optional, an NPSI device operating in NPE-Fabric Mode that supports the Multicast Bitmap option may also support one or both of the optional Sub-port addressing formats.

- A fabric may support one of the two sub-port addressing formats without regard to support for the other.
- An NPE shall support either none or both of the sub-port formats.

If an NPSI device supports more than one multicast bitmap addressing format, it shall provide a mechanism to configure the device to use one and only one multicast bitmap format.

8.3.3 Unicast Addressing

Each transfer of user payload (each data segment) that is going to a single destination (i.e., unicast traffic) is associated with a control sequence, comprising of a single Address Control Word followed by a single Payload Control Word (the Extended address data burst without ADW). The egress (destination) port of a packet conveyed via the unicast service is designated by a Port Identifier (which may be a physical port ID or a physical port ID with a sub-port ID).

The Address Type field in the Payload Control Word identifies the data transfer as a unicast burst. The Address Data Field of the two Control Words combined (i.e., the ACW and PCW) is 20 bits. It is partitioned into several sub-fields, some of which are required and some of which are optional. It always contains a Physical Port ID (i.e., destination port address) as well as a class field to indicate the class to which the data burst belongs. Optionally, it may include Sub-port ID's.

When the NPSI is used at an ingress port of a switch fabric (NPE to fabric data transfer), the Address Data Field contains the Egress Port ID and the class (and optionally an Ingress and Egress Sub-port ID). When the NPSI is used at an egress port of a switch fabric (fabric to NPE data transfer), the Address Data Field contains the Ingress Port ID and the class (and optionally an Egress and Ingress Sub-port ID). The switch fabric shall replace the Egress Physical Port ID (on the ingress interface)

[22] The value 'n' may be from 0 through 8, inclusive. If a switch fabric supports more than one value of 'n', it should provide a mechanism to configure that value. An NPE should support all values of 'n'.

with the Ingress Physical Port ID (on the egress interface), as described in Section 8.3.1.1, “Address Swapping Operation.” The switch fabric shall preserve the used bits of the class field and sub-port ID fields. If a switch fabric does not utilize all bits of the class (or sub-port) field, the switch fabric is not required to preserve the unused class bits. At the egress interface of a fabric, the ingress port ID may be used (along with the AT and class and possibly sub-port information) by the receiving NPE for demultiplexing of unicast packets that were multiplexed across the fabric from multiple sources.

The class number specifies traffic isolation within the fabric and possibly different quality of service handling. The class number should not be used to specify individual sub-ports within a switch fabric port, since the NPSI does not ensure interoperability if the class field is used to specify sub-ports.

8.3.3.1 Ingress Unicast Addressing

Figure 17. Unicast Ingress Address Format: Physical Port Addressing

	Bit Position															
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Address Control Word	0	EOPS		1	EPP ID (3:0)				Class (7:0)							
Payload Control Word	1	0	0	S	Egress Physical Port ID (11:4)								DIP-4			

Note: EPP ID: Egress Physical Port ID

Figure 18. Unicast Ingress Address Format: Sub-port Option 1

	Bit Position															
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Address Control Word	0	EOPS		1	ESP ID (3:0)				ISP ID (3:0)				Class (3:0)			
Payload Control Word	1	0	0	S	Egress Physical Port ID (7:0)								DIP-4			

Note: ESP ID: Egress Sub-Port ID

Note: ISP ID: Ingress Sub-Port ID

Figure 19. Unicast Ingress Address Format: Sub-port Option 2

	Bit Position															
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Address Control Word	0	EOPS		1	EPP ID (1:0)		ESP ID (1:0)		ISP ID (1:0)		Class (5:0)					
Payload Control Word	1	0	0	S	Egress Physical Port ID (9:2)								DIP-4			

Note: EPP ID: Egress Physical Port ID

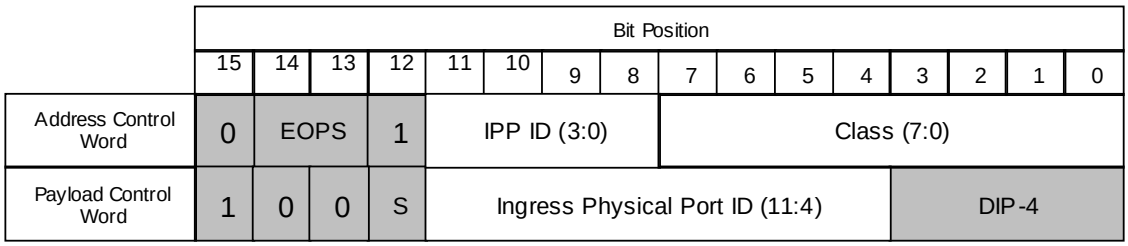
Note: ESP ID: Egress Sub-Port ID

Note: ISP ID: Ingress Sub-Port ID

At the fabric ingress interface, the egress port ID indicates a target across a switch fabric, at the level of resolution supported by the fabric. For example, a target may be a port to an NPE, or may be a logical interface (a sub-port of a physical interface). The switch fabric shall use the Egress Physical Port ID (and Egress Sub-port ID, if applicable) to determine to which output port it sends the data.

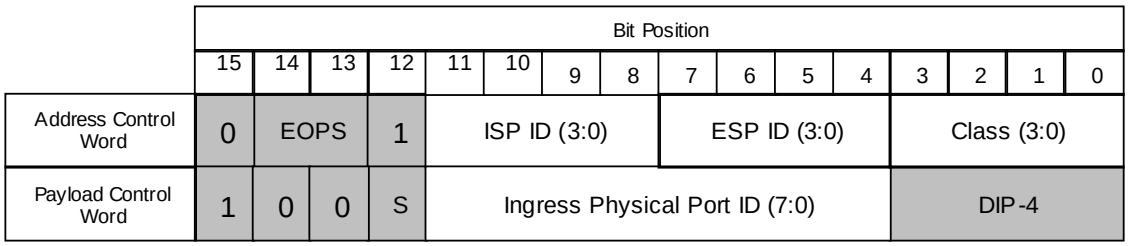
8.3.3.2 Egress Unicast Addressing

Figure 20. Unicast Egress Address Format: Physical Port Addressing



Note: IPP ID: Ingress Physical Port ID

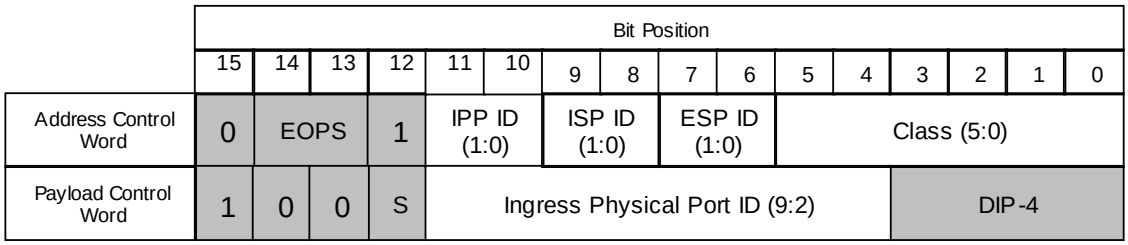
Figure 21. Unicast Egress Address Format: Sub-port Option 1



Note: ESP ID: Egress Sub-Port ID

Note: ISP ID: Ingress Sub-Port ID

Figure 22. Unicast Egress Address Format: Sub-port Option 2



Note: IPP ID: Ingress Physical Port ID

Note: ESP ID: Egress Sub-Port ID

Note: ISP ID: Ingress Sub-Port ID

8.3.4 Multicast Addressing

A switch fabric may optionally support multicasting of data. Each transfer of user payload that is to be transported to multiple egress ports of the fabric is associated with a control sequence that includes the information regarding how the data should be replicated. There are two multicast options defined for providing multicast service across the interface: bitmap based or multicast ID (tag-based). The

Address Type field in the Payload Control Word identifies the type of multicast (Multicast Identifier versus Multicast Bitmap). Although there is an AT code point that can differentiate between the two types of multicast traffic on a single interface, on a given instance of the NPE-Fabric interface, at most one multicast address format shall be active.[23]

Support for multicast is optional for both an NPE and a fabric.

8.3.4.1 Multicast ID Addressing (Ingress Only)

The multicast ID option uses a label that identifies a set of egress ports. Each transfer of user payload (each data segment) using this option is associated with a control sequence, comprising of a single Address Control Word followed by a single Payload Control Word (the Extended Address data burst without ADW). The set of egress ports of a packet conveyed via the multicast ID service is designated by a label, called a Multicast Identifier. (The fabric maps the label to the set of egress ports.)

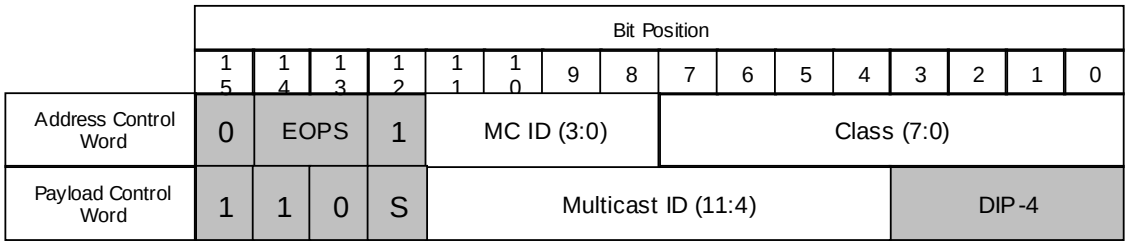
The Address Type field in the Payload Control Word identifies the corresponding segment as multicast traffic, using the Multicast ID to identify the set of egress ports. The Address Data Field of the control sequence is 20 bits. It is partitioned into several sub-fields, some of which are required and some of which are optional. It always contains a Multicast ID (i.e., the tag designating a set of Egress Port ID's) as well as a class field to indicate the class to which the data burst belongs. If the switch fabric supports multicasting to a sub-port level[24], the Multicast ID shall be a tag that designates a set of Egress Port ID's, where the Port ID's include the sub-port ID. Otherwise, the Multicast ID shall designate a set of Egress Physical Port ID's.

When the NPSI is used at an ingress port of a switch fabric (NPE to fabric data transfer), the Address Data Field contains the Multicast ID and the class (and optionally an Ingress Sub-port ID). The Multicast ID formats, when supported, shall be used only at the fabric ingress interface.

The switch fabric shall use the Multicast ID to determine to which egress ports it sends the data. Note, however, that the data may be queued in the switch fabric based solely on the class field before replication across the fabric. Also, in order for the egress NPE to be able to reassemble multicast packets (as explained below), the ingress NPE shall not interleave segments of packets from multiple multicast ID's within a multicast class and ingress sub-port (if applicable).

The class number specifies traffic isolation within the fabric and possibly different quality of service handling. The class number should not be used to specify individual sub-ports within a switch fabric destination port, since the NPSI does not ensure interoperability if the class field is used to specify sub-ports.

Figure 23. Multicast ID Address Format: Format 1



Note: MC ID: Multicast ID

[23] A given switch fabric may choose to accept only one of the two types of multicast traffic.

[24] If a fabric supports multicasting to a sub-port level, it supports sending one copy of the segment per sub-port (i.e., one copy per sub-port associated with a given physical port, as designated by the multicast ID or bitmap).

Figure 24. Multicast ID Address Format: Format 2

	Bit Position															
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Address Control Word	0	EOPS		1	Multicast ID (7:0)								Class (3:0)			
Payload Control Word	1	1	0	S	Multicast ID (15:8)								DIP-4			

Figure 25. Multicast ID Address Format: Sub-port Option 1

	Bit Position															
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Address Control Word	0	EOPS		1	MC ID (3:0)				ISP ID (3:0)				Class (3:0)			
Payload Control Word	1	1	0	S	Multicast ID (11:4)								DIP-4			

Note: MC ID: Multicast ID

Note: ISP ID: Ingress Sub-Port ID

Figure 26. Multicast ID Address Format: Sub-port Option 2

	Bit Position															
	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
Address Control Word	0	EOPS		1	MC ID (3:0)				ISP ID (1:0)		Class (5:0)					
Payload Control Word	1	1	0	S	Multicast ID (11:4)								DIP-4			

Note: MC ID: Multicast ID

Note: ISP ID: Ingress Sub-Port ID

8.3.4.2 Multicast Bitmap Addressing (Ingress Only)

Multicast bitmap is intended for use in switch fabrics with 128 or fewer ports.

The multicast bitmap option uses a bit vector (or bitmap) that identifies the members of a multicast group. The bits are in big-endian order (within words). The words of the bitmap are transferred in little endian order. Each transfer of user payload (each data segment) using this option is associated with a control sequence, using the Extended Address data burst without ADW or the Extended Address data burst with ADW.

The set of egress ports of a packet conveyed via the Multicast Bitmap service is designated by a bitmap. Each bit in the bitmap corresponds to an egress port ID.

The Address Type field in the Payload Control Word identifies the corresponding segment as multicast traffic, using the multicast bitmap to identify the egress ports. The Address Data Field is partitioned into several sub-fields, some of which are required and some of which are optional. It always contains a Multicast Bitmap, as well as a class field to indicate the class to which the data

burst belongs. If the switch fabric supports multicasting to a sub-port level, the multicast bitmap designates a set of Egress Port ID's, where the Port ID's include the sub-port ID. Otherwise, the bitmap shall designate a set of Egress Physical Port ID's.

- Format 1 and 2 or when physical ports are specified in the bitmap.

  Bitmap Index = Physical Port ID

- Subport Option 1 when subports are specified in the bitmap

  Bitmap Index = Physical Port ID * 16 + Egress Subport ID

- Subport Option 2 when subports are specified in the bitmap

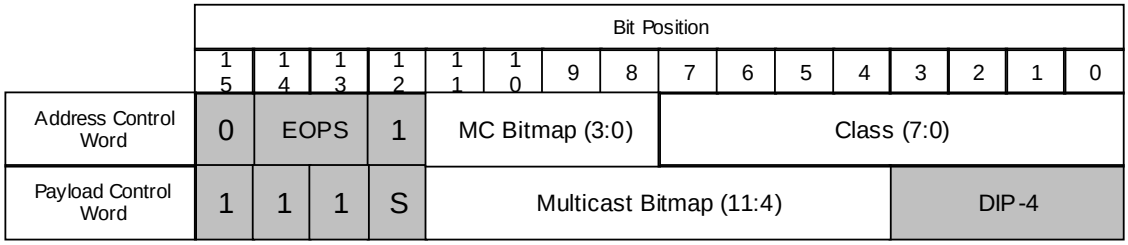
  Bitmap Index = Physical Port ID * 4 + Egress Subport ID

When the NPSI is used at an ingress port of a switch fabric (NPE to fabric data transfer), the Address Data Field contains the Multicast Bitmap and the Class fields (and optionally an Ingress Sub-port ID field). The Multicast Bitmap Addressing formats, when supported, shall be used only at the fabric ingress interface.

