US 20250260758A1

(19) United States
(12) Patent Application Publication
George et al.
(10) Pub. No.: US 2025/0260758 A1
(43) Pub. Date: Aug. 14, 2025

(54) RDMA IN DATA CENTER APPLICATIONS

(71) Applicants: **George Madathilparambil George**, Bangalore (IN); **Susan George**, Bangalore (IN); **Mammen Thomas**, Seattle, WA (US)

(72) Inventors: **George Madathilparambil George**, Bangalore (IN); **Susan George**, Bangalore (IN); **Mammen Thomas**, Seattle, WA (US)

(21) Appl. No.: **19/192,102**

(22) Filed: **Apr. 28, 2025**

Related U.S. Application Data

(63) Continuation of application No. 18/991,816, filed on Dec. 23, 2024, now Pat. No. 12,341,703, which is a continuation of application No. 18/900,714, filed on Sep. 28, 2024, now Pat. No. 12,218,848, which is a continuation of application No. 18/648,425, filed on Apr. 28, 2024, now Pat. No. 12,126,537, which is a continuation of application No. 18/600,441, filed on Mar. 8, 2024, now Pat. No. 12,074,801, which is a continuation of application No. 18/201,779, filed on May 25, 2023, now Pat. No. 11,956,154, which is a continuation of application No. 17/834,097, filed on Jun. 7, 2022, now Pat. No. 11,706,148, which is a continuation of application No. 17/062,594, filed on Oct. 4, 2020, now Pat. No. 11,398,985.

Publication Classification

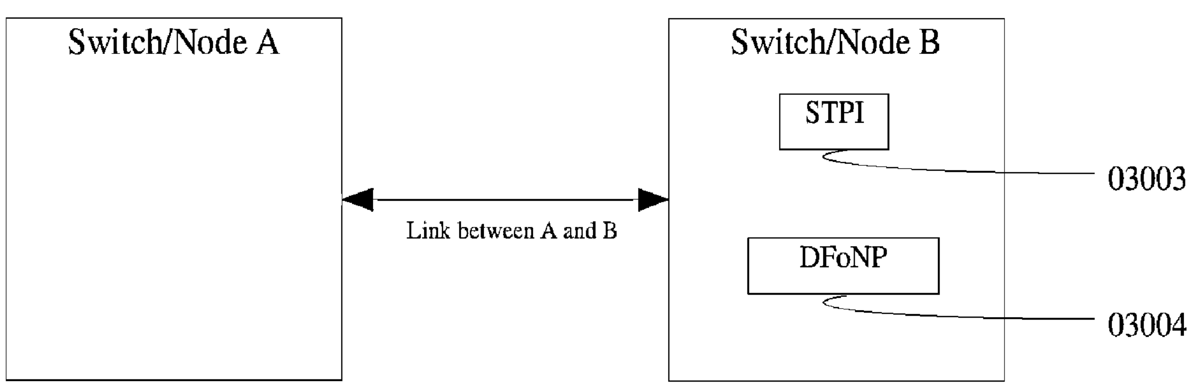
(51) Int. Cl.

H04L 69/324	(2022.01)
G06F 13/40	(2006.01)
G06F 13/42	(2006.01)
H04L 45/74	(2022.01)
H04L 47/125	(2022.01)
H04L 47/24	(2022.01)
H04L 49/25	(2022.01)
H04L 69/32	(2022.01)

(52) U.S. Cl.
CPC ***H04L 69/324*** (2013.01); ***G06F 13/4022*** (2013.01); ***G06F 13/4282*** (2013.01); ***H04L 45/74*** (2013.01); ***H04L 47/125*** (2013.01); ***H04L 47/24*** (2013.01); ***H04L 49/25*** (2013.01); ***H04L 69/32*** (2013.01)

(57) **ABSTRACT**

A network system in a data center comprises of a plurality of interconnected nodes with network switches that enable the interconnection between the nodes that also include end nodes that comprise at least a processor and at least a memory. The end nodes are enabled to act as source nodes for data or destination nodes for data being transferred over the network system. Remote Direct memory access (RDMA) technology enable data to be accessed directly from a first memory address at the source node and transmitted over the interconnected network nodes to deliver the data to a memory address at the destination node without engaging the processors at either end nodes. Bypassing the processors reduce the latency associated with data retrieval, data transfer and data storage thereby improving the efficiency of the data center. The plurality of switches are configured for per-flow congestion control (PFC) to limit data loss.



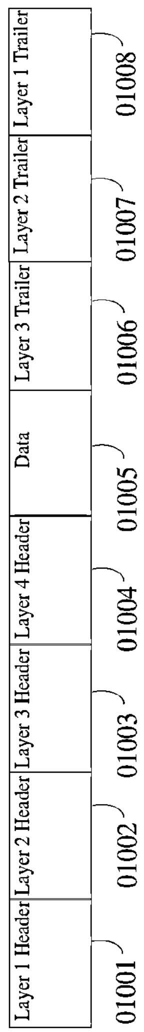


FIG. 1A PRIOR ART

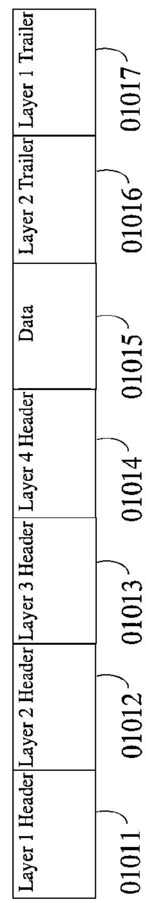


FIG. 1B PRIOR ART

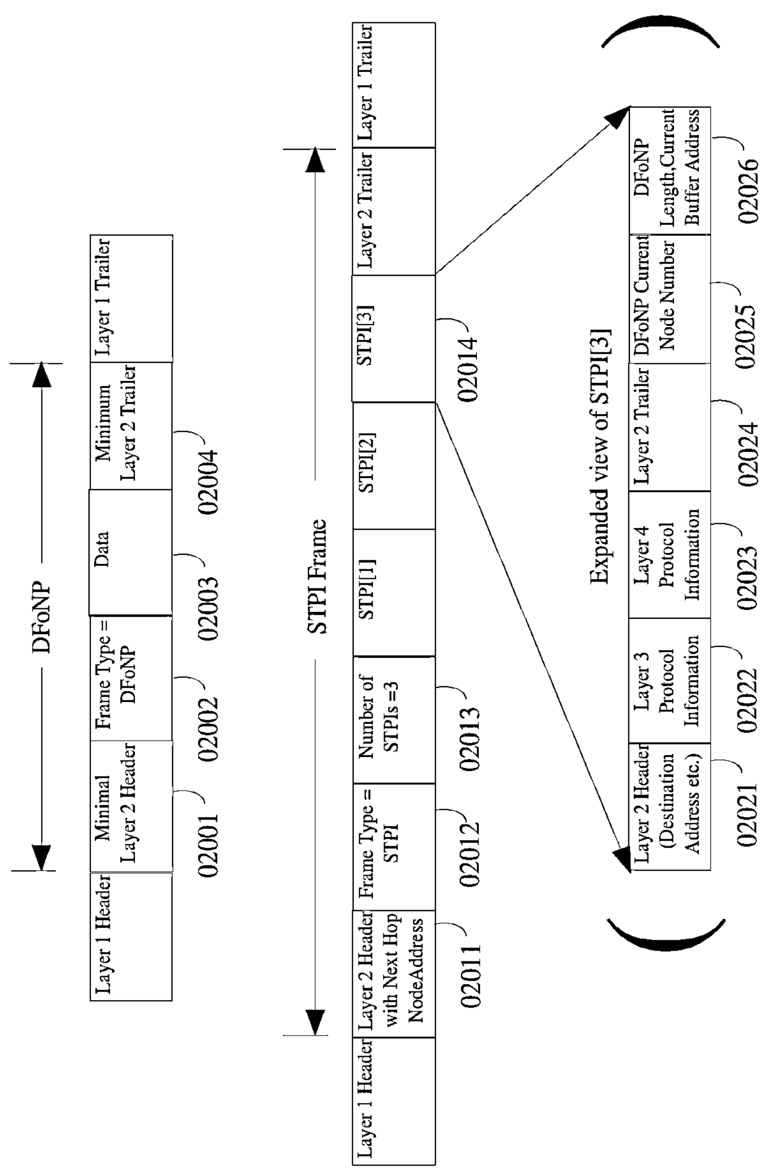

DFoNP
Layer 1 Header
Minimal Layer 2 Header
Frame Type = DFoNP
Data
Minimum Layer 2 Trailer
Layer 1 Trailer
02001
02002
02003
02004
STPI Frame
Layer 1 Header
Layer 2 Header with Next Hop NodeAddress
Frame Type = STPI
Number of STPIs =3
STPI[1]
STPI[2]
STPI[3]
Layer 2 Trailer
Layer 1 Trailer
02011
02012
02013
02014
Expanded view of STPI[3]
Layer 2 Header (Destination Address etc.)
Layer 3 Protocol Information
Layer 4 Protocol Information
Layer 2 Trailer
DFoNP Current Node Number
DFoNP Length,Current Buffer Address
02021
02022
02023
02024
02025
02026