The switch fabric shall use the Multicast Bitmap to determine to which egress ports it sends the data. Note, however, that the data may be queued in the switch fabric based solely on the class field before replication across the fabric. Also, in order for the egress NPE to be able to reassemble multicast packets (as explained below), the ingress NPE shall not interleave segments of multicast packets within a multicast class and ingress sub-port.

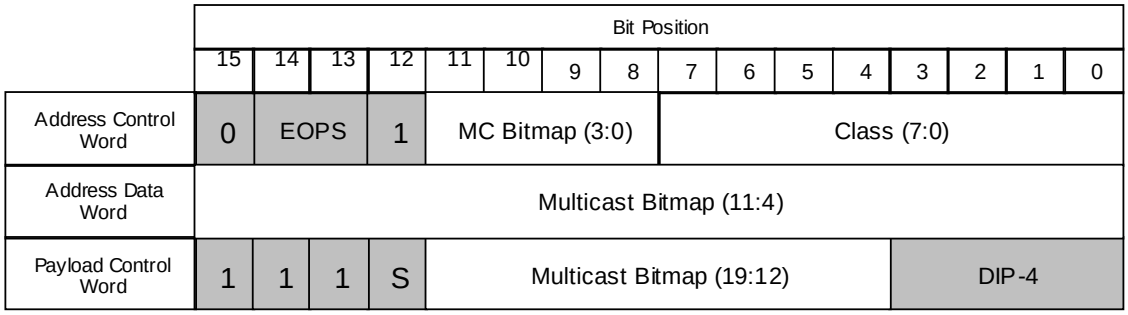
The class number specifies traffic isolation within the fabric and possibly different quality of service handling. The class number should not be used to specify individual sub-ports within a switch fabric destination port, since the NPSI does not ensure interoperability if the class field is used to specify sub-ports.

Figure 27. Multicast Bitmap Address Format: Format 1



Note: MC: Multicast

Figure 28. Multicast Bitmap Address Format: Format 1 with ADW (example with 1 ADW)[25]



Note: MC: Multicast

[25] Note that up to 8 ADW's may be used with any multicast bitmap format. Only one example is shown here.

Figure 29. Multicast Bitmap Address Format: Format 2

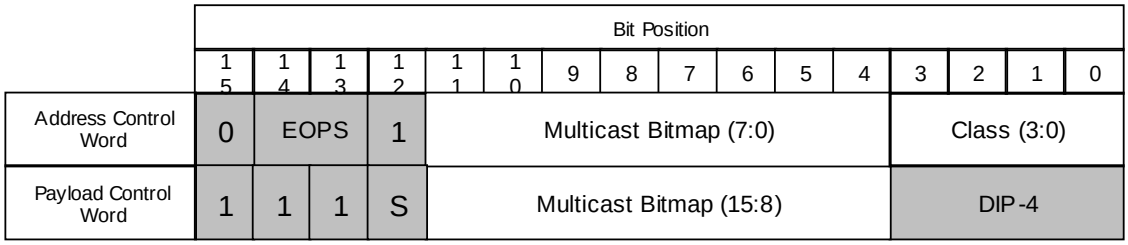
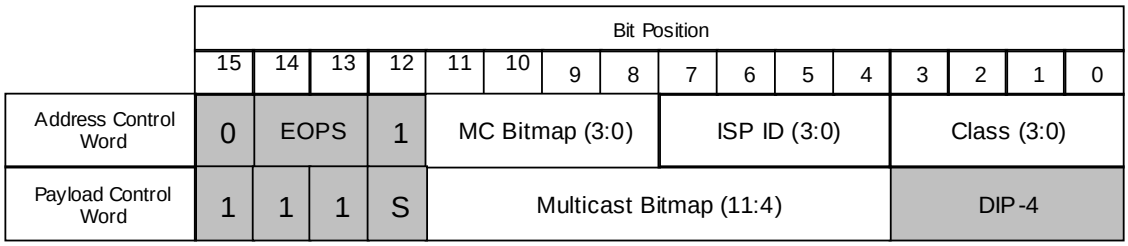


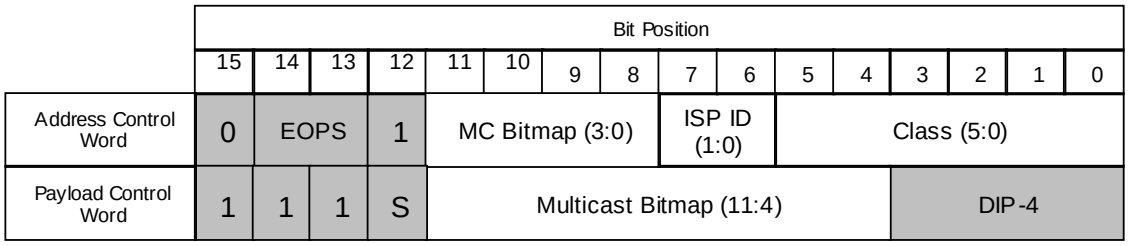
Figure 30. Multicast Bitmap Address Format: Sub-port Option 1



Note: MC: Multicast

Note: ISP ID: Ingress Sub-Port ID

Figure 31. Multicast Bitmap Address Format: Sub-port Option 2



Note: MC: Multicast

Note: ISP ID: Ingress Sub-Port ID

8.3.4.3 Multicast Egress Addressing

This format is used at the egress ports of multicast flows.

The Address Data Field contains the Ingress Port ID and the class (and optionally an Egress and Ingress Sub-port ID). The switch fabric shall preserve the used bits of the class field and sub-port ID fields. If a switch fabric does not utilize all bits of the class (or sub-port) field, the switch fabric is not required to preserve the unused bits. Unused bits shall be transmitted as zeroes. At the egress interface of a fabric, the ingress port ID may be used (along with the AT and class and possibly sub-port information) by the receiving NPE for demultiplexing of multicast packets that were multiplexed across the fabric from multiple sources.

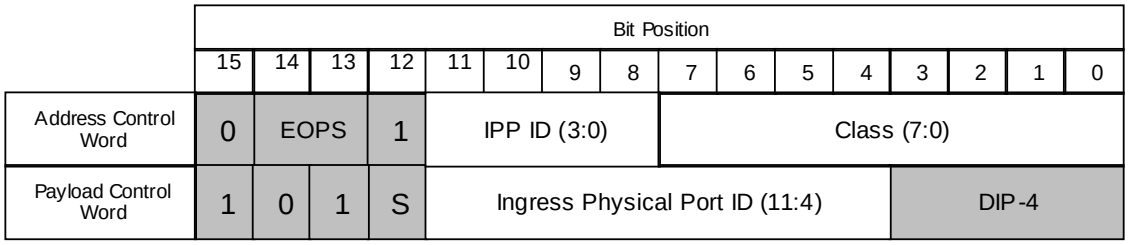
A switch fabric with egress sub-ports and multicast support may replicate multicast traffic across sub-ports as well as physical ports. (In this case, the fabric sends ‘n’ copies of the traffic across the egress physical port when configured to multicast to ‘n’ sub-ports of that physical port.) If the fabric does not

replicate the traffic to the sub-ports, it shall send the traffic once with the egress sub-port field set to zero. (Further replication is outside the scope of this specification.)

Note that there are similarities between the unicast and multicast egress address formats. Once data has been delivered to its destination, whether it's a single destination or multiple destinations, there is little distinction between the two types of flows except that their demultiplexing contexts must remain logically distinct from one another. This differentiation is provided by the AT code.

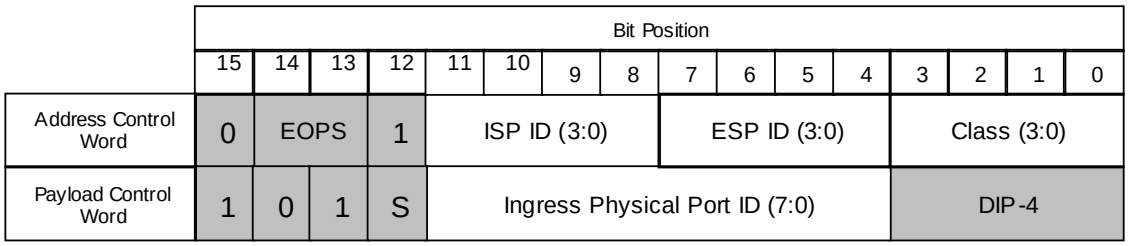
The class number specifies traffic isolation within the fabric and possibly different quality of service handling. The class number should not be used to specify individual sub-ports within a switch fabric port, since the NPSI does not ensure interoperability if the class field is used to specify sub-ports.

Figure 32. Multicast Egress Address Format: Physical Port Addressing



Note: IPP ID: Ingress Physical Port ID

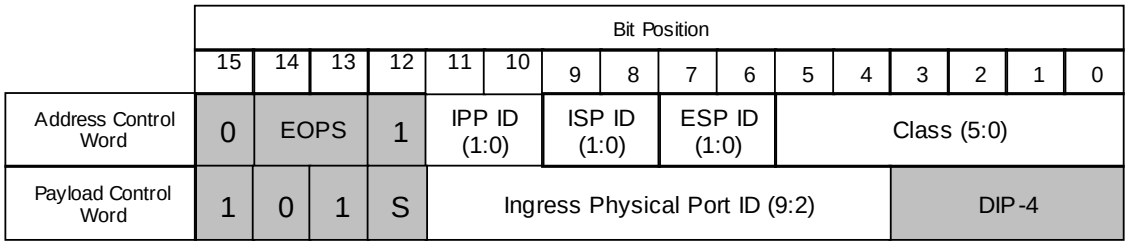
Figure 33. Multicast Egress Address Format: Sub-port Option 1



Note: ISP ID: Ingress Sub-Port ID

Note: ESP ID: Egress Sub-Port ID

Figure 34. Multicast Egress Address Format: Sub-port Option 2



Note: IPP ID: Ingress Physical Port ID

Note: ISP ID: Ingress Sub-Port ID

Note: ESP ID: Egress Sub-Port ID

8.4 Flow Control

Flow control information is transferred over a separate status bus, as described in Section 7.2, “Common Flow Control Path Operation.” The flow control channel carries only flow control messages (in addition to its required framing).

A device operating in the NPE-Fabric Mode shall support a status bus that is 2 bits wide and may support a status bus that is 4 bits wide. If a device implements a status bus that is 4 bits, it shall have a mechanism to configure the interface to operate in a 2-bit mode, by disabling the unused bits of the interface.

8.4.1 Flow Control Message Encoding and Framing

Flow control messages are assembled into flow control status frames. Each frame contains one framing code, one 36-bit (encoded) flow control message and one 2-bit diagonal in-line parity (DIP-2) code.

Frames are delineated by the 2-bit framing code (‘0b11’). To avoid framing code emulation, flow control messages are 3B4B encoded with the following mapping:

Table 10. Flow Control Message 3B4B Encoding

3B Data	4B Code
000	0000
001	0001
010	0010
011	1001
100	0100
101	0101
110	0110
111	1000

The following table shows the invalid code values.

Table 11. Invalid 3B4B Code Values

4B Invalid
1111
1110
1101
0011
1011
1010
1100
0111

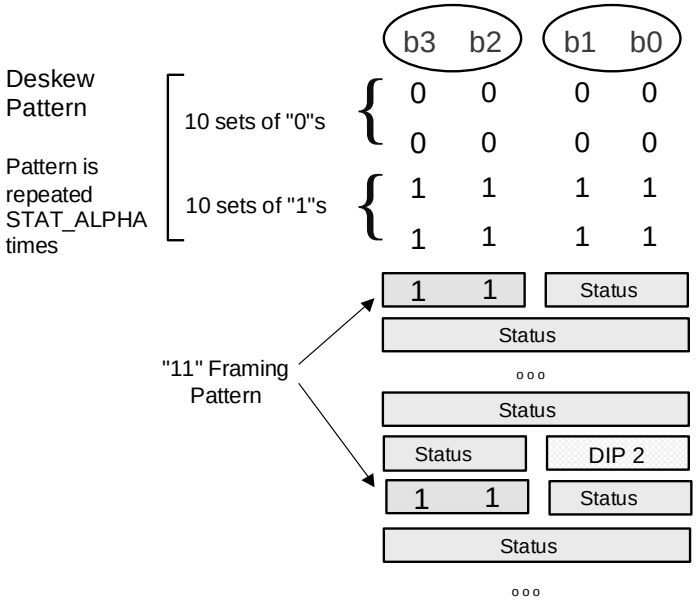
For the 3B4B encoding, each 27-bit message is divided into 9 sections and encoded starting with the most significant section first. The most significant 3-bit section of the flow control message maps to the most significant 4-bit section of the 36-bit encoded message. The 36-bit encoded message is transmitted most significant bit first.

The framing code and the DIP-2 field are not 3B4B encoded.

The framing of the status bus shall be validated before a flow control message is interpreted.

When a 4-bit status bus is used, the framing code shall always appear on bits (3:2), as shown in the following diagram.

Figure 35. Status Channel with Optional 4-Bit Wide Mode



8.4.2 Flow Control Mechanisms

The NPE-Fabric Mode of the NPSI provides several flow control mechanisms, some of which are required while others are optional.

A link-level flow control mechanism enables or disables data transmission on the entire data path. Only link-level flow control has a specified response requirement.

Directed status messages provide backpressure mechanisms that enable or disable transmission on a single flow, a logical grouping of flows or a subset of flows. Separate flow control mechanisms are provided on ingress and egress sides of the fabric.

Some ingress flow control mechanisms include sub-port granularity as an option.