FIG. 2A

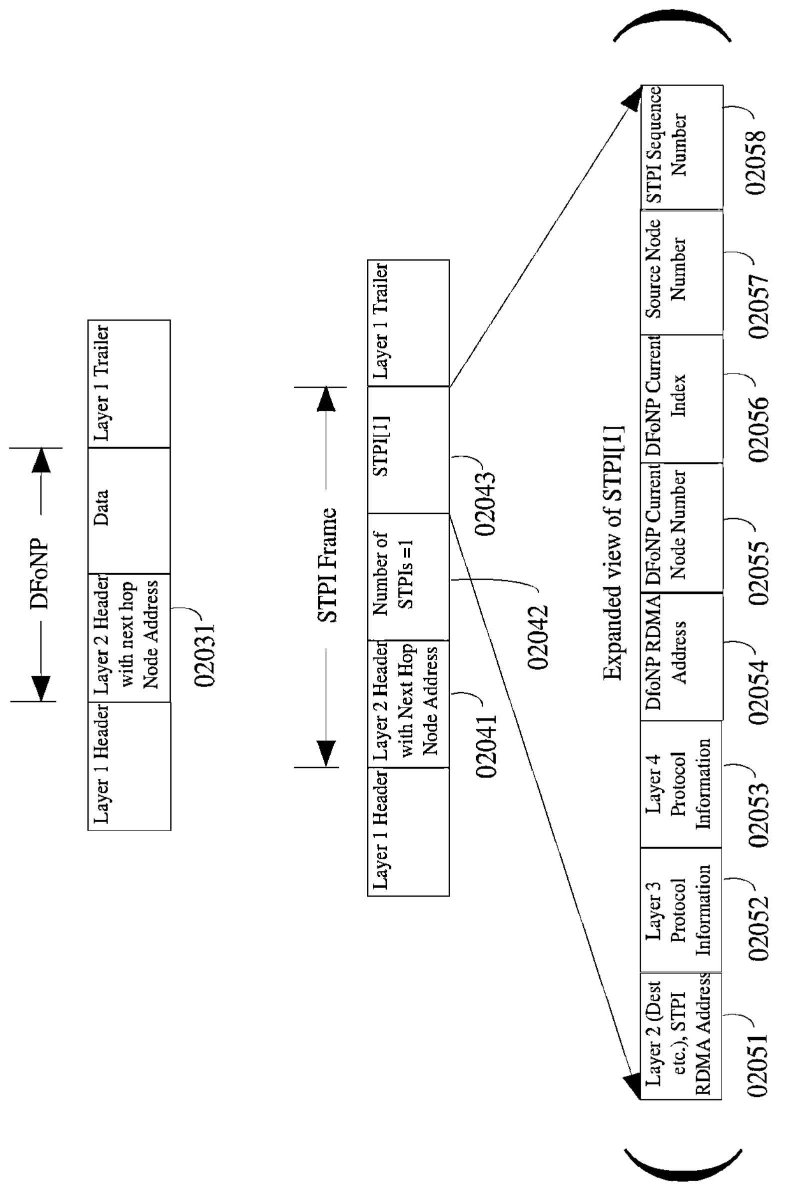

DFoNP
Layer 1 Header
Layer 2 Header with next hop Node Address
Data
Layer 1 Trailer
02031
STPI Frame
Layer 1 Header
Layer 2 Header with Next Hop Node Address
Number of STPIs =1
STPI[1]
Layer 1 Trailer
02041
02042
02043
Expanded view of STPI[1]
Layer 2 (Dest etc.), STPI RDMA Address
Layer 3 Protocol Information
Layer 4 Protocol Information
DfoNP RDMA Address
DFoNP Current Node Number
DFoNP Current Index
Source Node Number
STPI Sequence Number
02051
02052
02053
02054
02055
02056
02057
02058

FIG. 2B

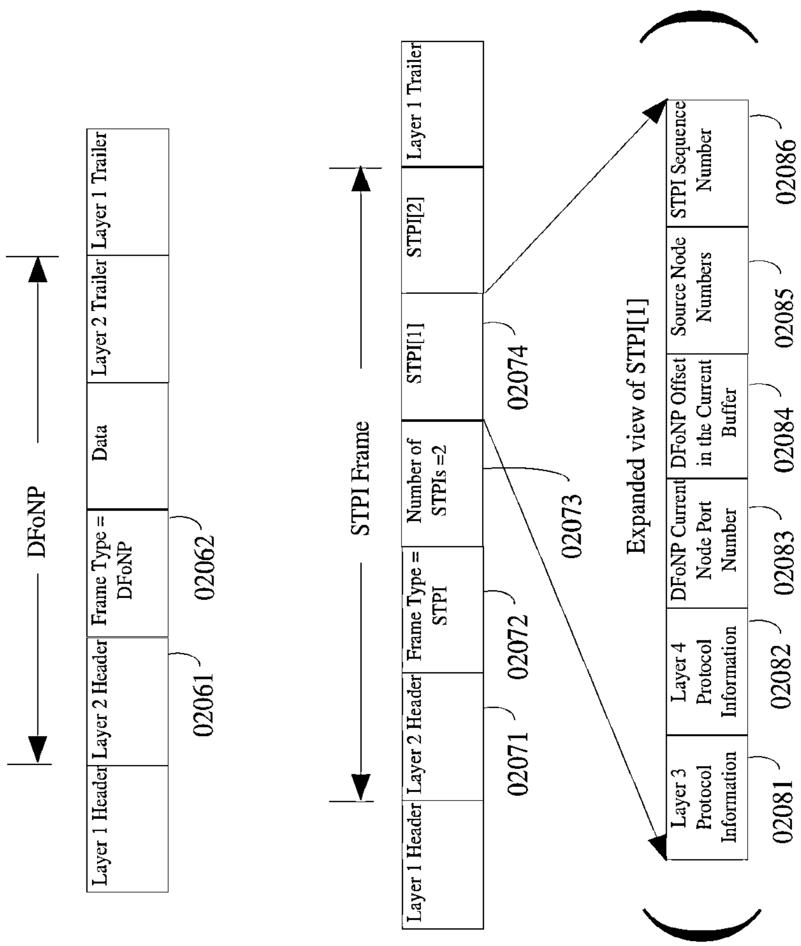

DFoNP
Layer 1 Header
Layer 2 Header
Frame Type = DFoNP
Data
Layer 2 Trailer
Layer 1 Trailer
02061
02062
STPI Frame
Layer 1 Header
Layer 2 Header
Frame Type = STPI
Number of STPIs =2
STPI[1]
STPI[2]
Layer 1 Trailer
02071
02072
02073
02074
Expanded view of STPI[1]
Layer 3 Protocol Information
Layer 4 Protocol Information
DFoNP Current Node Port Number
DFoNP Offset in the Current Buffer
Source Node Numbers
STPI Sequence Number
02081
02082
02083
02084
02085
02086

FIG. 2C

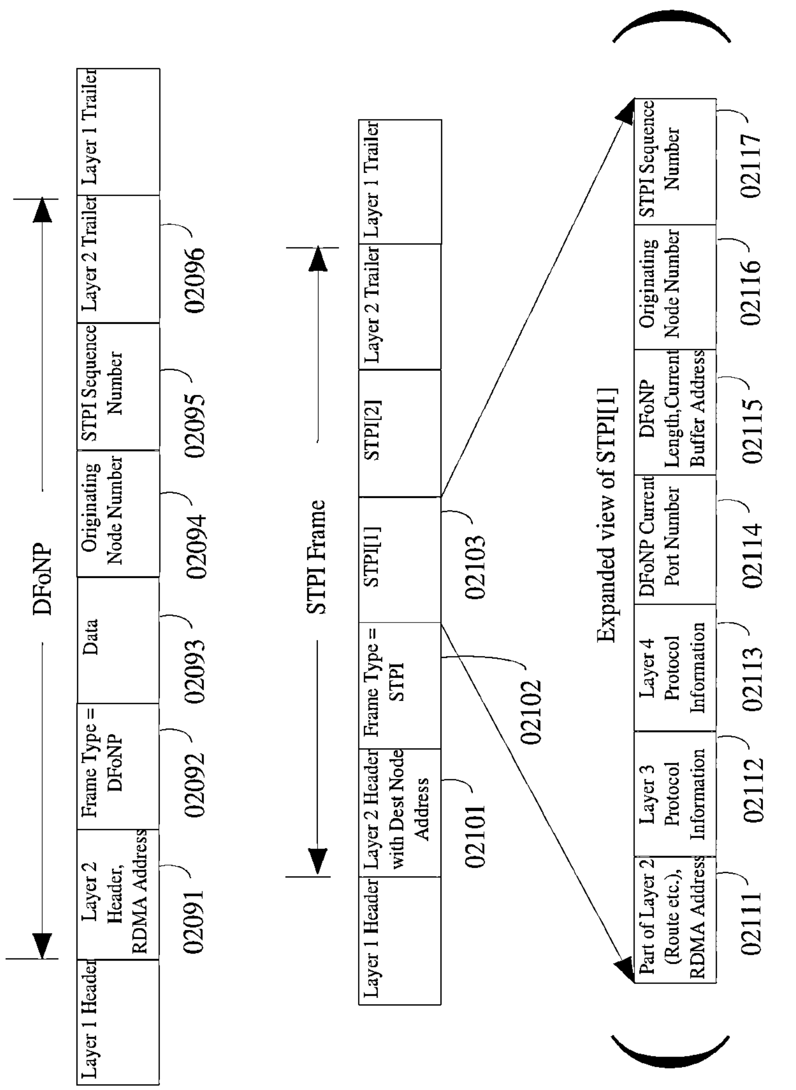

DFoNP
Layer 1 Header
Layer 2 Header, RDMA Address
Frame Type = DFoNP
Data
Originating Node Number
STPI Sequence Number
Layer 2 Trailer
Layer 1 Trailer
02091
02092
02093
02094
02095
02096
STPI Frame
Layer 1 Header
Layer 2 Header with Dest Node Address
Frame Type = STPI
STPI[1]
STPI[2]
Layer 2 Trailer
Layer 1 Trailer
02101
02102
02103
Expanded view of STPI[1]
Part of Layer 2 (Route etc.), RDMA Address
Layer 3 Protocol Information
Layer 4 Protocol Information
DFoNP Current Port Number
DFoNP Length,Current Buffer Address
Originating Node Number
STPI Sequence Number
02111
02112
02113
02114
02115
02116
02117

FIG. 2D

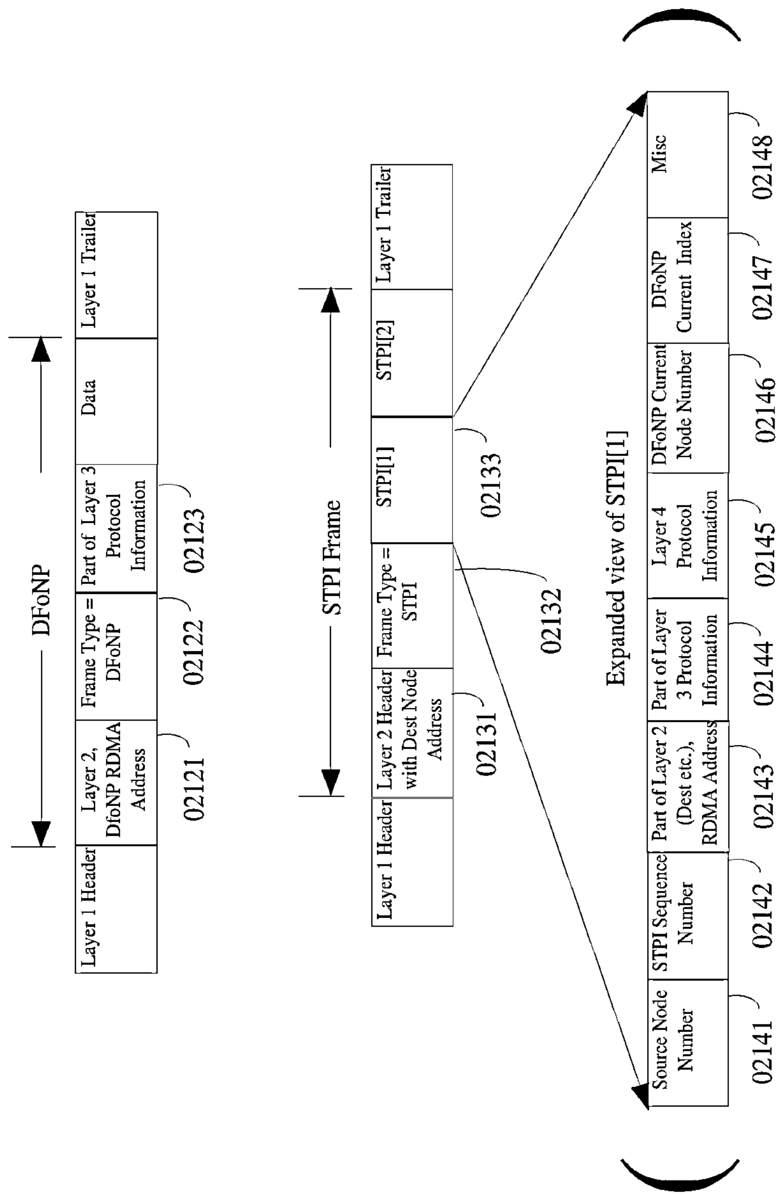

DFoNP
Layer 1 Header
Layer 2, DfoNP RDMA Address
Frame Type = DFoNP
Part of Layer 3 Protocol Information
Data
Layer 1 Trailer
02121
02122
02123
STPI Frame
Layer 1 Header
Layer 2 Header with Dest Node Address
Frame Type = STPI
STPI[1]
STPI[2]
Layer 1 Trailer
02131
02132
02133
Expanded view of STPI[1]
Source Node Number
STPI Sequence Number
Part of Layer 2 (Dest etc.), RDMA Address
Part of Layer 3 Protocol Information
Layer 4 Protocol Information
DFoNP Current Node Number
DFoNP Current Index
Misc
02141
02142
02143
02144
02145
02146
02147
02148