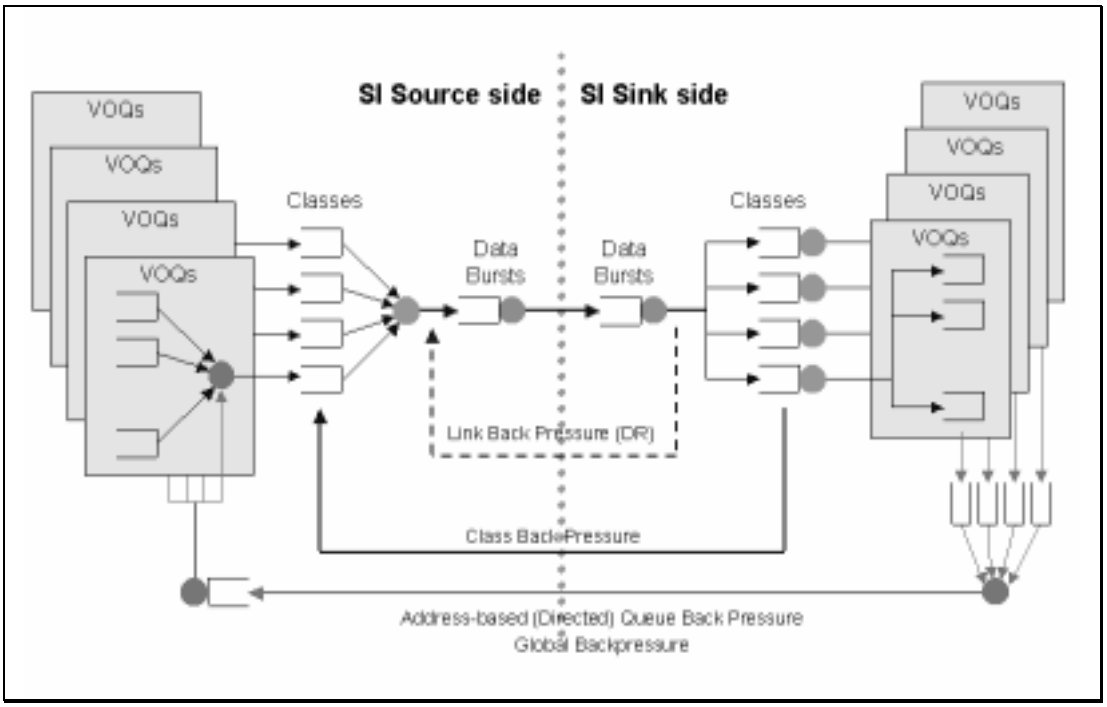
Flow control mechanisms are available for multicast traffic.

Flow control information can be aggregated into classes using the class flow control mechanism. Similarly, ingress global congestion information may be passed back to the source using a global flow mechanism.

Finally, an ingress queue map flow control mechanism provides an optional bitmap-based transfer of flow control information.

The diagram below conceptually illustrates the flow control model supported by the NPSI. Note that this is a logical view and is not meant to imply an implementation, nor does it include all of the flow control mechanisms.

Figure 36. NPE-Fabric Flow Control Reference Model



8.4.3 Link-Level Flow Control

Link-level flow control provides a single, low latency mechanism to protect the receive interface from overflow. It is controlled by a Data Ready (DR) bit from the NPSI sink to the NPSI source carried in a fixed location of every flow control message. When XOFF (DR deasserted) is recognized by the NPSI source, any segment transfer in progress shall complete and terminate with an idle control word. Idle control words and/or training sequences shall be generated as long as the XOFF condition persists.

Although XOFF takes precedence over all other flow control mechanisms discussed below, the state of these mechanisms shall be maintained and updated even when the data path interface is in the XOFF state. When XOFF reverts to XON (DR asserted is received), control returns to each of the individual mechanisms.

Link-level flow control is required in both ingress and egress directions. The default transmit state of link-level flow control shall be XOFF following reset and until updated by a flow control message (normal operation).

8.4.4 Sub-Port Flow Control

When the configured mode of NPSI segment addressing provides sub-port specification, there are three mutually exclusive modes of flow control operation, one of which shall be configured:

- Both ingress and egress sub-port fields are set to zero on transmit and ignored on reception. Flow control is exercised only on physical port granularity.
- The ingress sub-port field is set to zero on transmit and ignored on reception. The egress sub-port field specifies the actual egress sub-port. Flow control is exercised independently for each egress sub-port.
- Both the egress and ingress sub-port fields specify the actual respective sub-port. Flow control is exercised independently for each egress and ingress sub-port.

If an NPE supports sub-port granularity addressing, it shall be capable of operating in the first mode by appropriate configuration. Support for the second and third mode is optional. If a fabric supports sub-port granularity addressing, it may support any subset of the three modes.

8.4.5 Ingress Flow Control

A switch fabric typically receives data from an NPE into queues, with a queue designated for a flow or aggregation of flows[26]. A switch fabric typically uses the information in the control word (ADF) of the data it receives to determine into which queue the data is stored.

8.4.5.1 Unicast Flow Control

Individual unicast queues are flow controlled with directed status messages. Each message includes information (from the ADF) indicating the queue for which status is reported as well as the present state of that queue.

The 20-bit ADF of the unicast flow control message is formatted the same as the Unicast Address Format. The MSB of the ADF (bit 11 of the PCW) is transmitted first in the flow control message.

State is persistent until the next directed status for that queue. Accordingly, a message should be sent whenever queue status changes. For robustness, various mechanisms are possible (such as periodic updates of queue status) to recover from lost or corrupted messages. The choice to implement such updates and/or the schedule with which they are performed is implementation specific.

Flow control is XON/XOFF, using a 4-bit code point. Two of the sixteen code values are used to indicate the XON and XOFF states, and the remaining 14 code values are reserved. Unlike link level flow control, there is no specified bound on the required response latency in the data source from the reception of a unicast flow control message. (However, an excessive response latency of the source may adversely affect the performance of the system. An overly aggressive response time requirement may limit the throughput of the NPE.)

The default state of all unicast queues in the transmitting device is XON after reset (until updated by a flow control message).

All device interfaces operating in NPE-Fabric mode shall support unicast flow control using the Unicast Flow Control message format.

8.4.5.2 Multicast Flow Control

Multicast queues are also flow controlled with directed status messages. Each message includes information indicating the queue for which status is reported as well as the present state of that queue.

Multicast flow control contains the multicast class value of the multicast queue and an ingress sub-port. If sub-port addressing granularity is not configured (or not supported), the sub-port field shall be set to zero on transmit and ignored on reception. State is persistent until the next directed status for that queue. Accordingly, a message should be sent whenever queue status changes. The choice to update the queue status more often than required is optional and implementation specific.

Flow control is XON/XOFF, using a 4-bit code point. Two of the sixteen code values are used to indicate the XON and XOFF states, and the remaining 14 code values are reserved. There is no specified bound on the required response latency in the data source from the reception of a multicast flow control message.

The default state of all multicast queues in the transmitting device is XON after reset and until updated by a flow control message.

If a device supports multicast (either via multicast ID or multicast bitmap), it shall support multicast flow control using the Multicast Flow Control message format.

8.4.5.3 Class-Based Flow Control

Class-based flow control is an optional mechanism for backpressuring an entire class (or multiple classes) of traffic with a single flow control message. When implemented, each class maintains a persistent XON/XOFF state. The state defaults to XON unless updated by a class flow control

[26] On the ingress of a fabric, these may be Output Queues or Virtual Output Queues, for example.

message. This message is structured to allow the state of 16 consecutive classes to be updated at once. Since the state is persistent, a message should be sent whenever class flow control status changes. For robustness, various mechanisms are possible (such as periodic updates of class status) to recover from lost or corrupted messages. The choice to implement such updates and/or the schedule with which they are performed is implementation specific.

The Class-Based Flow Control message contains an ingress sub-port. If ingress sub-port flow control granularity is not configured (or not supported), the sub-port field shall be set to zero on transmit and ignored on reception.

A class XOFF state overrides the flow control state of all unicast queues belonging to that class, though unicast flow control status shall be maintained and updated even when the class is in the XOFF state. When a class is returned to the XON state, flow control reverts to the individual unicast queue states.

Class-based flow control is optional. If an NPSI device supports class-based flow control, it shall support the Ingress Class Flow Control message format.

There are two modes of Class Flow Control. The first mode provides support for flow control of unicast traffic only. (Flow control of multicast traffic is possible using the Multicast Flow Control message.) The second mode maps the unicast class and the multicast class with the same value to the same message bit. If an NPSI device supports class-based flow control, support must be provided for the first mode and support for the second mode is optional[27].

8.4.5.4 Global Flow Control

Global flow control is an optional mechanism on the NPE-fabric interface that allows the fabric to announce the free space in its global queue (e.g., VOQ) memory. In shared memory fabrics where resources can be dynamically allocated, this allows the NPE to determine how to best use remaining memory space.

When implemented, a global flow control message includes an N-bit code point that corresponds to a thresholded region within the fabric. By way of configuration in both the source and sink devices, this message conveys the total number of segments (of MAX_SEGMENT_SIZE) the fabric can accept before it asserts link level (DR) backpressure. A lower code point corresponds to a lower amount of available memory space. This number is persistent until the next global flow control message, so a message should be generated whenever state changes.

The reaction to Global Flow Control is implementation-specific. For example, an NPE may adjust its total transmission rate or its bandwidth distribution among classes based on the congestion level advertised by the fabric.

Global flow control is optional. If a fabric is capable of generating global flow control messages, it shall also be capable of generating class based flow control messages, and it shall provide a mechanism for enabling by configuration the generation of one or both of these flow control message types. If an NPE is capable of responding to one or both of these flow control messages, it shall provide a mechanism by configuration for responding to the enabled flow control message type(s).

8.4.5.5 Queue Map Flow Control

Queue map flow control is an optional flow control mechanism.

If a device supports Queue Map Flow Control, it shall provide a mechanism for enabling or disabling the Queue Map Flow Control. When Queue Map Flow Control is used, the other flow control message formats shall not be used.

If the fabric is configured such that it is using only the queue map message formats and mechanisms, the fabric shall continuously refresh the queue states by cycling through the unicast queue groups in order, followed by the multicast queue map, sending idle messages only when it can not transmit flow

[27] This is optional because it is not required that a given class value for unicast traffic correlate to the same class value of multicast traffic. If unicast traffic of a given class value is given different treatment through the fabric than multicast traffic of the same class value, then this second mapping option is not possible.

control information. Only flow control messages corresponding to existing queues need to be sent. Each bit in the unicast queue map field represents the XON/XOFF status of a unicast queue in a particular queue group, while each bit in the Multicast queue map field represents the XON/XOFF status of a multicast class.

If Queue Map Flow Control is used to implement egress subport granularity flow control, the index within the queue map shall be calculated by the following equation:

- Bitmap Index = (Class * MaxSubPort * MaxPhysicalPort) + Physical Port ID * MaxSubPort + Egress Sub-Port ID

Otherwise

- Bitmap Index = (Class * MaxPhysicalPort) + Physical Port ID

The product of MaxClass and MaxPorts shall be 4096 or less. MaxSubPort, MaxPhysicalPort, and MaxClass are configuration parameters. MaxPort shall be 2^n. MaxClass shall be 2^n, where $n \varepsilon \{0,8\}$, inclusive. MaxSubPort shall be 2, 4, 8, or 16.

8.4.6 Egress Flow Control

8.4.6.1 Class-Based Flow Control

Egress class-based flow control is an optional mechanism for providing backpressure from an egress NPE to an egress fabric for all traffic of an entire class (or multiple classes) with a single flow control message. When implemented, each class maintains a persistent XON/XOFF state. (The state defaults to XON unless updated by a class flow control message.) This message is structured to allow the state of 16 consecutive classes to be updated at once. Since the state is persistent, a message should be sent whenever class flow control status changes. For robustness, various mechanisms are possible (such as periodic updates of class status) to recover from lost or corrupted messages. The choice to implement such updates and/or the schedule with which they are performed is implementation specific.

The Class-Based Flow Control message contains an egress sub-port. If egress sub-port flow control granularity is not configured (or not supported), the sub-port shall be set to zero on transmit and ignored on reception.

Egress class-based flow control is optional. If an NPSI device supports egress class-based flow control, it shall support the Egress Class Flow Control message format.

There are two modes of Class Flow Control. The first mode maps the unicast class and the multicast class with the same value to the same message bit. The second mode provides support for flow control of unicast and multicast classes independently. Flow control messages for unicast use a value for M (bit 24) of zero, while those for multicast use a value for M of one. In the first mode, M is set to zero on transmit and ignored on receive. If an NPSI device supports class-based flow control, support must be provided for the first mode and support for the second mode is optional.

8.4.7 Flow Control Message Format

All flow control messages are 27 bits in length. Each message contains a dedicated bit for link level flow control (DR) in order to minimize latency. Based on the differing flow control requirements at the ingress and egress NPE-Fabric interfaces, two sets of messages are defined.

8.4.7.1 Summary of Flow Control Message Formats

Ingress Formats

The following table shows the ingress flow control message formats.

Table 12. Ingress Flow Control Message Summary

	6	5	4	3	2	1	0	9	8	7	6	5	4	3	2	1	0	9	8	7	6	5	4	3	2	1	0
Unicast	d	0	0	Address Data Field [20b]																				Status [4b]			
Class	d	0	1	Ing Sb-Pt [4b]				Cls_Grp [4b]				Class_Map [16b]															
Multicast	d	1	0	0	*reserved*							Ing Sb-Pt [4b]		Multicast Class [8b]										Status [4b]			
Global	d	1	0	1	*reserved*															Global Status [8b]							
Idle	d	1	1	*reserved*																							

The following table shows the ingress flow control message formats when a device is configured for the Queue Map Option. They are mutually exclusive with the above formats.

Table 13. Ingress Flow Control Message Summary (Queue Map Option)

	6	5	4	3	2	1	0	9	8	7	6	5	4	3	2	1	0	9	8	7	6	5	4	3	2	1	0
Unicast map	d	0	0	Q_Grp [8b]								Unicast Queue_Map [16b]															
Multicast map	d	0	1	*reserved*				Q_Grp [4b]				Multicast Queue_Map [16b]															
Reserved	d	1	0	*reserved*																							
Idle	d	1	1	*reserved*																							

Egress Formats

The following table shows the egress flow control message formats.

Table 14. Egress Flow Control Message Summary

	6	5	4	3	2	1	0	9	8	7	6	5	4	3	2	1	0	9	8	7	6	5	4	3	2	1	0
Class	d	0	M	Eg Sb-Pt [4]				Cls_Grp [4b]				Class_Map [16b]															
Reserved	d	1	0	*reserved*																							
Idle	d	1	1	*reserved*																							

8.4.7.2 Ingress Formats

The following table shows the format of the Ingress Unicast Flow Control message. Its usage is described in Section 8.4.5.1.

Table 15. Ingress Unicast Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready (DR) Indicator This bit indicates if the fabric is ready to receive data.[28] d = 0 indicates that the fabric is not ready to receive data. d = 1 indicates that the fabric is ready to receive data.
25:24	type	The type code for a unicast queue flow control message is 00.
23:4	ADF[19:0]	The used bits of the ADF correspond to the unicast queue for which this status message applies. If the fabric only supports a subset of the queues addressable by this field, any unused bits shall be set to 0.
3:0	Status[3:0]	This field indicates the ability for the target queue to receive data. 1111 indicates ready (XON). 0000 indicates not ready (XOFF). All other code values are reserved.

The following table shows the format of the Ingress Multicast Flow Control message. Its usage is described in Section 8.4.5.2.