FIG. 2E

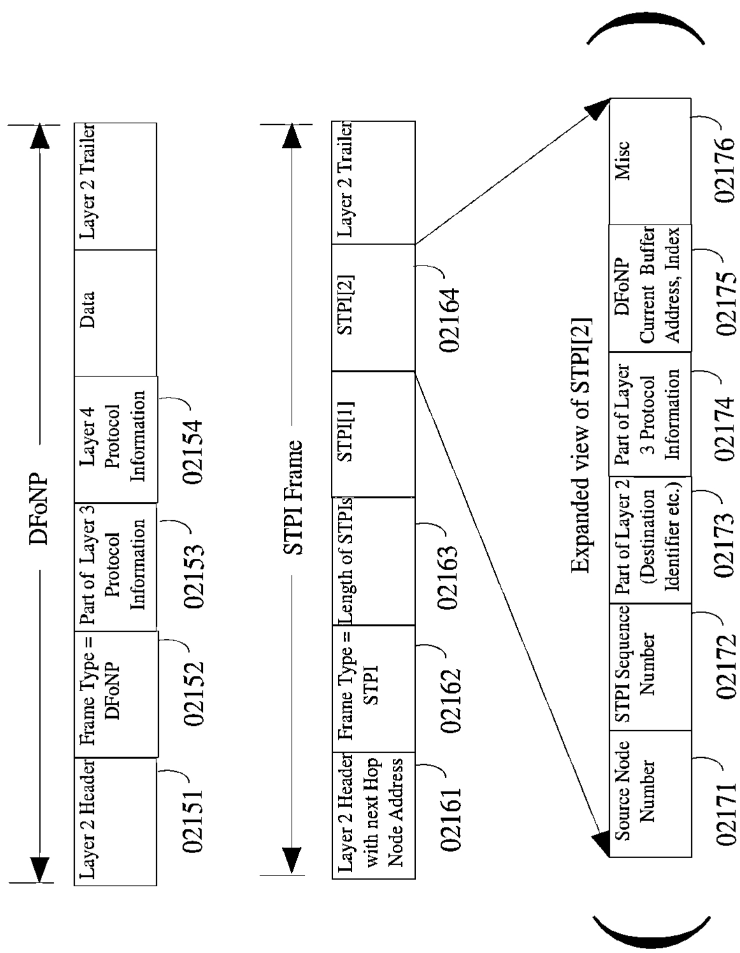

DFoNP
Layer 2 Header
Frame Type = DFoNP
Part of Layer 3 Protocol Information
Layer 4 Protocol Information
Data
Layer 2 Trailer
02151
02152
02153
02154
STPI Frame
Layer 2 Header with next Hop Node Address
Frame Type = STPI
Length of STPIs
STPI[1]
STPI[2]
Layer 2 Trailer
02161
02162
02163
02164
Expanded view of STPI[2]
Source Node Number
STPI Sequence Number
Part of Layer 2 (Destination Identifier etc.)
Part of Layer 3 Protocol Information
DFoNP Current Buffer Address, Index
Misc
02171
02172
02173
02174
02175
02176

FIG. 2F

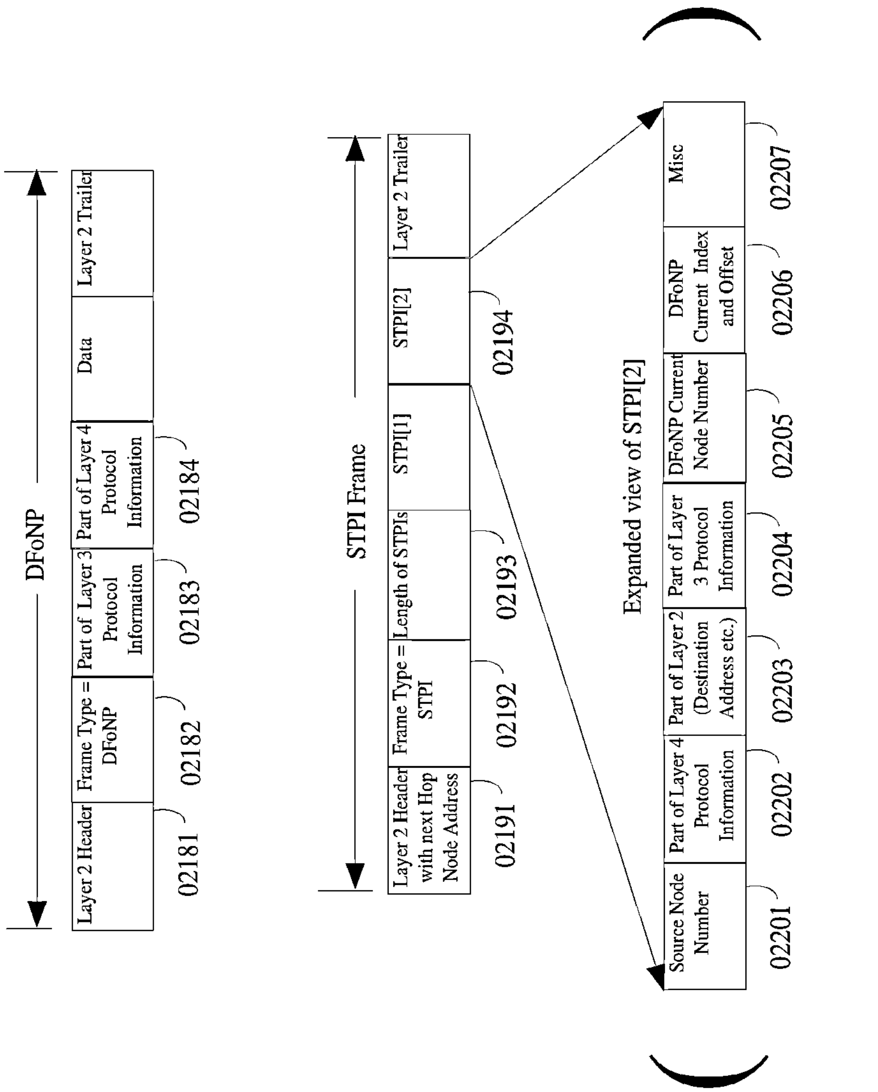

DFoNP
Layer 2 Header
Frame Type = DFoNP
Part of Layer 3 Protocol Information
Part of Layer 4 Protocol Information
Data
Layer 2 Trailer
02181
02182
02183
02184
STPI Frame
Layer 2 Header with next Hop Node Address
Frame Type = STPI
Length of STPIs
STPI[1]
STPI[2]
Layer 2 Trailer
02191
02192
02193
02194
Expanded view of STPI[2]
Source Node Number
Part of Layer 4 Protocol Information
Part of Layer 2 (Destination Address etc.)
Part of Layer 3 Protocol Information
DFoNP Current Node Number
DFoNP Current Index and Offset
Misc
02201
02202
02203
02204
02205
02206
02207

FIG. 2G

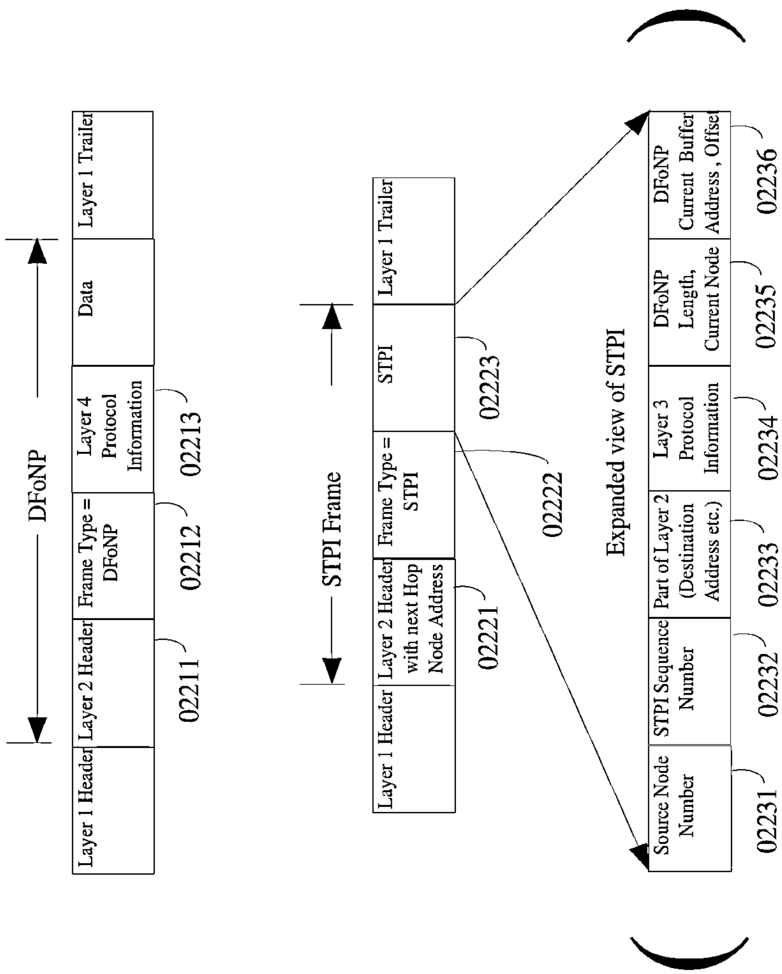

DFoNP
Layer 1 Header
Layer 2 Header
Frame Type = DFoNP
Layer 4 Protocol Information
Data
Layer 1 Trailer
02211
02212
02213
STPI Frame
Layer 1 Header
Layer 2 Header with next Hop Node Address
Frame Type = STPI
STPI
Layer 1 Trailer
02221
02222
02223
Expanded view of STPI
Source Node Number
STPI Sequence Number
Part of Layer 2 (Destination Address etc.)
Layer 3 Protocol Information
DFoNP Length, Current Node
DFoNP Current Buffer Address , Offset
02231
02232
02233
02234
02235
02236

FIG. 2H

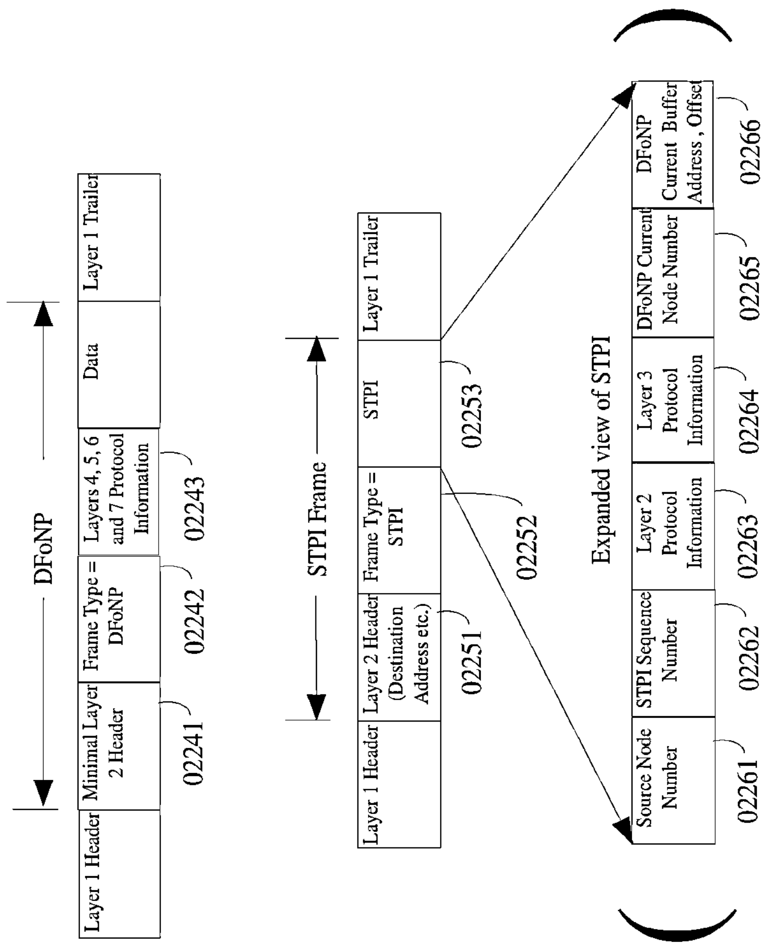

DFoNP
Layer 1 Header
Minimal Layer 2 Header
Frame Type = DFoNP
Layers 4, 5, 6 and 7 Protocol Information
Data
Layer 1 Trailer
02241
02242
02243
STPI Frame
Layer 1 Header
Layer 2 Header (Destination Address etc.)
Frame Type = STPI
STPI
Layer 1 Trailer
02251
02252
02253
Expanded view of STPI
Source Node Number
STPI Sequence Number
Layer 2 Protocol Information
Layer 3 Protocol Information
DFoNP Current Node Number
DFoNP Current Buffer Address , Offset
02261
02262
02263
02264
02265
02266

FIG. 2I

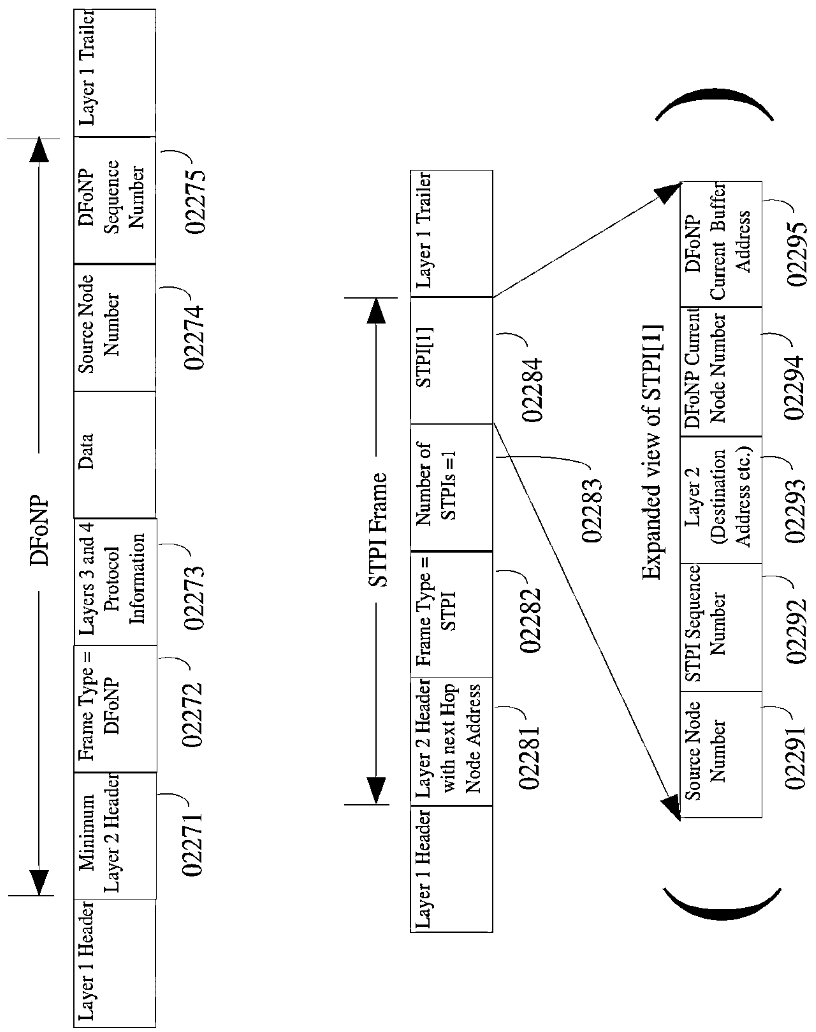

DFoNP
Layer 1 Header
Minimum Layer 2 Header
Frame Type = DFoNP
Layers 3 and 4 Protocol Information
Data
Source Node Number
DFoNP Sequence Number
Layer 1 Trailer
02271
02272
02273
02274
02275
STPI Frame
Layer 1 Header
Layer 2 Header with next Hop Node Address
Frame Type = STPI
Number of STPIs =1
STPI[1]
Layer 1 Trailer
02281
02282
02283
02284
Expanded view of STPI[1]
Source Node Number
STPI Sequence Number
Layer 2 (Destination Address etc.)
DFoNP Current Node Number
DFoNP Current Buffer Address
02291
02292
02293
02294
02295

FIG. 2J

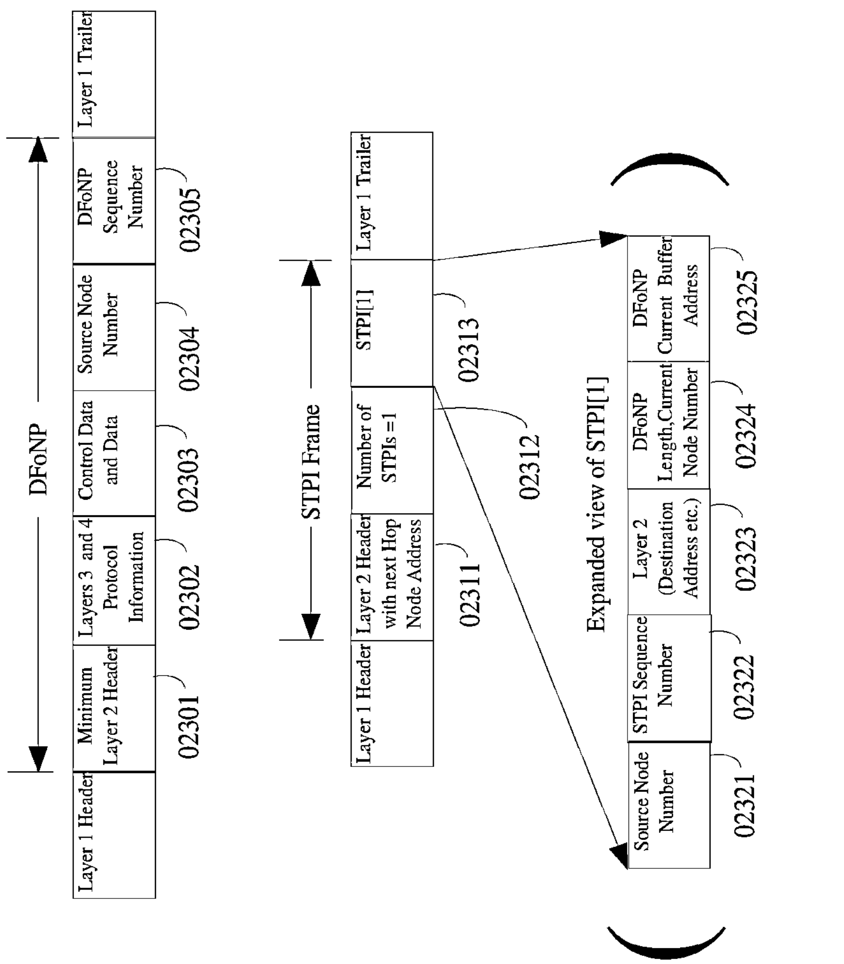

DFoNP
Layer 1 Header
Minimum Layer 2 Header
Layers 3 and 4 Protocol Information
Control Data and Data
Source Node Number
DFoNP Sequence Number
Layer 1 Trailer
02301
02302
02303
02304
02305
STPI Frame
Layer 1 Header
Layer 2 Header with next Hop Node Address
Number of STPIs =1
STPI[1]
Layer 1 Trailer
02311
02312
02313
Expanded view of STPI[1]
Source Node Number
STPI Sequence Number
Layer 2 (Destination Address etc.)
DFoNP Length, Current Node Number
DFoNP Current Buffer Address
02321
02322
02323
02324
02325