Table 16. Ingress Multicast Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready (DR) Indicator This bit indicates if the fabric is ready to receive data. d = 0 indicates that the fabric is not ready to receive data. d = 1 indicates that the fabric is ready to receive data.
25:23	type	The type code for a multicast queue flow control message is 100.
22:16	rsvd	Reserved. Set to 0.
15:12	Ing Sb-Pt	Ingress Sub-Port If Sub-Port granularity addressing is not configured (or not supported), this field is set to zero on transmit and ignored on receive.
11:4	Multicast Class[7:0]	This field indicates the multicast queue for which this status message applies. It is derived from the class field of the Ingress Multicast ADF. A multicast queue is designated for each class. If the fabric only supports a subset of the queues addressable by this field, any unused bits shall be set to 0.
3:0	Status[3:0]	This field indicates the ability of the target queue to receive data. 1111 indicates ready (XON). 0000 indicates not ready (XOFF). All other code values are reserved.

[28] For those familiar with CSIX, this bit is bridgeable to DRDY.

The following table shows the format of the Ingress Class Flow Control message. Its usage is described in Section 8.4.5.3.

Table 17. Ingress Class Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready (DR) Indicator This bit indicates if the fabric is ready to receive data. d = 0 indicates that the fabric is not ready to receive data. d = 1 indicates that the fabric is ready to receive data.
25:24	type	The type code for an ingress class flow control message is 01.
23:20	Ing Sub-Port	Ingress Sub-Port If Sub-Port granularity addressing is not configured (or not supported), this field is set to zero on transmit and ignored on receive.
19:16	Cls_Grp[3:0]	This field indicates the group of 16 classes covered by this message. The range is from {Cls_Grp [3:0], 0b1111} to {Cls_Grp [3:0], 0b0000}.
15:0	class_map[15:0]	This bit map indicates the XON/XOFF state for the 16 classes indicated by Cls_Grp. Bit 15 corresponds to {Cls_Grp [3:0], 0b1111}. Bit 0 corresponds to {Cls_Grp [3:0], 0b0000}. 1 indicates XON. 0 indicates XOFF. Note that support for class based flow control is optional.

The following table shows the format of the Ingress Global Flow Control message. Its usage is described in Section 8.4.5.4.

Table 18. Ingress Global Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready (DR) Indicator This bit indicates if the fabric is ready to receive data. d = 0 indicates that the fabric is not ready to receive data. d = 1 indicates that the fabric is ready to receive data.
25:23	type	The type code for a global flow control message is 101.
22:8	rsvd	Reserved. Set to 0.
7:0	global_status[7:0]	This field indicates the number of segments a switch fabric can accept before it asserts link level flow control. It carries a unit-less code mapped to identical quantities on both the sending NPE and the receiving fabric. 11111111 = all segments in shared memory are available. 00000000 = no segments are available. Support for global flow control messages is optional. If it's supported, any number of bits from 1 to 8 may be used, though, for interoperability, monotonic increases in space must result in monotonic code increases. For the same reason, active bits must be aligned to the most significant positions and unused bits must be aligned to the least significant positions and set to 0.

The following table shows the format of the Ingress Unicast Queue Map Flow Control message. Its usage is described in Section 8.4.5.5.

Table 19. Ingress Unicast Queue Map Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready (DR) Indicator This bit indicates if the fabric is ready to receive data. d = 0 indicates that the fabric is not ready to receive data. d = 1 indicates that the fabric is ready to receive data.
25:24	type	The type code for a unicast queue map flow control message is 00.
23:16	Q_Grp[7:0]	This field indicates the group of 16 queues covered by this message[29]. The range is from {Q_Grp [7:0], 0b1111} to {Q_Grp [7:0], 0b0000}.
15:0	Unicast queue_map [15:0]	This bit map indicates the XON/XOFF state for the 16 queues indicated by Q_Grp. Bit 15 corresponds to {Q_Grp [7:0], 0b1111}. Bit 0 corresponds to {Q_Grp [7:0], 0b0000}. 1 indicates XON. 0 indicates XOFF. Note that support for unicast queue map flow control is optional.

The following table shows the format of the Ingress Multicast Queue Map Flow Control message. Its usage is described in Section 8.4.5.5.

Table 20. Ingress Multicast Queue Map Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready (DR) Indicator This bit indicates if the fabric is ready to receive data. d = 0 indicates that the fabric is not ready to receive data. d = 1 indicates that the fabric is ready to receive data.
25:24	type	The type code for a multicast queue map flow control message is 01.
23:20	rsvd	Reserved. Set to 0.
19:16	Q_Grp[3:0]	This field indicates the group of 16 queues covered by this message. The range is from {Q_Grp [3:0], 0b1111} to {Q_Grp [3:0], 0b0000}.
15:0	Mcast queue_map[15:0]	This bit map indicates the XON/XOFF state for the 16 multicast queues indicated by Q_Grp. Bit 15 corresponds to {Q_Grp [3:0], 0b1111}. Bit 0 corresponds to {Q_Grp [3:0], 0b0000}. A multicast queue is designated per class. 1 indicates XON. 0 indicates XOFF. Note that support for multicast queue map flow control is optional.

The following table shows the format of the Ingress Idle Flow Control message. The Idle flow control message is sent whenever the status channel is not in training (or sending an LODS alarm) and no

[29] Computing the Q_Group and Index within the Ucast_queue_map for a unicast queue designated by {Port, Class}

MaxClass	Q_Group	Index for queue within Ucast_queue_map
256	Class[7:0]	Port mod 16
128	Class[6:0],Port[4]	Port mod 16
64	Class[5:0],Port[5:4]	Port mod 16
32	Class[4:0],Port[6:4]	Port mod 16
16	Class[3:0],Port[7:4]	Port mod 16
8	Class[2:0],Port[8:4]	Port mod 16
4	Class[1:0],Port[9:4]	Port mod 16
2	Class[0],Port[10:4]	Port mod 16
1	Port[11:4]	Port mod 16

flow control messages are available to send (or when the transmission of flow control messages is backpressured via the SNR bit).

Table 21. Ingress Idle Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready (DR) Indicator This bit indicates if the fabric is ready to receive data. d = 0 indicates that the fabric is not ready to receive data. d = 1 indicates that the fabric is ready to receive data.
25:24	type	The type code for an idle flow control message is 11.
23:0	rsvd	Reserved. Set to 0.

8.4.7.3 Egress Formats

The following table shows the format of the Egress Class Flow Control message. Its usage is described in Section 8.4.6.1.

Table 22. Egress Class Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready (DR) Indicator This bit indicates if the NPE is ready to receive data. d = 0 indicates that the NPE is not ready to receive data. d = 1 indicates that the NPE is ready to receive data.
25	type	The type code for an egress class flow control message is 0.
24	M	Multicast If the interface is configured to specify flow control separately for multicast classes and unicast classes, this bit is set to zero for unicast classes and set to one for multicast classes. If the interface is configured to map equal class values for multicast and unicast to the same bit in the message, M shall be set to zero on transmit and ignored on receive.
23:20	Eg Sb-Pt	Egress Sub-Port If Sub-Port granularity addressing is not configured, this field is set to zero on transmit and ignored on receive.
19:16	Cls_Grp[3:0]	This field indicates the group of 16 classes covered by this message. The range is from {Cls_Grp [3:0], 0b1111} to {Cls_Grp [3:0], 0b0000}.
15:0	class_map[15:0]	This bit map indicates the XON/XOFF state for the 16 classes indicated by Cls_Grp. Bit 15 corresponds to {Cls_Grp [3:0], 0b1111}. Bit 0 corresponds to {Cls_Grp [3:0], 0b0000}. 1 indicates XON. 0 indicates XOFF. Note that support for class based flow control is optional.

The following table shows the format of the Egress Idle Flow Control message.

Table 23. Egress Idle Flow Control Message

Bits	Mnemonic	Description
26	d	Data Ready Indicator This bit indicates if the NPE is ready to receive data. d = 0 indicates that the NPE is not ready to receive data. d = 1 indicates that the NPE is ready to receive data.
25:24	type	The type code for an idle flow control message is 11.
23:0	rsvd	Reserved. Set to 0.

8.4.8 Flow Control Response Requirements

After power-up, the data source shall initialize its internal DR state to DR de-asserted (XOFF). All other flow control states shall initialize as XON or Ready.

8.4.8.1 Link Level Flow Control

A data source shall respond to the reception of a deasserted DR bit according to the following rule. A data source should respond to the reception of an asserted DR bit with a similar response.

From the clock tick that a DIP-2 code of a directed status message with the DR bit de-asserted is received by the data transmitting device of an NPSI instance, after a maximum response of (n*T) has elapsed, no data segment burst shall be transmitting from the data transmitting component.

From the tick that the DIP-2 code is received by the data transmitting device, the maximum response time for the suspension of data transfers is defined as: n*T, *where* n=C+M;

* T is the clock tick period of half a clock cycle (a single word time on the data path)

* n is the maximum number of ticks for the response

* C is a constant for propagating the DR bit within the data transmitting device (or chipset as the case may be) to the interface logic controlling the data transmission. C is defined to be 192 clock ticks. (After C clock ticks, no new data transfer should start.)

* M is the number of ticks required to transport the maximum size segment. (For example, for a 64-byte MAX_SEGMENT_SIZE, M is 34 for unicast segments.)

The DIP-2 code of a flow control message shall be checked before the DR bit is interpreted. When a DIP-2 error is detected on a flow control message, the DR bit from the flow control message shall be ignored and the device shall interpret the DR bit as not ready. This means that the device that detects the parity error shall stop sending further data on the data path according to the required response latency for receiving a valid, de-asserted DR bit, until the device receives a valid DIP-2 code and the following valid flow control message without any errors and corrects the state of the DR bit.

When the data source detects transmission of a training sequence on its status channel, it should interpret this immediately as the reception of a de-asserted DR bit.It should then stop transmission of data within the same required response latency as if a de-asserted DR bit had been received (substituting the reception of the DIP-2 code with a single full training sequence). Once the data source detects the termination of a training sequence, it should wait until the first valid flow control message is received (qualified by the DIP-2 code) and react according to the DR bit of that message.

As described in section7.3, when a framing error is detected on the status interface, the device shall react as if a de-asserted DR were received. This means that the device that detects the framing error shall stop sending further data on its data path according to the required response latency for receiving a valid, de-asserted DR bit, starting from the time the third consecutive framing code is received. The internal DR state shall remain de-asserted until the device receives a message with a valid DIP-2 code, and updates the state of the DR bit.

8.4.8.2 Class Flow Control

In order for class-based flow control to be effective, an NPE sending data to a fabric should be able to react to class flow control messages quickly. For example, it should be able to react to class-based flow control messages faster than if the information were sent as a series of directed flow control messages (for example, for all of the ports of a given class).

8.4.9 Flow Control of Directed Status

An optional mechanism is provided for flow control of directed status. The SNR bit reflects the ability of the data transmitter side of the interface to accept flow control messages (i.e., any message other than Idle messages) on its corresponding status interface. A value of ‘0’ for the SNR bit indicates the ability to accept all flow control messages, while a value of 1 for the SNR bit indicates the ability to only accept Idle messages.

Bit 11 of the Idle Control Word is used as the Status Not Ready (SNR) bit, as shown in Figure 15. It is de-asserted to indicate Ready (XON) and asserted to indicate Not Ready (XOFF). For NPE-Framer and NPE-NPE modes of the NPSI, the bit shall be zero in normal operation. The address field of the Training Control Word does not contain a Not Ready bit.

The time allowed for responding to transitions in the state of the SNR bit is identical to that of Link Level flow control, described above.

An NPSI device operating in NPE-Fabric Mode may support the use of the SNR. If an NPSI device supports the SNR bit, it shall provide a mechanism to configure the device to enable it.

If Directed Status Flow Control is enabled, the data sink should, upon receiving a training sequence on the data interface, act as if a Status Not Ready bit had been received. The data source should always follow the transmission of a training sequence with the transmission of an Idle Control Word to provide an updated SNR bit.

The Streaming Interface does not specify any frequency for transmitting Idle Control Words to update the SNR bit state.

When the data source has asserted the SNR bit, it shall still react to the DR bit of the idle flow control messages.

8.5 Summary of Start-up Parameters

Several interface parameters need to be initialized after reset and before normal operation of the interface begins.

The DATA_ALPHA and STAT_ALPHA (if applicable) required by a receiver shall be initialized in the corresponding transmitter. Likewise, DATA_MAX_T and STAT_MAX_T (if applicable) shall be initialized in the transmitter.

A summary of the start-up parameters for the NPE-Fabric mode is listed in Table 24, below.

Table 24. Summary of Start-up Parameters.

Parameter	Definition	P	Units	Min	Max
DATA_ALPHA	Number of repetitions of the data training sequence that must be scheduled every DATA_MAX_T cycles.	✔	repetitions	0	255
DATA_MAX_T	Maximum interval between scheduling of training sequences on Data Path interface.	✔	2^8 clock cycles		2^{32}-1
STAT_ALPHA	Number of repetitions of the status training sequence that must be scheduled every STAT_MAX_T cycles.	✔	repetitions	0	255
STAT_MAX_T	Maximum interval between scheduling of training sequences on Status Path interface.	✔	2^8 clock cycles		2^{32}-1
MAX_SEGMENT_SIZE	Length of non-EOP segments	✔	Bytes		

P – Provisionable

9. NPE-NPE Mode

9.1 Functional Description

The NPE-NPE mode of operation of the NPSI can be used to support connectivity between two adjacent network processing elements. An NPE may be a network processor, coprocessor, or traffic manager.

The LVDS electrical specifications from Section 10 “Physical Layer” shall be used for the NPE-NPE data and status paths.

The NPE-NPE Mode uses data framing and flow control mechanisms consistent with the SPI-4 Phase 2 [1] and SPI-5 [2] implementation agreements. However, this document fully specifies the operation of the NPE-NPE Mode of the NPSI.

The interface supports a flat addressing space of up to 256 ports (channels) for connecting NPEs in its basic mode, and channelization using up to 148 bits of addressing (8 ADW’s) in its extended mode. The uses of ports for multiplexing of data from multiple packets across the interface are defined by the application and not by the NPSI. The support of an 8-bit address space is mandatory; extended addressing is optional.

Data is transferred in segments as described in Section 7.1.

The NPSI uses the SPI-5 Pool Status Mechanism, which allows the flow control granularity to be different from the full data multiplexing capability. Using this mechanism, backpressure for multiple ports can be aggregated to create “pool” flow control. The TDM calendar function specified in the SPI-4 Phase 2 implementation agreement, however, is a proper subset of this functionality.

Support for an optional narrow bus mode of operation is provided.

9.2 Data Path Operation

Refer to Section 7.1, “Common Data Path Operation” for a description of the fundamentals of the data path operation.