FIG. 2K

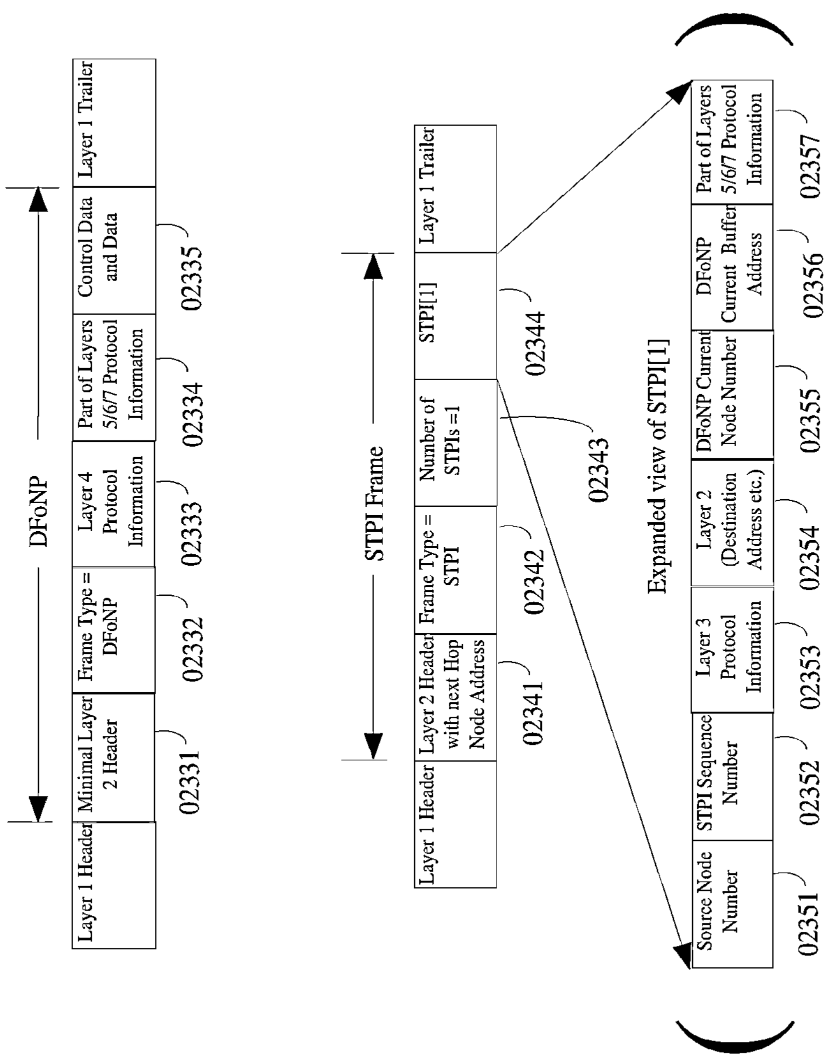

DFoNP
Layer 1 Header
Minimal Layer 2 Header
Frame Type = DFoNP
Layer 4 Protocol Information
Part of Layers 5/6/7 Protocol Information
Control Data and Data
Layer 1 Trailer
02331
02332
02333
02334
02335
STPI Frame
Layer 1 Header
Layer 2 Header with next Hop Node Address
Frame Type = STPI
Number of STPIs =1
STPI[1]
Layer 1 Trailer
02341
02342
02343
02344
Expanded view of STPI[1]
Source Node Number
STPI Sequence Number
Layer 3 Protocol Information
Layer 2 (Destination Address etc.)
DFoNP Current Node Number
DFoNP Current Buffer Address
Part of Layers 5/6/7 Protocol Information
02351
02352
02353
02354
02355
02356
02357

FIG. 2L

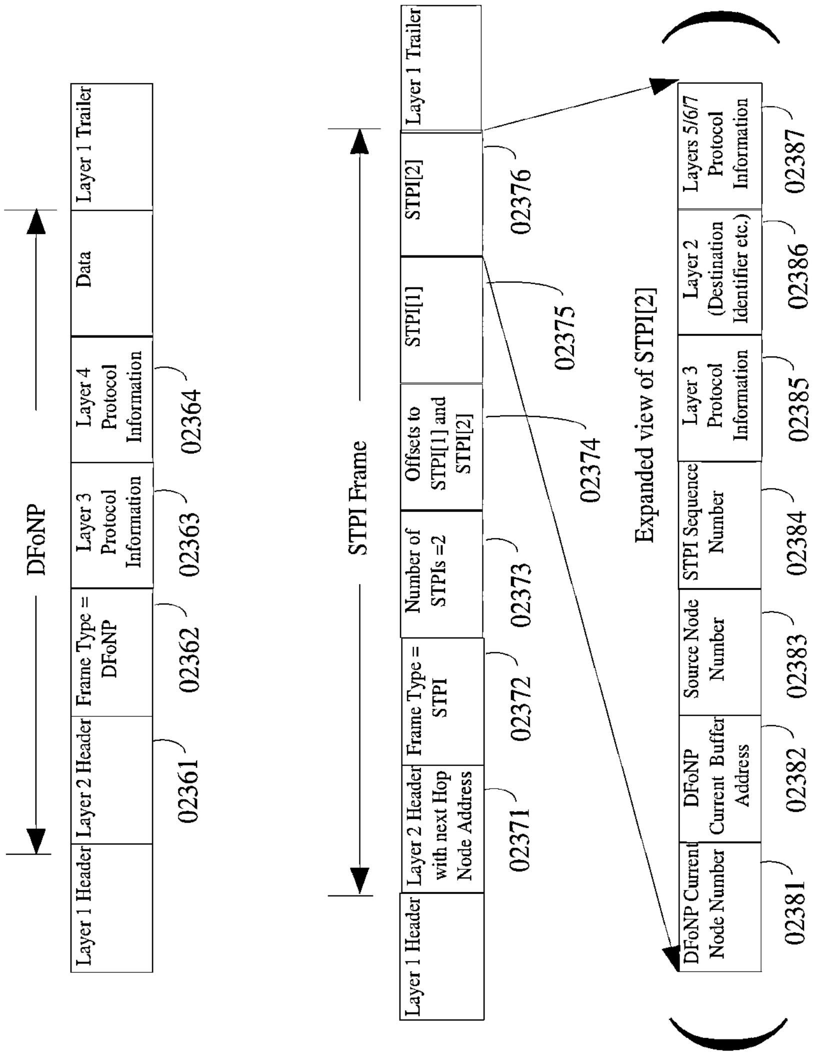

DFoNP
Layer 1 Header
Layer 2 Header
Frame Type = DFoNP
Layer 3 Protocol Information
Layer 4 Protocol Information
Data
Layer 1 Trailer
02361
02362
02363
02364
STPI Frame
Layer 1 Header
Layer 2 Header with next Hop Node Address
Frame Type = STPI
Number of STPIs =2
Offsets to STPI[1] and STPI[2]
STPI[1]
STPI[2]
Layer 1 Trailer
02371
02372
02373
02374
02375
02376
Expanded view of STPI[2]
DFoNP Current Node Number
DFoNP Current Buffer Address
Source Node Number
STPI Sequence Number
Layer 3 Protocol Information
Layer 2 (Destination Identifier etc.)
Layers 5/6/7 Protocol Information
02381
02382
02383
02384
02385
02386
02387

FIG. 2M

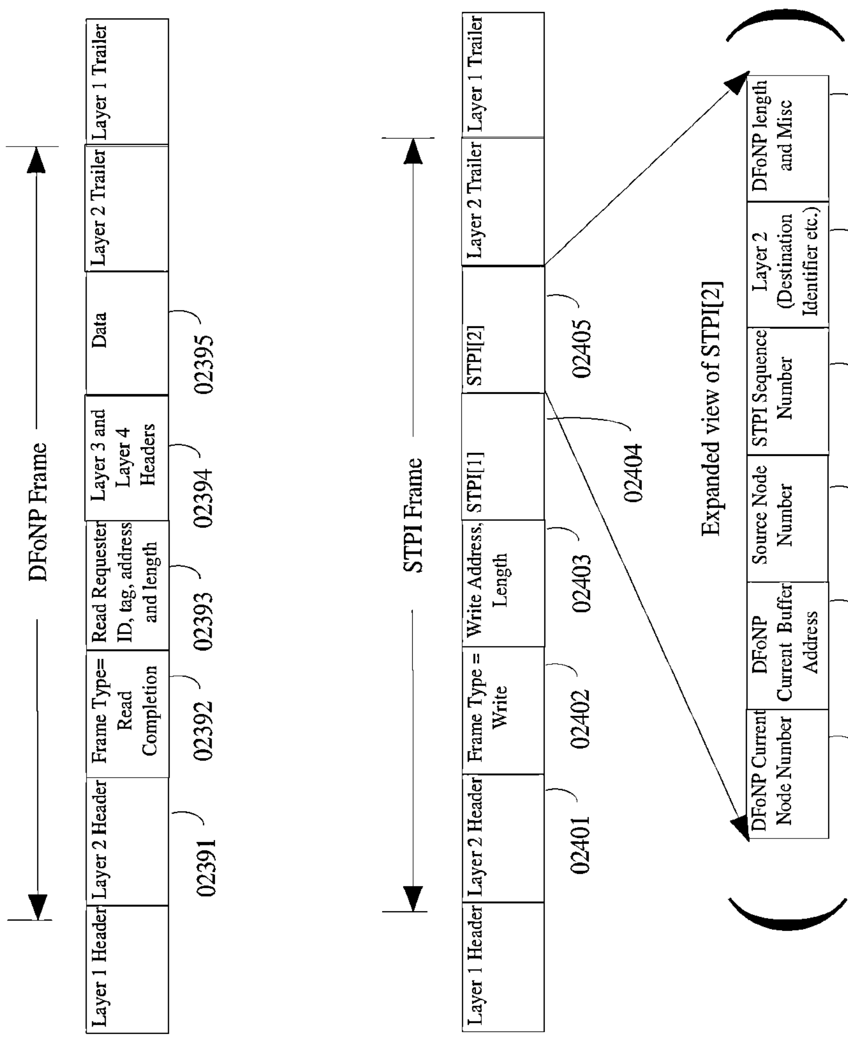

DFoNP Frame
Layer 1 Header
Layer 2 Header
Frame Type= Read Completion
Read Requester ID, tag, address and length
Layer 3 and Layer 4 Headers
Data
Layer 2 Trailer
Layer 1 Trailer
02391
02392
02393
02394
02395
STPI Frame
Layer 1 Header
Layer 2 Header
Frame Type = Write
Write Address, Length
STPI[1]
STPI[2]
Layer 2 Trailer
Layer 1 Trailer
02401
02402
02403
02404
02405
Expanded view of STPI[2]
DFoNP Current Node Number
DFoNP Current Buffer Address
Source Node Number
STPI Sequence Number
Layer 2 (Destination Identifier etc.)
DFoNP length and Misc
02411
02412
02413
02414
02415
02416