An NPE sending data to or receiving data from another NPE shall support a MAX_SEGMENT_SIZE of 64 bytes. Support for any other MAX_SEGMENT_SIZE is optional. An NPE may support multiple values of MAX_SEGMENT_SIZE; if an NPE supports more than one value of MAX_SEGMENT_SIZE, it shall provide a mechanism for configuring this value before normal operation of the interface begins.

Consistent with SPI-4 Phase 2, successive Start of Packets must occur not less than 8 Wcycles apart, where a Wcycle is one 16-bit control or data word.

Narrow mode operation, as described in the appendix of SPI-4 Phase 2, is optional[30].

The NPE-NPE data path uses the training operation defined in Section 7.1.5, “Training Sequence for Data Path .”

9.2.1 Data Framing

The NPE-NPE mode supports both the Basic and Extended Address data burst formats defined in Section 7, “Common Functions for the NPE-NPE and NPE-Fabric Mode.”

In the NPE-NPE Mode, there are three possible data burst formats, depending on the control sequence, as described in Section 7.1, “Common Data Path Operation:”

- Basic Data Burst
- Extended Address Data Burst without ADW
- Extended Address Data Burst with ADW

[30] Narrow mode operation is not specified here; refer to the SPI-4 Phase 2 specification for information on this mode.

An NPE operating in NPE-NPE Mode of the NPSI shall support the Basic data burst format and may support the optional Extended Address data burst formats. The required (Basic) format is consistent with SPI-4 Phase 2, with an 8-bit address field. If an NPSI device supports the Extended Address data burst format, it shall provide a mechanism to configure the device to use either the Basic or Extended Address data burst format, including the number of ADW's (a value from 0 through 8, inclusive, up to the number of ADW's supported) in the Extended Address format.

9.2.2 Data Transfer Procedure

The NPE-NPE Mode shall use the Basic Data Burst as described in Section 7.1.1, "Data Framing Formats." The NPE-NPE Mode may use the Extended Address Data Burst Formats (described in the same section).

9.3 Addressing

The NPE-NPE mode uses a flat address space. Other than the mapping of address pools, as defined in the next section, the NPE-NPE addresses have no specified structure. The entire ADF is used as a single Port ID. The use of the address for classes or other classifications is defined by the application and not by this specification.

- Support for 8-bit address field (Basic data burst) is required.
- Extended address support is optional.

The AT field is reserved. This field shall be transmitted as zeroes and ignored on reception (except for DIP-4 checking).

If the Extended Address format is used, the MSB of the address is sent in the PCW (little endian order).

9.4 Flow Control

The NPE-NPE mode uses the pool status mechanism and framing from SPI-5. (The TDM calendar functionality of the SPI-4 Phase 2 implementation agreement is a subset of this.) The status channel implements a credit-based flow control scheme. Each credit represents a 16-byte block of transmitted or received data.

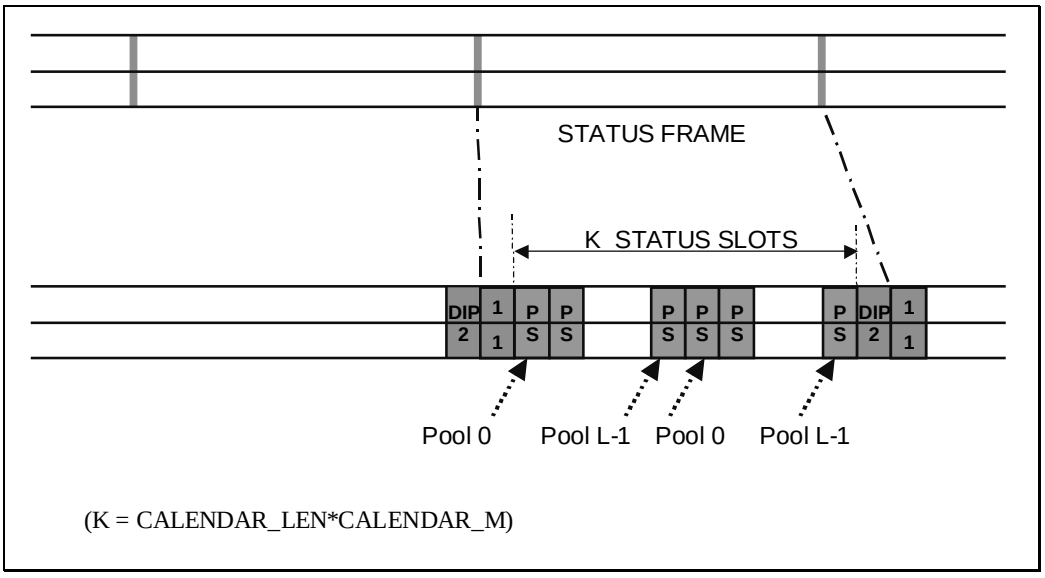
In full duplex applications (i.e., two NPSI's carrying traffic in opposite directions), each direction of the NPE-NPE bus has its own independent status channel.

9.4.1 Flow Control Message Encoding and Framing

The NPE-NPE mode uses the framing described in Section 7.2.1, "Flow Control Status Framing."

The Pool Status Channel message is organized in a logical frame structure, as shown in Figure 37. A status frame begins with the framing code. The framing code is followed by a number of pool status report words, defined by a TDM calendar of Pools Status (K status slots consisting of M repetitions of the L status reports) as described in Section 9.4.3. Each status slot is 2 bits. The frame ends with a DIP-2 odd parity checksum.

Figure 37. Pool Status Frame



The length of a frame, in bits, is 4 overhead bits plus the number of Calendar entries times the Calendar repetition count times two bits per pool status report word. The 4 bits for overhead consist of the framing code (2 bits) and the DIP-2 checksum (2 bits).

The 2-bit framing code (‘0b11’) is reserved to delimit the beginning/end of a status frame. The three other 2-bit values (’0b01’, ’0b10’ and ’0b00’) are used for the body of the status frame to indicate the fill level of each pool, and are placed on the channel in a time-division-multiplexed manner in accordance with the Calendar.

A framer in the sink device of a Pool Status Channel monitors the received pool status frame for the framing code. When the framer is in-frame, the pool status report words are used to update the credits of the Pools. The framer monitors for framing pattern errors and DIP-2 errors. Occurrence of multiple framing errors places the framer in the out-of-frame state. When the framer is in the out-of-frame state, all previously granted credits are cancelled and remain so until the framer returns to the in-frame state. The framer returns to the in-frame state when it observes multiple consecutive error-free pool status frames. Although no specific number of framing errors is specified, it is a requirement that any single error event shall not cause loss of synchronization.

Note that the pattern ’0b11’ is a valid DIP-2 codeword and is also part of the training pattern. The framer must not lock onto these occurrences as frame boundaries.

9.4.2 Credit Pools

To allow the flow control granularity to be different from the full data multiplexing capability on an NPE-to-NPE data interface, multiple ports (channels) may be collected into what is called a Pool (equivalent to SPI-5 pools). Credits are granted and consumed on a per-Pool basis and all ports in a Pool share a common bank of credits. The pool number associated with a port address (PA) shall be uniquely identified by the lower order port address bit values PA[POOL_LEN-1:0]. Applications with a small number of ports may choose to have one port per pool, giving a simple one-to-one mapping between ports and credit pools.

Credits are granted by data path sink devices based on their ability to accept data. Each Pool Status report word grants credits to the data source device, allowing it to transfer data for the corresponding Pool across the interface. Each unit of credit permits a single 16-byte block of data to be sent to one of the ports sharing the Pool.

Credits are consumed by data path source device whenever Payload Data words or Address Data words are sent. Credits are not consumed for Control words or for Training Data words. A credit is consumed for each 16-byte block of data that is transferred. A block containing fewer than 16 bytes consumes one whole credit. Thus, a partial block at the end of a packet consumes a whole credit

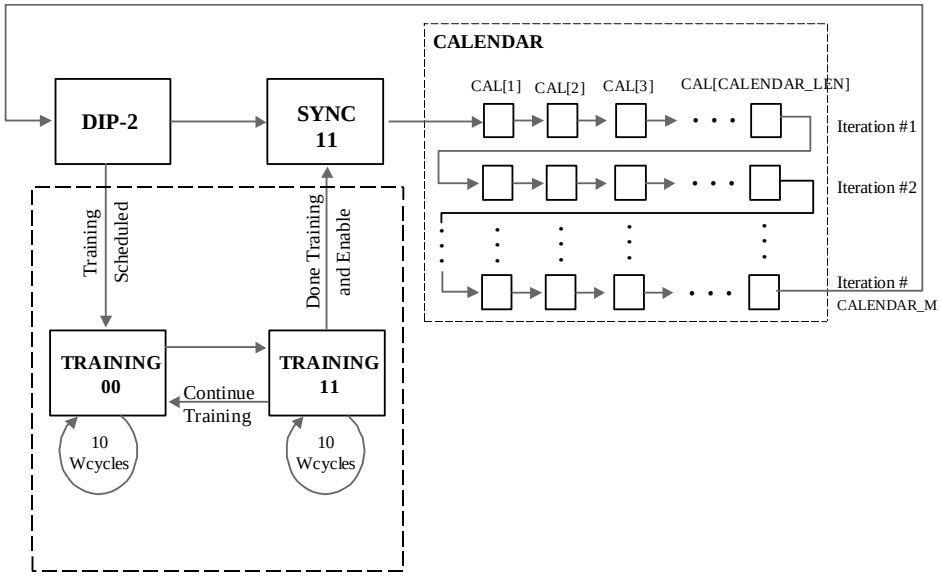
even if it is shorter than 16 bytes in length. Similarly, a sequence of Address Data words at the beginning of a transfer always consumes a whole credit regardless of its length.[31]

Credits are granted to Pools and one credit is consumed for each 16-byte block of data transferred from any port within the Pool. The source of a data interface only schedules transfers for ports whose associated Pools have sufficient credits to support the entire transfer.

9.4.3 Pool Status Calendar

The order of pool status report words in a frame is defined by a Calendar structure, as shown in Figure 38. The number of entries in the Calendar is given by the bus parameter CALENDAR_LEN. The number of times that the calendar is traversed for each frame is given by the bus parameter CALENDAR_M. Entries in the calendar are traversed sequentially.

Figure 38. Pool Status Calendar



Each Calendar entry has an associated pool address (CAL[i]), which may optionally be provisionable.[32] As the Calendar is traversed, the status of the Pool associated with each Calendar entry is placed on the Pool Status Channel. The status is encoded using one of the three possible 2-bit Pool status report word values defined in the following table.

[31] For example, a transfer with only one ADW and only one PDW consumes two credits.

[32] For example, in the case of a single port device, CAL[i] need not be provisionable.

Table 25. Pool Status Format

STAT[1]	STAT[0]	**Description**
1	1	In-band framing code The framing code is sent once per pool status frame in normal operation. This status value is also repeated as part of the training pattern or to indicate a disabled link.
1	0	SATISFIED The SATISFIED status indicates that the receiver of the corresponding sink Pool is almost full. When the SATISFIED status is received, no further credits are granted. Credits granted previously by HUNGRY or STARVING status reports remain available.
0	1	HUNGRY The HUNGRY status indicates that the receiver of the corresponding sink Pool is partially empty. When a HUNGRY status is received, the amount of credits at the source Pool is increased to MAXBURST2 credits if the current value is less than MAXBURST2. If the amount of credit remaining is currently greater than MAXBURST2, due to a previous STARVING status report, the credit count is left unchanged.
0	0	STARVING The STARVING status indicates that the FIFO of the corresponding sink Pool is almost empty. When a STARVING status is received, the amount of credits at the source Pool is set to MAXBURST1 credits. This status value is also repeated as part of the training pattern.

The indicated FIFO status is based on the latest available information. A STARVING indication provides additional feedback information, so that transfers can be scheduled accordingly. Applications that do not need to distinguish between HUNGRY and STARVING may only examine the most significant FIFO status bit.

Note that a pool address may appear more than once in a calendar. This is helpful in situations where the pools have unequal bandwidths. For example, a high bandwidth pool can have a proportionally greater share of entries in the calendar to allow more frequent status updates.

Note that if a calendar entry is unpopulated or unprovisioned it should be sent with STAT='0b10' ("Satisfied").

It is a requirement that the source and sink devices of a given interface have identical Calendar configurations in order to operate properly.

9.4.4 Transmission

The 2-bit Pool status words from the Pool Status Frame are transmitted on STAT.

An error monitor in the sink device continuously verifies the DIP-2 code words received. After multiple DIP-2 errors, the monitor enters the out-of-frame state and all credits previously granted to Pools are cancelled and remain cancelled. After multiple consecutive error-free codewords are received, the monitor returns to the in-frame state and credits can be accumulated. Although no specific number of DIP-2 errors is specified, it is a requirement that any single error event shall not cause a transition to the out-of-frame state.

Note that a receiver should not wait for the DIP-2 validation of the contents of a status frame before acting on the status reports contained in the status frame.

The product of CALENDAR_LEN and CALENDAR_M (CALENDAR_LEN * CALENDAR_M) must be greater than or equal to sixteen to be able to distinguish between status information and the training sequence.

9.5 Summary of Start-up Parameters

Several interface parameters need to be initialized after reset and before normal operation of the interface begins.

The DATA_ALPHA and STAT_ALPHA (if applicable) required by a receiver shall be initialized in the corresponding transmitter. Likewise, DATA_MAX_T and STAT_MAX_T (if applicable) shall be initialized in the transmitter.

The length of the calendar and the number of repetitions of the calendar, CALENDAR_LEN and CALENDAR_M, shall be initialized in both the receiver and the transmitter.

The number of bits of the port address that are used to determine the credit pool shall be initialized in the transmitter and the receiver. The number of credits (16-byte blocks) for MaxBurst1 and (if applicable) MaxBurst2 shall be configured in the transmitter per credit pool (channel).

A summary of the start-up parameters for the NPE-NPE mode is listed in Table 26, below.

Table 26. Summary of Start-up Parameters.

Parameter	Definition	P	CH	Units	Min	Max
CALENDAR[i]	Pool address at calendar location i.	✔	I	(N/A)		
CALENDAR_LEN	Length of the calendar sequence.	✔	I	(N/A)		
CALENDAR_M	Number of times calendar sequence is repeated between insertions of framing pattern.	✔	I	(N/A)		
MAX_CALENDAR_LEN	Maximum supported value of CALENDAR_LEN		I	(N/A)		
MaxBurst1	Maximum number of 16 byte blocks that the receiver can accept when Status channel indicates Starving.	✔	C / I	16 byte blocks		
MaxBurst2	Maximum number of 16 byte blocks that the receiver can accept when Status channel indicates Hungry. MaxBurst2 <= MaxBurst1	✔	C / I	16 byte blocks		
DATA_ALPHA	Number of repetitions of the data training sequence that must be scheduled every DATA_MAX_T cycles.	✔	I	(N/A)	0	255
DATA_MAX_T	Maximum interval between scheduling of training sequences on Data Path interface.	✔	I	2^8 clock cycles		2^{32}-1
STAT_ALPHA	Number of repetitions of the status training sequence that must be scheduled every STAT_MAX_T cycles.	✔	I	(N/A)	0	255
STAT_MAX_T	Maximum interval between scheduling of training sequences on Status Path interface.	✔	I	2^8 clock cycles		2^{32}-1
POOL_LEN	The number of bits of the port address that are used to uniquely decode the associated credit pool.	✔	I	bits		
MAX_SEGMENT_SIZE	Length of non-EOP segments	✔	I	Bytes		

P – Provisionable

CH – Per channel (C) or per interface (I)

10. Physical Layer

A block diagram depicting the interface signals is shown in Section 5.4, “Interface Signals.”

In the electrical specification tables below (Table 27, “Data and Status Path DC Specifications,” and Table 28, “Data and Status Path AC Specifications”), the values listed in the tables take precedence over [4] for NPSI devices. Parameters that are not specified here are the same as [4].

10.1 Data and Status Path DC Specifications

Table 27. Data and Status Path DC Specifications

Symbol	Parameter	Conditions	Min	Max	Units
Voh	Output voltage high	Differential load R_{load}=100Ω		1475	mV
Vol	Output voltage low	Differential load R_{load}=100Ω	235		mV
$\|V_{OD}\|$	Output Differential voltage	Differential load R_{load}=100Ω	250	600	mVpkd
V_{OS}	Output Common Mode Voltage	Differential load R_{load}=100Ω	540	1275	mV
R_O	Differential Output Impedance		80	120	Ω
R_I	Differential Input Impedance		90	110	Ω
Vi	Input Voltage Range	$\|V_{gpd}\|$ < 50mV	160	1625	mV
V_{hyst}	Input differential hysteresis		Note 3	Note 3	
$\|V_{ID}\|$	Input Differential Voltage	$\|V_{gpd}\|$ < 50mV, Note 2	80	600	mVpkd
C_{in}	Input Capacitance			10	pF

Notes:
1. Values are measured with each LVDS output DC-coupled into a 50-Ohm impedance (100 Ohms differential impedance).
2. The minimum of 80 mVpkd (millivolt peak differential) is only required for devices supporting operation at or above 900 Mbps operation. Otherwise, the minimum of 100 mVpkd (from [4]) is required.
3. The EIA/TIA 644 definition of V_{hyst} does not apply to the NPSI.

System-level reference points for specified parameters in this section are shown in Figure 39.

10.2 Data and Status Path AC Specifications

An NPSI-conforming implementation should specify its maximum data rate. This data rate must be at least 622 Mbps (i.e., 311 MHz DDR clock). The device shall operate over the entire frequency range between 622 Mbps and its maximum rate.

The timing parameters are specified to support two different bit alignment schemes at the receiver. Some of the parameters in Table 28 below correspond to the case of "static alignment", in which the receiver latches data at a fixed point in time relative to clock (requiring a more precisely specified sampling window). Some of the parameters in Table 28 below correspond to the case of "dynamic alignment", in which the receiver has the capability of centering the data and control bits relative to clock. From an AC timing perspective, a compliant receiver only needs to meet the parameters at the data path for either static or dynamic alignment, but may also comply to both sets of parameters. A compliant driver shall meet both timing specifications to be interoperable with both types of receivers.

For operation above 900 Mbps (450 MHz), the receiver should implement dynamic alignment. All devices shall be capable of generating the training sequence.

The following electrical specifications are qualified over the operating range of 622 Mbps to 1.3 Gbps (311 MHz to 650 MHz DDR). However, the parameters are defined to allow operation above 1.3 Gbps (650 MHz DDR).

Table 28. Data and Status Path AC Specifications

Symbol	Parameter	Conditions	Min	Max	Units
fD	DCLK frequency		311		MHz
Dcoc	Output Clock Duty Cycle		45	55	%
tskew2[33]	Output Clock lane to any Data lane Skew	Specified to support alignment		0.2 (222 ps)	UI
tskew1	Output Differential Skew	Below 1.3 Gbps		50	ps
		At or above 1.3 Gbps		0.065	UI
Jclk (Note 9)	Output clock peak-to-peak jitter (Note 8)[34]	Specified to support alignment		0.15	UI
Jdata1 (Note 11)	Output data jitter relative to the clock edge from which the data was transmitted (Note 8)	Specified to support alignment		0.14	UI
Jdata2 (Note 12)	Output data jitter relative to the opposite clock edge from which the data was transmitted (Note 8)	Specified to support alignment		0.29	UI
tR, tF	Output 20%-80% Rise & Fall Times	Below 1.3 Gbps	75	 0.3	ps UI
		At or above 1.3 Gbps	0.1	0.3	UI
Dcic	Input Clock Duty Cycle		40	60	%
tskew3	Worst-case cumulative skew and jitter any Data lane	Not applicable for dynamic alignment		0.55 (614ps)	UI
tR, tF	Input 20%-80% Rise & Fall Times	Below 1.3 Gbps	75	 0.36	ps UI
		At or above 1.3 Gbps	0.1	0.36	UI

Notes:
1. Rise and fall times assume nominal 100-Ohm differential termination and exclude reflections.
2. The Unit Interval (UI) is the reciprocal of the symbol rate for both clock and data.
3. All timing parameters are measured relative to the differential crossing point of the corresponding clock signal.
4. Total jitter includes both deterministic jitter and random jitter.

[33] Tskew2 does not have to include substrate offsets or other deterministic skew components as long as the offsets are documented and compensated for by the board layout.

[34] Jclk is peak-to-peak consistent with SPI-4 Phase 2 [1].

5. The transmitter shall conform to all output AC specifications, supporting both static and dynamic alignment
6. Values are measured with each LVDS output DC coupled into a 50-Ohm impedance (100 Ohms differential impedance).
7. Jitter and skew are specified between crossings of the 50% threshold of the reference signal.
8. Only frequency components greater than the data_rate/2000 shall be included in the jitter measurement.
9. Jclk is calculated from the rising edge to the rising edge and falling edge to falling edge across an appropriate interval. Jitter of the rising edge is considered independent of the jitter of the falling edge.
11. Jdata1 does not include cancelled, correlated jitter shared between clock and data. Static skew between data and clock are treated as cancelled by dynamic alignment.
12. Jdata2 makes no assumptions about cancelled, correlated jitter between clock and data. Static skew between data and clock are treated as cancelled by dynamic alignment.

Corresponding reference points are shown in Figure 39 and Figure 40.

Figure 39. Reference Points for Electrical Specification Parameters

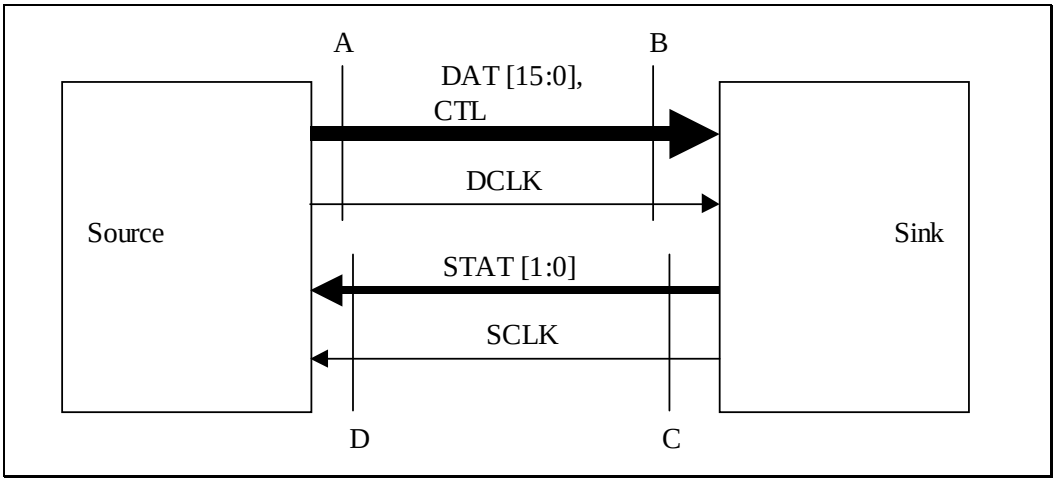
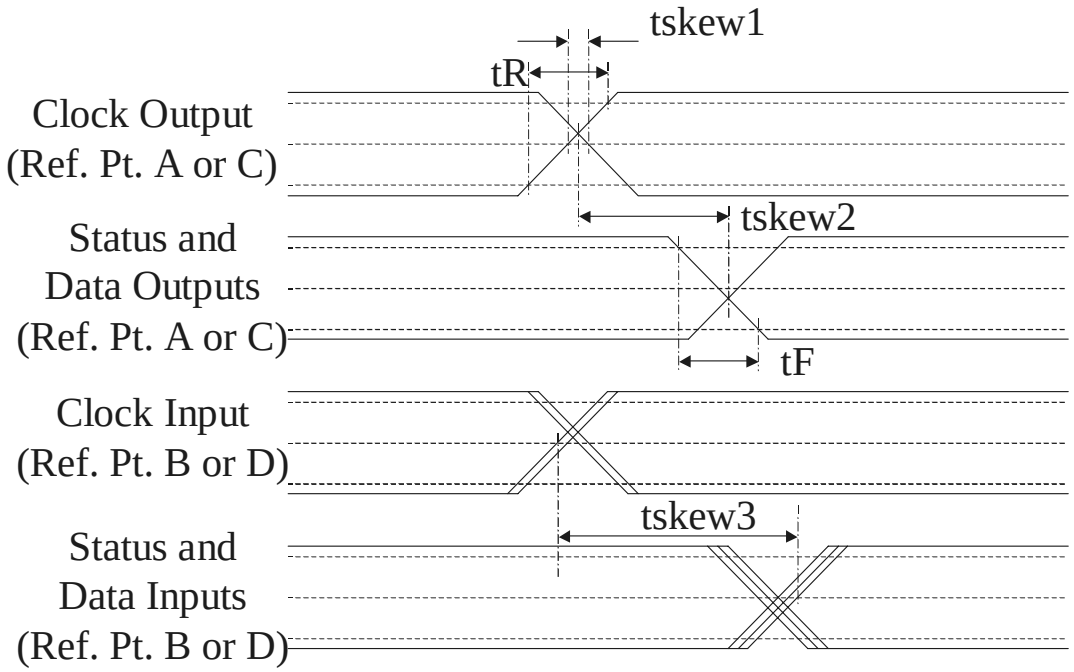


Figure 40. Reference Points for AC Timing Specifications



Though only shown for clock outputs in Figure 40, tskew1 also applies to data outputs.

11. Appendix A. Aggregation of NPSI Interfaces for Higher Bandwidth Applications (Informative)

11.1 Introduction

This appendix defines a dual-word or quad-word version of the Streaming Interface, called NPSI-Wide Mode (NPSI-W). Whereas the 16-bit wide version of the Streaming Interface (NPSI) transfers a single Control or Data word on the data path and one status report on the Status Channel at each clock cycle, NPSI-W behaves as two or four NPSI buses operating side-by-side. Two or four 16-bit words and two or four status reports are transferred at each clock cycle.

All operations of the NPSI-W behave in accordance to the NPSI definition. All control words and addressing capabilities are supported.

All data structures, bus parameters and AC timing parameters associated with the NPSI-W data path and the status channels are the same as for the NPSI. Payload data is sent in transfers containing multiples of 16 data bytes.

The NPSI-W bus is one recommended method of doubling or quadrupling the effective bandwidth on the Streaming Interface. However, all Streaming Interface compliant devices shall be capable of operating in NPSI mode; NPSI-W is not intended as an alternate default configuration for the bus.

11.2 Wide Bus Physical Interface

NPSI-W behaves as two side-by-side NPSI buses or four side-by-side buses. All NPSI interface signals are replicated except the Status Clocks. A transmitter shall source one clock per 16 data bits.

Two words are transferred at each clock cycle in the dual version, and four words are transferred at each clock cycle in the quad version. Thus, two or four Wcycles transpire at each clock cycle.

- In the dual version, the more significant portion of the bus (DAT[31:16], CTL[1]) carries the first Control or Data word while the less significant portion of the bus (DAT[15:0], CTL[0]) carries the second word.
- In the quad version, the most significant portion of the bus (DAT[63:48], CTL[3]) carries the first Control or Data word while the least significant portion of the bus (DAT[15:0], CTL[0]) carries the fourth word.

An NPSI-W bus shall have the same number of status channel bits per 16-bit data bus as the equivalent NPSI mode. That is, it shall have 2 bits of status per 16-bit data bus, but may have 4 bits of status per 16-bit data bus. A transmitter of the status bus shall source one clock for the entire NPSI-W status bus.

For the NPE-NPE interface, two or four status reports are transferred at each clock cycle.

- In the dual version, the more significant portion of the Status Channel (STAT[3:2]) carries the odd Calendar entries while the less significant portion (STAT[1:0]) carries the even Calendar entries.
- In the quad mode version, the most significant portion of the status channel (STAT[7:6]) carries the first calendar entry of each transfer. Each sequential calendar entry is placed on the status channel in a round-robin fashion. If a transfer group consists of 4 calendar entries, the most significant portion of the status channel (STAT[7:6]) carries the first calendar entry and the least significant portion of the status channel (STAT[1:0]) carries the last calendar entry of the group.

For the NPE-Fabric interface the 3B4B encoded data is spread across the status bits as defined in Section 8.4.1, “Flow Control Message Encoding and Framing.”