FIG. 2N

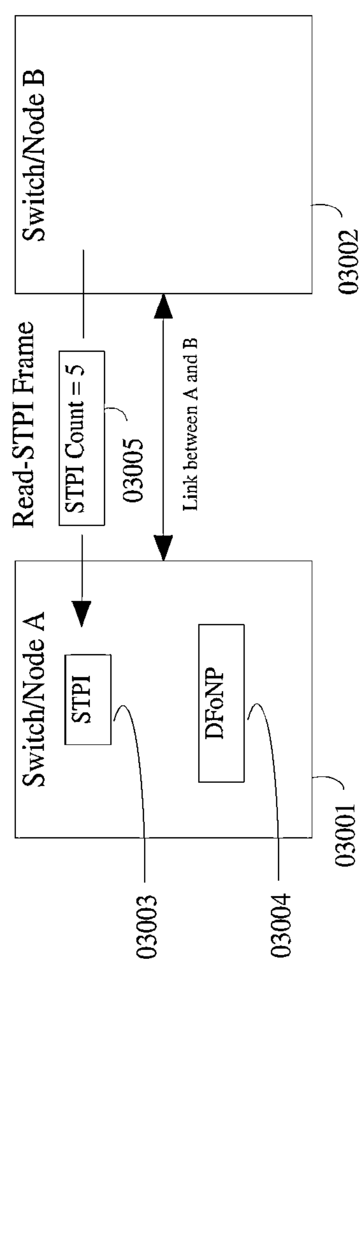

Switch/Node A
STPI
DFoNP
Switch/Node B
Read-STPI Frame
STPI Count = 5
Link between A and B
03001
03002
03003
03004
03005

FIG. 3A

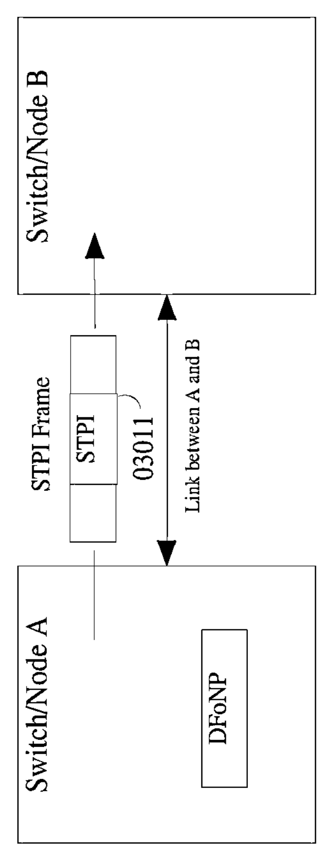

Switch/Node A
DFoNP
Switch/Node B
STPI Frame
STPI
Link between A and B
03011

FIG. 3B

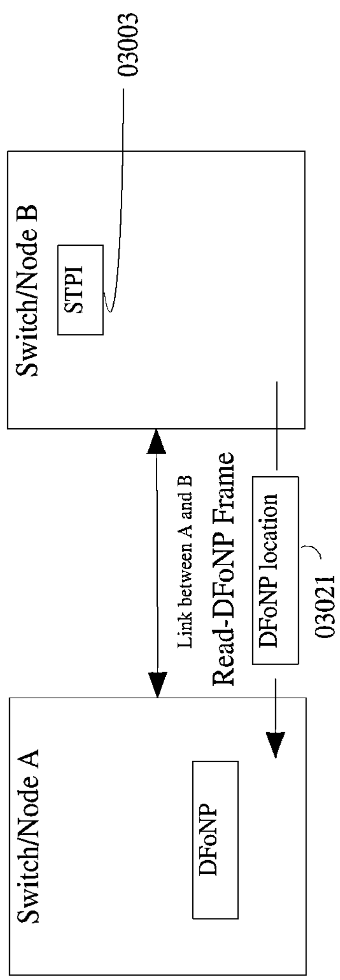

03003
Switch/Node B
STPI
Switch/Node A
DFoNP
Link between A and B
Read-DFoNP Frame
DFoNP location
03021

FIG. 3C

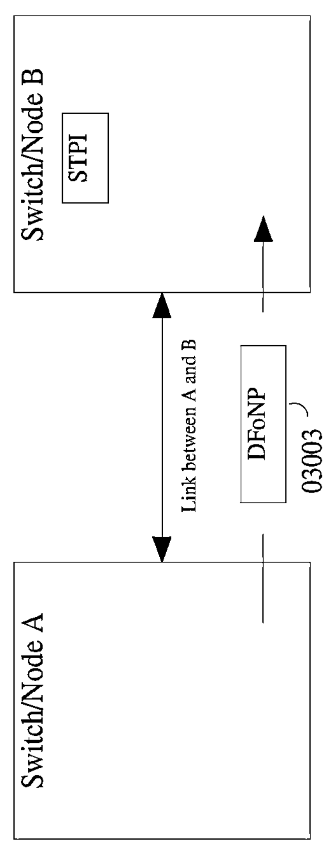

Switch/Node B
STPI
Switch/Node A
Link between A and B
DFoNP
03003

FIG. 3D

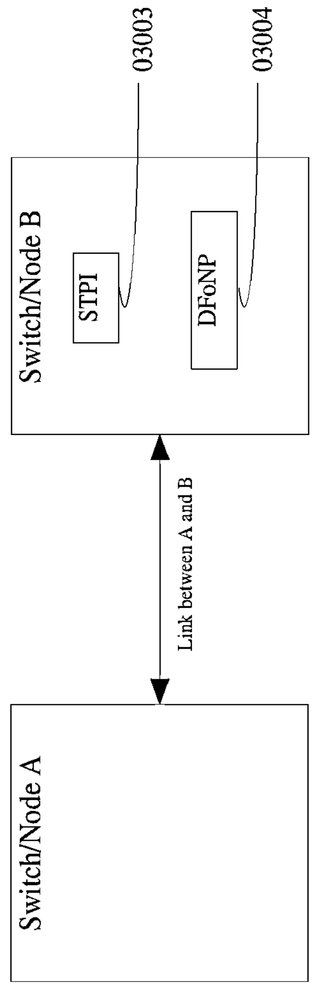

Switch/Node A
Switch/Node B
STPI
DFoNP
03003
03004
Link between A and B

FIG. 3E

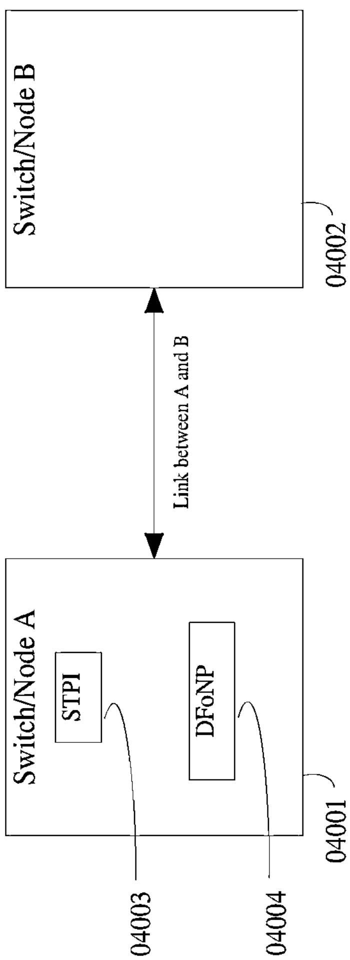

Switch/Node A
STPI
DFoNP
Switch/Node B
Link between A and B
04001
04002
04003
04004

FIG. 4A

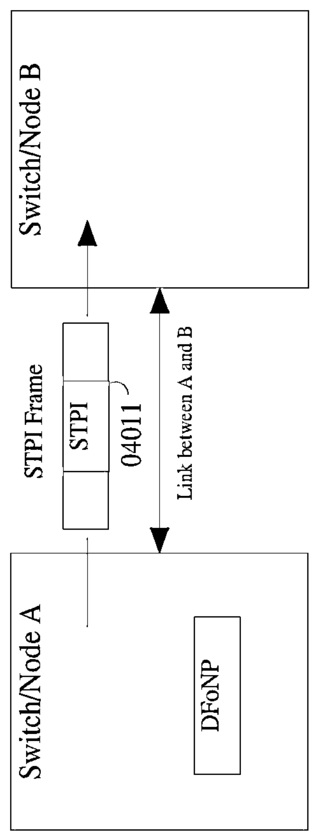

Switch/Node A
DFoNP
Switch/Node B
STPI Frame
STPI
04011
Link between A and B

FIG. 4B

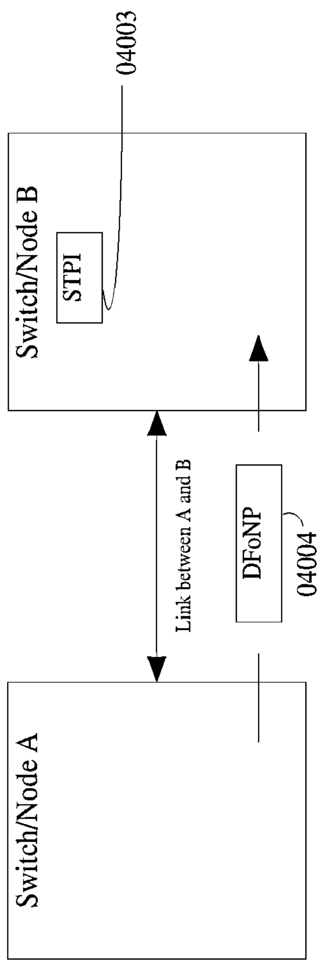

Switch/Node A
Switch/Node B
STPI
04003
Link between A and B
DFoNP
04004

FIG. 4C

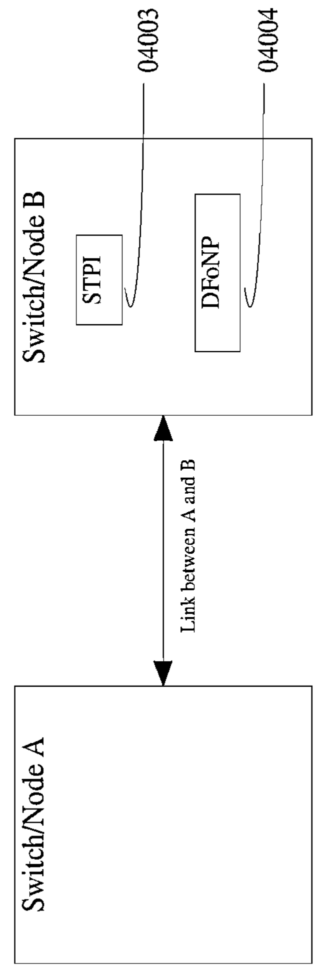

Switch/Node A
Switch/Node B
STPI
DFoNP
04003
04004
Link between A and B

FIG. 4D

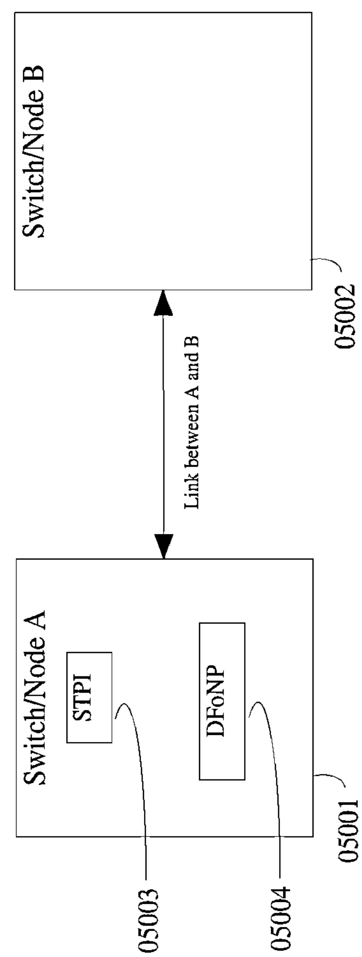

Switch/Node A
STPI
DFoNP
Switch/Node B
Link between A and B
05001
05002
05003
05004