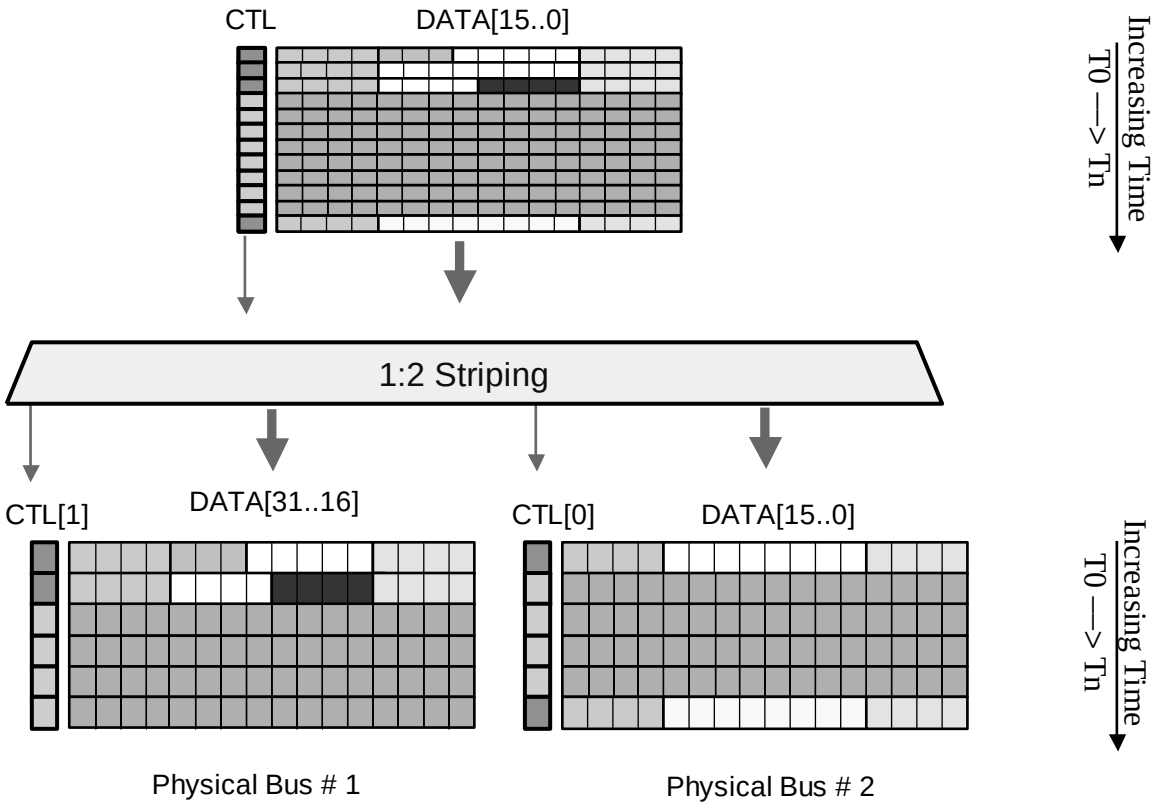
A word on the data bus may be either a Control word or a Data word. A control word is denoted by a high logic level on the associated CTL signal. A data word is denoted by a low logic level on the associated CTL signal. Thus, a single 32-bit or 64-bit transfer can contain any combination of control and data words.

11.3 Operation of the NPSI-W

There is a single protocol state machine handling both or all four physical buses.

The Control and Data Words (2 bytes) are striped back and forth across both buses in a TDM fashion from LEFT to RIGHT. This is depicted in the following figure. This is extensible to four buses by making Physical Bus #1 bits 63:48 and Physical Bus #4 bits 15:0.

Figure 41. 1:2 Striping Across Two Physical Buses



11.4 NPSI-W Signals

All interface signals other than the ones listed below operate in accordance with the NPSI definition in Section 5.4, “Interface Signals.”

Table 29. NPSI-W Interface Signals, Dual Mode (Modified Signals Only)

Signal	Direction	Description
DAT[31:0]	Source to Sink	The 32-bit Data bus (DAT) carries Data and Control words from the Source Device to the Sink Device DAT[31:24] carries the first byte transmitted, while DAT[7:0] the last byte transmitted. Within each byte, the most significant bit (DAT[31], DAT[23], DAT[15], and DAT[7]) is transmitted first, while the least significant bit (DAT[24], DAT[16], DAT[8], and DAT[0]) is transmitted last. The minimum data rate of each bit lane is 622 Mbps. Each group of 16 DAT bits is frequency locked to its associated DCLK.
CTL[1:0]	Source to Sink	The Control signals (CTL) identifies control words on the corresponding 16-bit section of the Data bus (DAT). In NPSI-W, CTL[1] is set high when a Control word is present on DAT[31:16]. CTL[0] is set high when a Control word is present on DAT[15:0]. The corresponding bit of CTL[1:0] is set low when the associated portion of DAT[31:0] contains a Data word. The minimum data rate of CTL is 622 Mb/s. CTL[i] is frequency locked to DCLK[i]
STAT[3:0] STAT[7:0]	Sink to Source	The status lines are replicated. The NPSI-W concatenates the status from each NPSI interface together as described in Section 7.2.1 and Section 8.4.1 (4 bits required and optionally extended to 8 bits).

Table 30. NPSI-W Interface Signals, Quad Mode (Modified Signals Only)

Signal	Direction	Description
DAT[63:0]	Source to Sink	The 64-bit Data bus (DAT) carries Data and Control words from the Source Device to the Sink Device DAT[63:56] carries the first byte transmitted, while DAT[7:0] the last byte transmitted. Within each byte, the most significant bit (DAT[n*8-1]) is transmitted first, while the least significant bit (DAT[(n-1)*8]) is transmitted last (where n is from 1 to 8, inclusive). The minimum data rate of each bit lane is 622 Mbps. Each group of 16 DAT bits is frequency locked to its associated DCLK.
CTL[3:0]	Source to Sink	The Control signals (CTL) identifies control words on the corresponding 16-bit section of the Data bus (DAT). In NPSI-W, CTL[3] is set high when a Control word is present on DAT[63:48]. CTL[0] is set high when a Control word is present on DAT[15:0]. The corresponding bit of CTL[3:0] is set low when the associated portion of DAT[63:0] contains a Data word. The minimum data rate of CTL is 622 Mb/s. CTL[i] is frequency locked to DCLK[i]
STAT[7:0] STAT[15:0]	Sink to Source	The status lines are replicated. The NPSI-W concatenates the status from each NPSI interface together as described in Section 7.2.1 and Section 8.4.1 (8 bits required and optionally extended to 16 bits).

11.5 De-skewing

The same de-skewing pattern and procedure is used for the NPSI-W mode as the NPSI mode except that the pattern is duplicated on all sections of the data/status bus interface. That is, the training pattern is always run across the full width of the data interface and the full width of the status interface. Though the training pattern is run across the full bus width, each 16-bit data lane is de-skewed using its associated source clock (e.g., DAT[15:0] and CTL[0] are de-skewed using DCLK[0]).

On the data interface, the Training Sequence consists of simultaneous series of 10 (repeated) training control words on all 16-bit words of the interface, followed by a simultaneous series of 10 (repeated) training data words on all 16-bit words of the interface. If the previous frame did not end on a full data width boundary, the transmitter shall insert Idle Control Words to ensure that the training sequence runs across the full width of the status interface.[35]

On the status interface, the Training Sequence consists of simultaneous series of 10 (repeated) words of all zeroes across the entire the interface, followed by a simultaneous series of 10 (repeated) words of all ones across the entire interface. If the previous status frame did not end on a full status width boundary, the transmitter shall insert an all zeroes pattern after the DIP-2 code to ensure that the training sequence runs across the full width of the status interface.[36]

[35] In this case, there may be two or more consecutive ICW's on the data interface.

[36] Care should be taken in the design so that the padding with zeros does not cause a false Loss of Status Path Synchronization condition.

12. Appendix B: NPE-NPE Narrow Interface Applications

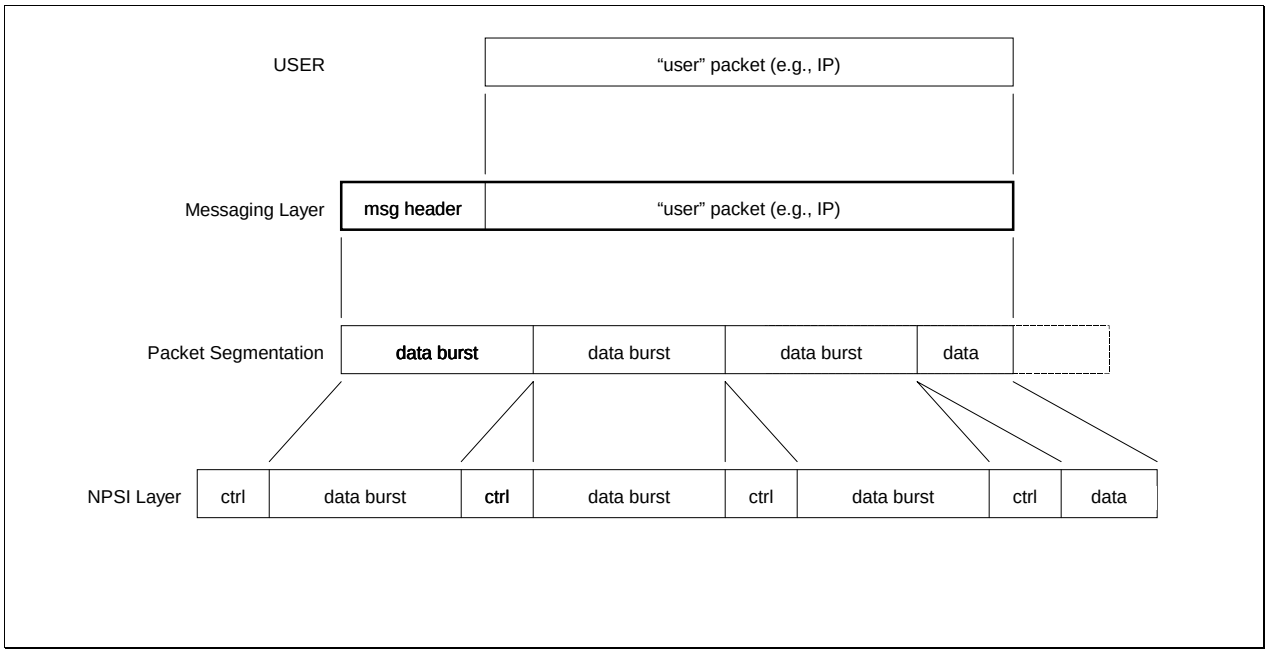
An NPSI device operating in NPE-NPE Mode shall support the 16-bit data path specified in Section 9.2 and may support the Narrow Mode Operation defined in Appendix F “Narrow Interface Applications” of the SPI-4 Phase 2 Specification. If an NPSI device supports the narrow mode, it shall provide a mechanism to configure the device to use either the required mode or the narrow mode.

13. Appendix C: NPSI Architectural Relationship Figures (Informative)

13.1 NPSI as Part of the NPF Layered Communication Model

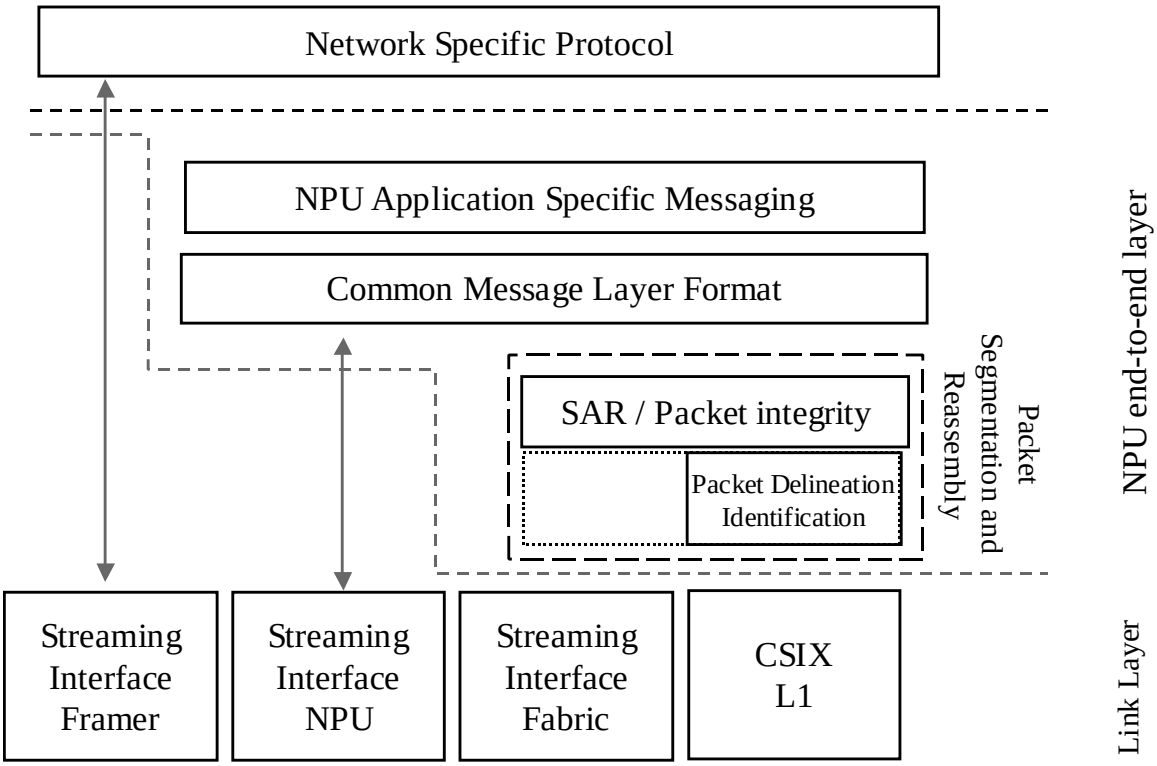
The NP Forum utilizes a layered communication model, of which NPSI is a component. (Note that the message layer is not required for the NPSI.)

Figure 42. NPF Layered Communication Model



13.2 NPF Streaming Interface and L2 Reference Model

Figure 43. Streaming Interface and L2 Reference Model



14. Appendix D. Features in CSIX-L1 Modified or Not Supported (Informative)

This appendix is an informative, not normative and not exhaustive discussion of the differences between the NPSI NPE-Fabric Mode and the CSIX-L1: Common Switch Interface Specification-L1.

14.1 Fabric Assumptions

As with CSIX-L1, the NPSI is intended to support board level connections of eight inches, although it is likely to support greater distances. Additionally, it is intended to support at least 1 connector between the NPE and the Fabric. It is intended to support the fabric architectures that were supported by CSIX-L1.

Unlike CSIX-L1, the NPSI supports packet delineation. It also supports sub-port addressing at the ingress and egress interfaces to a fabric, supporting a maximum of 4 or 16 logical interfaces per physical interface. The logical ports are addressable end-points of the NPSI protocol. Optionally, the NPSI supports flow control granularity per egress sub-port or per ingress-egress sub-port pair.

The NPSI defines interfaces on ingress and egress to a fabric. Correct system level behavior is likely to only occur if the fabric delivers data within a specific class between an ingress logical port and an egress logical port in the order in which the data was presented at the ingress interface (in-order delivery). This is consistent with the requirements on a CSIX fabric.

14.2 Unicast Operations

The NPSI supports an 8-bit class field for unicast traffic. Optionally, it supports a 6-bit and a 4-bit class field. It supports sub-port addressing as an option, providing for segment multiplexing to packets to different TM Ports and, as a further option, flow control per TM Port. The NPSI does not comment on TM Port addressing or flow control beyond the capabilities its sub-port features. (CSIX-L1 specifies use of the payload portion of the CFrame.)

The NPSI receives an Egress Physical Port Identifier as part of the ingress address format and replaces it with an Ingress Physical Port Identifier for transmission as part of the egress address format, making it available to identify the re-assembly context. The ingress Physical Port Identifier is inferred by the fabric. CSIX-L1 does not provide the ingress port address and considers the egress port address as undefined on egress.

14.3 Multicast Operations

As in CSIX-L1, support of multicast operations with the NPSI is optional. The NPSI supports a bitmap and an identifier representation of the multicast destinations. These roughly correspond to the bitmask and multicast ID representations in CSIX-L1. The NPSI does not support the binary copy mechanism of CSIX-L1.

The protocol formats on ingress and egress for multicast operations are different for the NPSI. They are identical for CSIX-L1, although some fields may be undefined on egress.

The NPSI supports 12 or 16 bit multicast identifiers. CSIX-L1 supports a 22-bit multicast ID. CSIX-L1 provides the multicast ID at the egress interface of the fabric, while the NPSI does not provide it.

The NPSI supports use of 128 bits for bitmap descriptions of multicast destinations. CSIX-L1 supports use of a 16-bit bitmask with an 8-bit offset (bitmask header). The CSIX-L1 mechanism supports addressing up to 4096 ports, while the NPSI mechanism supports only 128 ports. The CSIX-L1 mechanism is intended to support small systems of 16 to 32 ports.

14.4 Broadcast Operations

The NPSI uses multicast operations to support broadcast operations. CSIX-L1 has an independent mechanism for specifying the broadcast destination.

14.5 Flow Control

The NPSI provides flow control information for each simplex data interface via an independent, dedicated status interface of lesser bandwidth. CSIX-L1 provides similar information in-band, on the opposite half of the full duplex interface. In both cases, flow control information is delivered in messages.

The NPSI NPE-Fabric Mode may be configured to employ a diversity of mechanisms for the data interface. All of the mechanisms support a data ready bit on ingress and egress, providing link level flow control, similar to the DR bit in CSIX-L1.

The NPSI does not support a CR bit. Flow control messages are not passed on the data interface. (Control and status messages are expected to be layered on top of NSPI, not directly supported by it.) The NPSI does support an optional SNR bit to provide for link level backpressure of directed status flow control information.

CSIX-L1 provides for roughly symmetric flow control mechanisms at ingress and egress. CSIX-L1 provides for VOQ, Class wildcard and Port wildcard based flow control on ingress and egress. The NPSI provides for VOQ, Class (Port wildcard) and Global flow control on ingress. The NPSI provides for Class (Port wildcard) flow control on egress.

14.6 Physical Implementation

The NPSI specifies use of LVDS DDR signaling at a minimum clock rate of 311 MHz.

CSIX-L1 specifies use of LVCMOS or HSTL signaling at maximum clock rates of 166 MHz or 250 MHZ.

The NPSI provides an informative appendix as to how multiple instances of the interfaces may be integrated to support higher bandwidths.

CSIX-L1 specifies use of aggregated interfaces to support from 32 to 128 bits of data interface.

14.7 Message Formats

An NPSI burst corresponds to a CSIX-L1 CFrame. Most CFrames incur 8 bytes of frame overhead. Most NPSI bursts incur only 4 bytes of overhead.

CFrames may be any size up to 256 bytes, with the size of the payload specified in the header. The NPSI burst length is configured, with no maximum specified. The end of an NPSI burst is indicated by the EOPS in the subsequent control word, and not by a specified length field.

15. Appendix E. Differences Between SPI-4 Phase 2 and NPE-NPE mode (Informative)

This is the list of known differences between SPI-4 Phase 2 and the NPSI, NPE-NPE mode. It is not intended to be an exhaustive list.

The SI is capable of operating at data rates up to 650 MHz (DDR).

The SI uses dynamic alignment for data rates above 450 MHz (DDR).

The SI has burst sizes of MAX_SEGMENT_SIZE bytes or the last fragment of a packet; SPI-4 Phase 2 has burst sizes a multiple of 16 bytes up to a maximum configured payload data transfer size, or the last fragment of a packet.

The Control word formats have differing meanings and interpretations.

The status line uses LVDS, which is an optional mode in SPI-4 Phase 2. The status path includes an option for pooling of status.

16. Appendix F: Recommendations to Ensure Interoperability when Implementing NPE-Fabric Optional Features (Informative)

This appendix is an informative, not normative and not exhaustive discussion of the various implementation options described in the NPSI specification.

16.1.1 Implementation of Options and their dependencies

The tables below are guidelines for implementation of optional features intended to enable maximum interoperability. The row labeled required has the basic level of a feature, and it must be implemented. Features in rows with lower option levels should be implemented, if other features with a higher option level in the same vertical column are implemented. It is recommended that a device be able to operate at any lower option level than the highest implemented.

When deciding to implement a feature, only the column for that particular feature is applicable. There is no association between entries of the same option level in different columns. E.g. A switch fabric may implement Global flow control without supporting sub-ports, or it may implement sub-port support without supporting global flow control. Each table is independent of the other tables. An implementation may choose to support a different option level in each column of each table. If both multicast and sub-port options are supported, then the same sub-port configurations shall be supported for unicast and multicast.

16.1.2 Ingress Flow Control Options (NPE)

Option Level	**Unicast only (with or without sub-ports)**	
Required	Idle + DR bit + Unicast Queue Level	
Option Level 1	Class based flow control	Global flow control
Option Level 2	Class based flow control with sub-ports	

Option Level	**If you do Multicast, and sub-ports are not supported**	
Required	Idle + DR bit + Unicast Queue Level	
Option Level 1	Multicast class flow control (without Unicast / Multicast class mapping)	Global flow control
Option Level 2	Multicast class flow control (with Unicast / Multicast class mapping)	

Option Level	**If you do Multicast, and sub-ports are supported**		
Required	Idle + DR bit + Unicast Queue Level		
Option Level 1	Multicast class flow control (without Unicast / Multicast class mapping)		Global flow control
Option Level 2	Multicast class flow control with sub-ports (without Unicast / Multicast class mapping)	Multicast class flow control (with Unicast / Multicast class mapping)	
Option Level 3	Multicast class flow control with sub-ports (with Unicast / Multicast class mapping)	Multicast class flow control with sub-ports (with Unicast / Multicast class mapping)	

Option Level	**If you do Queue Map (with or without sub-ports)**	**Comment**
Required	Idle + DR bit + Unicast Queue Level and all other directed status options	
Option Level 1	Unicast Map	Used only when directed status (required option level) is disabled
Option Level 2	Multicast Map	Used only when directed status (required option level) is disabled

16.1.3 Ingress Flow Control Options (Fabric)

Option Level	**Unicast only (with or without subports)**	
Required	Idle + DR bit + Unicast Queue Level	
Option Level 1	Class based flow control	
Option Level 2	Class based flow control with sub-ports	Global flow control

Option Level	**If you do Multicast, and sub-ports are not supported**	
Required	Idle + DR bit + Unicast Queue Level	
Option Level 1	Multicast class flow control (without Unicast / Multicast class mapping)	
Option Level 2	Multicast class flow control (with Unicast / Multicast class mapping)	Global flow control

Option Level	**If you do Multicast, and sub-ports are supported**		
Required	Idle + DR bit + Unicast Queue Level		
Option Level 1	Multicast class flow control (without Unicast / Multicast class mapping)		
Option Level 2	Multicast class flow control with sub-ports (without Unicast / Multicast class mapping)	Multicast class flow control (with Unicast / Multicast class mapping)	Global flow control
Option Level 3	Multicast class flow control with sub-ports (with Unicast / Multicast class mapping)	Multicast class flow control with sub-ports (with Unicast / Multicast class mapping)	

Option Level	**If you do Queue Map**	**Comments**
Required	Idle + DR bit + Unicast Queue Level and all other directed status options	
Option Level 1	Unicast Map	Used only when directed status (required option level) is disabled
Option Level 2	Multicast Map	Used only when directed status (required option level) is disabled

16.1.4 Egress Flow Control Options

Option Level	**Unicast only (with and without subports)**
Required	Idle + DR bit
Option Level 1	Class based flow control
Option Level 2	Class and Sub-port based flow control

Option Level	**If you do Multicast, and sub-ports are not supported**
Required	Idle + DR bit
Option Level 1	Multicast class flow control (without Unicast / Multicast class mapping)
Option Level 2	Multicast class flow control (with Unicast / Multicast class mapping)

Option Level	**If you do Multicast, and sub-ports are supported**	
Required	Idle + DR bit	
Option Level 1	Multicast class flow control (without Unicast / Multicast class mapping)	
Option Level 2	Multicast class flow control with sub-ports (without Unicast / Multicast class mapping)	Multicast class flow control (with Unicast / Multicast class mapping)
Option Level 3	Multicast class flow control with sub-ports (with Unicast / Multicast class mapping)	Multicast class flow control with sub-ports (with Unicast / Multicast class mapping)

16.1.5 Ingress Address format Options (NPE)

Option Level	**Unicast**
Required	Unicast Physical Port Addressing Format (12.0.0.8)
Option Level 1	Unicast Subport addressing option 2 (10.2.2.6) AND Unicast Subport addressing option 1 (8.4.4.4)

Option Level	If you do Multicast ID	Comments
Required	Multicast ID Addressing Format 1 (12.0.0.8) AND Multicast ID Addressing Format 2 (16.0.0.4)	Configuration chooses one MC addressing mode
Option Level 1	Multicast ID Subport Addressing Option 1 (12.0.4.4) AND Multicast ID Subport Addressing Option 2 (12.0.2.6)	If Sub-ports are supported. Only one of them will be enabled at a time

Option Level	If you do Multicast Bitmap	Comments
Required	Multicast Bitmap Addressing Format 1 (12+16n.0.0.8) AND Multicast Bitmap Addressing Format 2 (16+16n.0.0.4)	Configuration chooses one MC addressing mode
Option Level 1	Multicast Bitmap Subport Addressing Format 1 (12+16n.0.4.4) AND Multicast Bitmap Subport Addressing Format 2 (12+16n.0.2.6)	If Sub-ports are supported. Only one of them will be enabled at a time

16.1.6 Ingress Address format Options (Fabric)

Option Level	Unicast
Required	Unicast Physical Port Addressing Format (12.0.0.8)
Option Level 1	Unicast Subport addressing option 2 (10.2.2.6) OR Unicast Subport addressing option 1 (8.4.4.4) OR BOTH

Option Level	**If you do Multicast ID**	**Comments**
Required	Multicast ID Addressing Format 1 (12.0.0.8) OR Multicast ID Addressing Format 2 (16.0.0.4) OR Both	Configuration chooses one MC addressing mode
Option Level 1	Multicast ID Subport Addressing Option 1 (12.0.4.4) OR Multicast ID Subport Addressing Option 2 (12.0.2.6)OR Both	If Sub-ports are supported. Only one of them will be enabled at a time

Option Level	**If you do Multicast Bitmap**	**Comments**
Required	Multicast Bitmap Addressing Format 1 (12+16n.0.0.8) OR Multicast Bitmap Addressing Format 2 (16+16n.0.0.4) OR Both	Configuration chooses one MC addressing mode
Option Level 1	Multicast Bitmap Subport Addressing Format 1 (12+16n.0.4.4) OR Multicast Bitmap Subport Addressing Format 2 (12+16n.0.2.6) OR Both	If Sub-ports are supported. Only one of them will be enabled at a time

16.1.7 Egress Address format Options (NPE)

Option Level	Unicast	Multicast	Comments
Required	Unicast Physical Port Addressing Format (12.0.0.8)	Multicast Physical Port Addressing Format (12.0.0.8)	
Option Level 1	Unicast Subport addressing option 2 (10.2.2.6) AND Unicast Subport addressing option 1 (8.4.4.4)	Multicast Subport addressing option 2 (10.2.2.6) AND Multicast Subport addressing option 1 (8.4.4.4)	If Sub-ports are supported. Only one of them will be enabled at a time

16.1.8 Egress Address format Options (Fabric)

Option Level	Unicast	Multicast	Comments
Required	Unicast Physical Port Addressing Format (12.0.0.8)	Multicast Physical Port Addressing Format (12.0.0.8)	
Option Level 1	Unicast Subport addressing option 2 (10.2.2.6) OR Unicast Subport addressing option 1 (8.4.4.4) OR BOTH	Multicast Subport addressing option 2 (10.2.2.6) OR Multicast Subport addressing option 1 (8.4.4.4) OR BOTH	If Sub-ports are supported. Only one of them will be enabled at a time

16.1.9 Flow Control Width Options

Option Level	Mode	Comment
Required	2 bits	
Option Level 1	4 bits	If 4 bits are implemented in an NPSI device, and it needs to interface to a device that has only two bits, it is recommended that bits [1:0] of the status bus be used.

16.1.10 Max_Segment_Size Options

Option Level	Size	Comment
Required	64 bytes	The rationale behind support for smaller sizes when implementing a larger size, is to enable using the largest possible segment size greater than 64. This does not preclude support for sizes that are not in the table.
Option Level 1	80	
Option Level 2	96	
Option Level 3	112	
Option Level 4	128	

16.1.11 Example Configuration Options

The table below summarizes various configuration attributes for the features listed in the tables above, the start-up parameters listed in section 8.5, and other features of the NPSI.

Note that some parameters are not applicable if the feature is not implemented.

Also, some of the parameters are not needed on egress. Ingress and egress may be independently configured.

This is a list of implementation choices. It is not intended to be an exhaustive list.

Feature	Attribute	Comments
Multicast	Bitmap Addressing Option	
	Bitmap size	Up to 128 bits
	Multicast ID Addressing Option	
	MultiCast Group ID Field size	Configure size of the field
	Number of ID's	
Flow Control	Multicast flow control option	
	Class flow control option	
	Number of UC classes supported	0 - 255
	Number of MC classes supported	0 - 255
	UC / MC Class Mapping Option	
	Global Flow Control Option	
	Number of Global Thresholds supported	Number of 8-bit code points supported
	Unicast Queue Map flow control option	
	Multicast Map flow control Option	
	Queue Map Size supported	Up to 4096
	SNR bit Option	
	Sub-ports flow control option	None, Egress Granularity, Ingress-Egress Granularity
	Status Bus Width	2, or 4 bits
Sub-ports	Sub-ports field size	2, or 4 bits
	Number of sub-ports supported	4 bit value
LODS	LODS response	
	Number of DIP-4 errors that trigger LODS	4 bit value
Training / Sync	Number of good control words needed to get back in sync	4 bit value