FIG. 5A

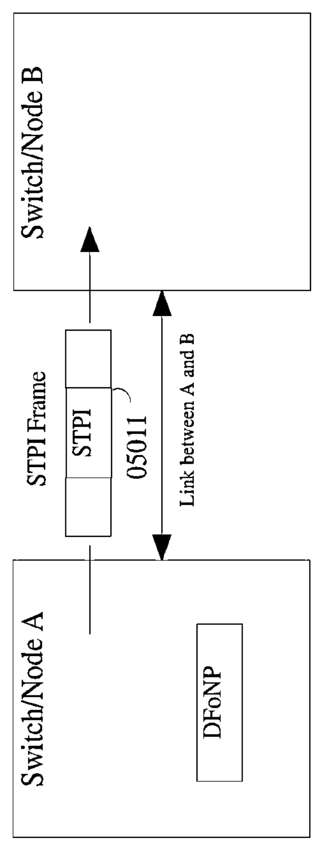

Switch/Node A
DFoNP
Switch/Node B
STPI Frame
STPI
05011
Link between A and B

FIG. 5B

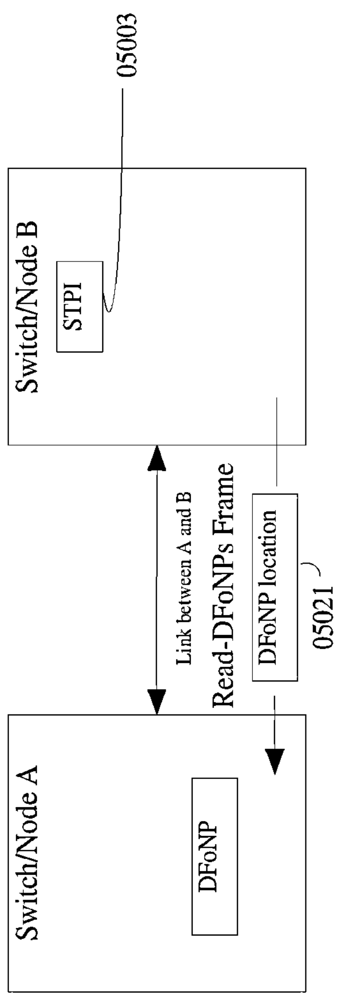

Switch/Node A
DFoNP
Switch/Node B
STPI
05003
Link between A and B
Read-DFoNPs Frame
DFoNP location
05021

FIG. 5C

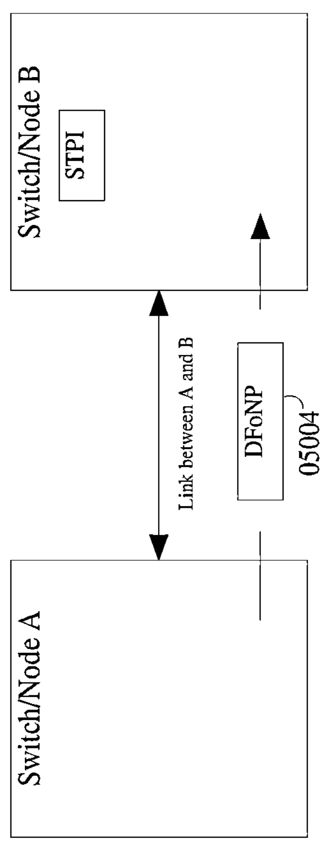

Switch/Node A
Switch/Node B
STPI
Link between A and B
DFoNP
05004

FIG. 5D

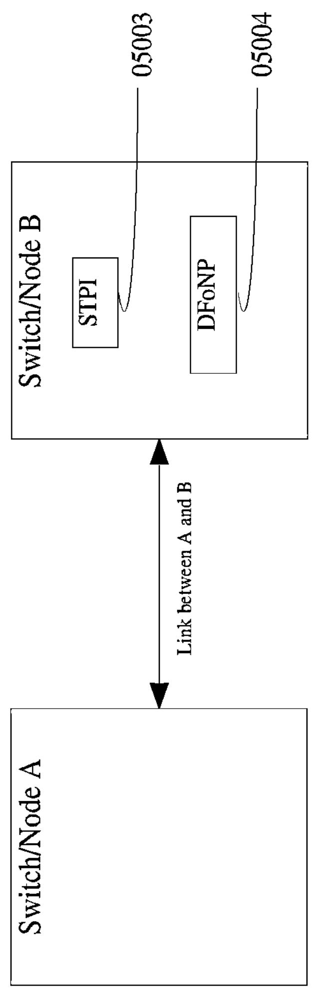

05003
05004
Switch/Node B
STPI
DFoNP
Link between A and B
Switch/Node A

FIG. 5E

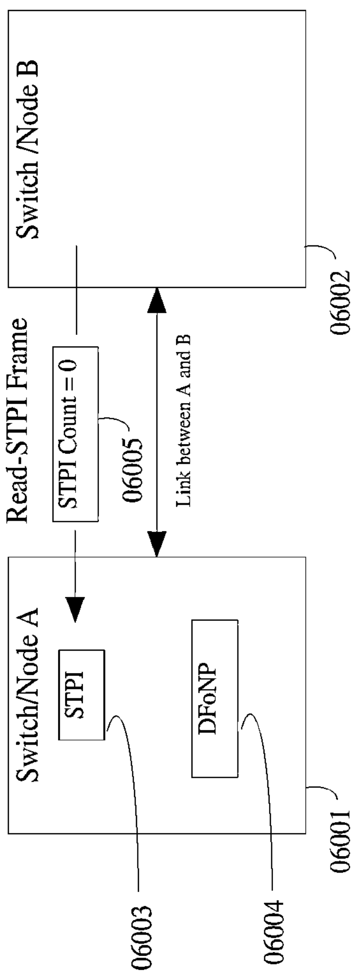

Switch/Node A
STPI
DFoNP
06001
06003
06004
Switch /Node B
06002
Read-STPI Frame
STPI Count = 0
06005
Link between A and B

FIG. 6A

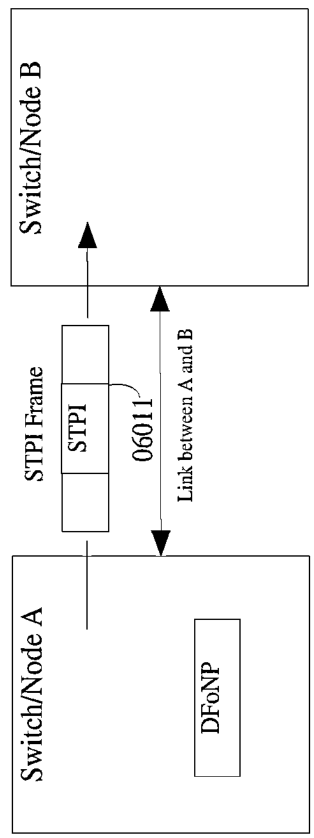

Switch/Node A
DFoNP
Switch/Node B
STPI Frame
STPI
06011
Link between A and B

FIG. 6B

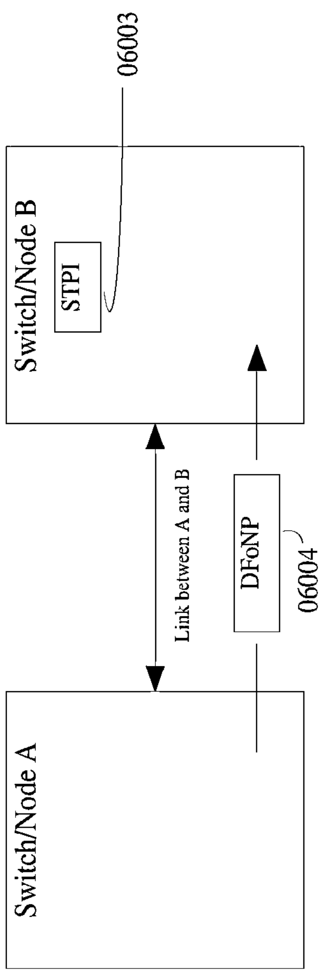

Switch/Node A
Switch/Node B
STPI
06003
Link between A and B
DFoNP
06004

FIG. 6C

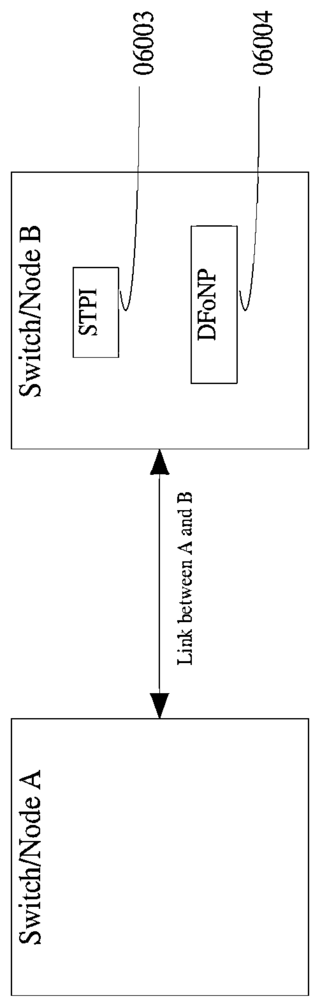

Switch/Node A
Switch/Node B
STPI
06003
DFoNP
06004
Link between A and B

FIG. 6D

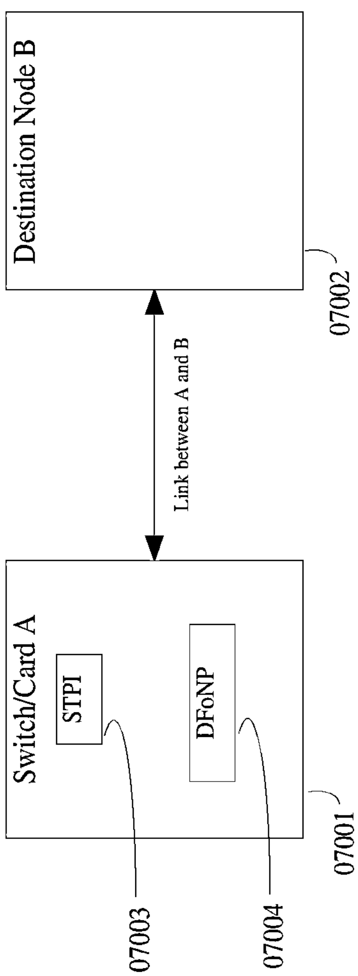

Destination Node B
Link between A and B
Switch/Card A
STPI
DFoNP
07001
07002
07003
07004

FIG. 7A

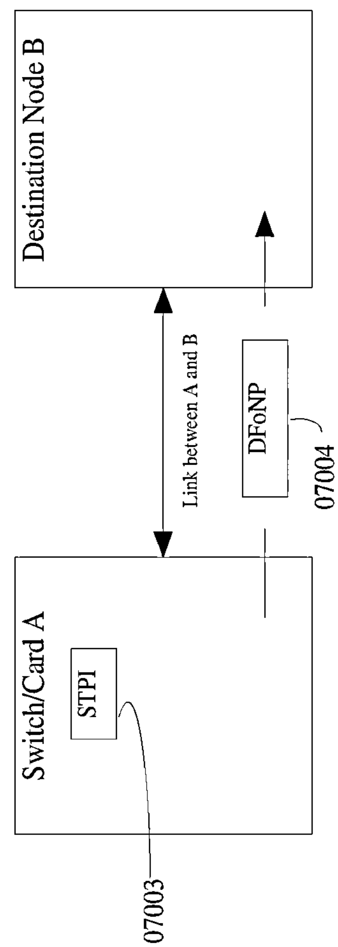

Destination Node B
Link between A and B
DFoNP
Switch/Card A
STPI
07003
07004

FIG. 7B

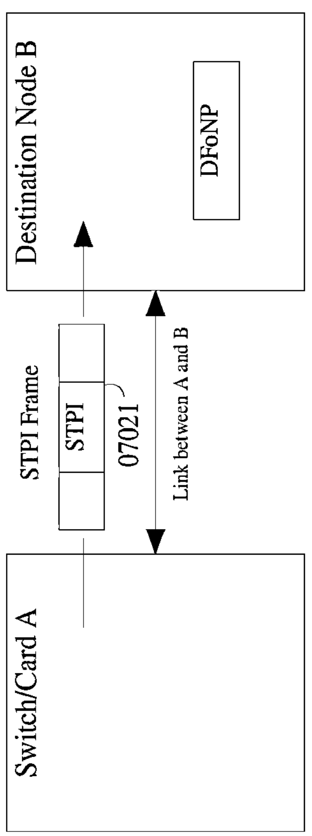

Switch/Card A
Destination Node B
DFoNP
STPI Frame
STPI
07021
Link between A and B

FIG. 7C

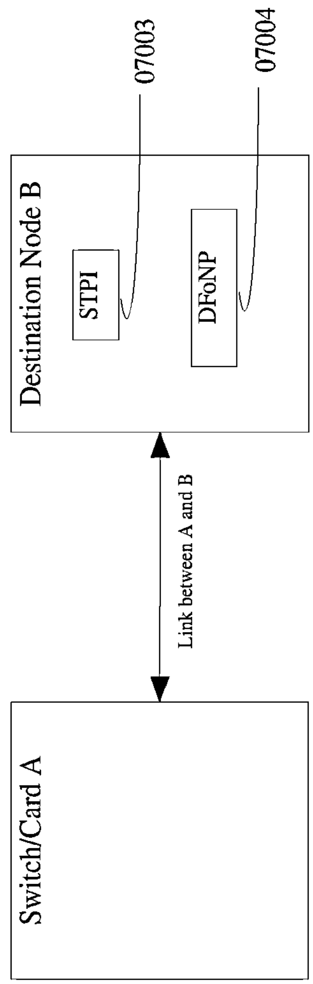

Switch/Card A
Destination Node B
STPI
DFoNP
07003
07004
Link between A and B

FIG. 7D

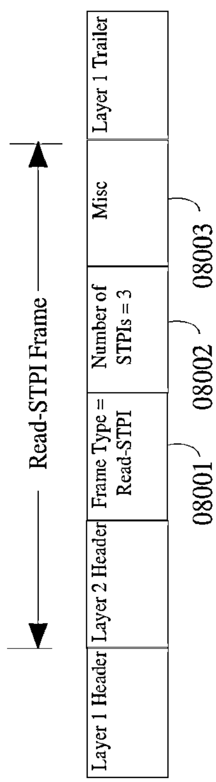

Read-STPI Frame
Layer 1 Header
Layer 2 Header
Frame Type = Read-STPI
Number of STPIs = 3
Misc
Layer 1 Trailer
08001
08002
08003

FIG. 8A

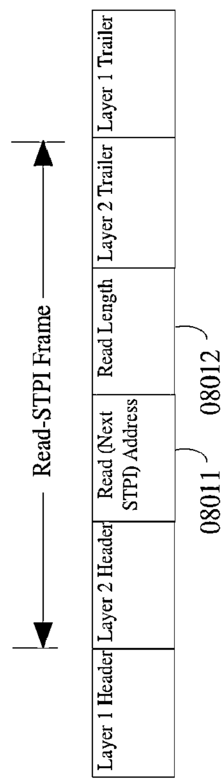

Read-STPI Frame
Layer 1 Header
Layer 2 Header
Read (Next STPI) Address
Read Length
Layer 2 Trailer
Layer 1 Trailer
08011
08012

FIG. 8B

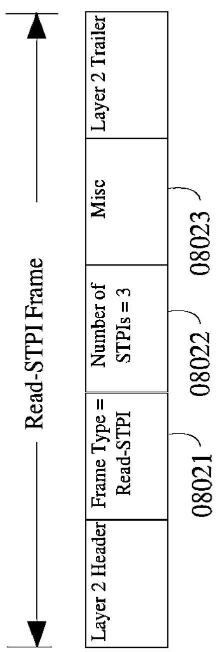

Read-STPI Frame
Layer 2 Header
Frame Type = Read-STPI
Number of STPIs = 3
Misc
Layer 2 Trailer
08021
08022
08023

FIG. 8C

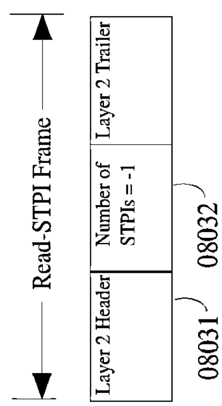
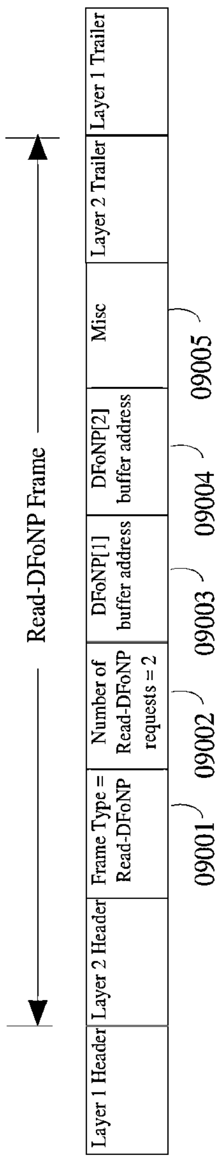
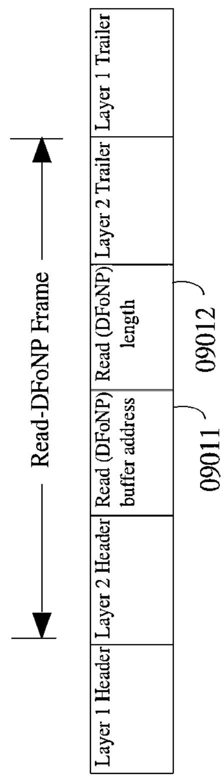

Read-STPI Frame
Layer 2 Header
Number of STPIs = -1
Layer 2 Trailer
08031
08032

FIG. 8D



Read-DFoNP Frame
Layer 1 Header
Layer 2 Header
Frame Type = Read-DFoNP
Number of Read-DFoNP requests = 2
DFoNP[1] buffer address
DFoNP[2] buffer address
Misc
Layer 2 Trailer
Layer 1 Trailer
09001
09002
09003
09004
09005

FIG. 9A



Read-DFoNP Frame
Layer 1 Header
Layer 2 Header
Read (DFoNP) buffer address
Read (DFoNP) length
Layer 2 Trailer
Layer 1 Trailer
09011
09012

FIG. 9B

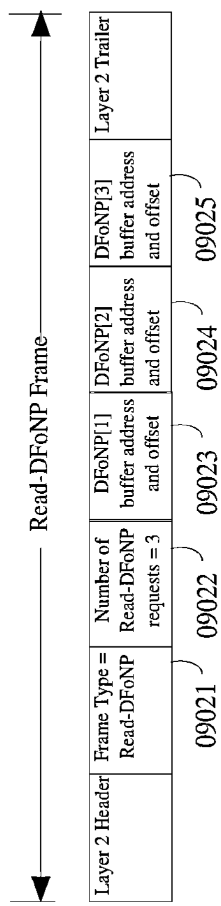

Read-DFoNP Frame
Layer 2 Header
Frame Type = Read-DFoNP
Number of Read-DFoNP requests = 3
DFoNP[1] buffer address and offset
DFoNP[2] buffer address and offset
DFoNP[3] buffer address and offset
Layer 2 Trailer
09021
09022
09023
09024
09025

FIG. 9C

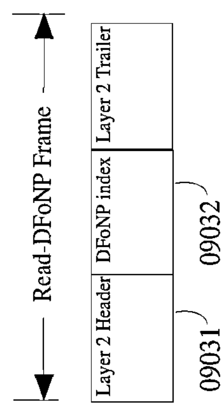

Read-DFoNP Frame
Layer 2 Header
DFoNP index
Layer 2 Trailer
09031
09032

FIG. 9D

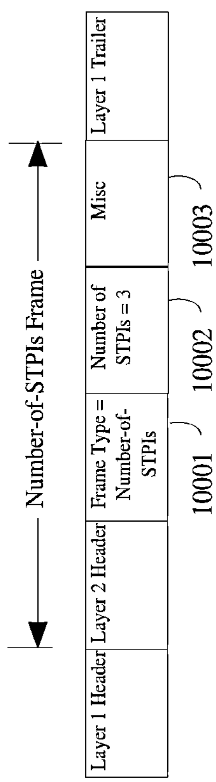


FIG. 10A

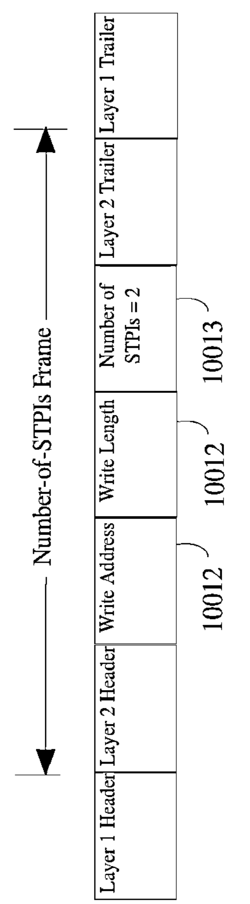


FIG. 10B

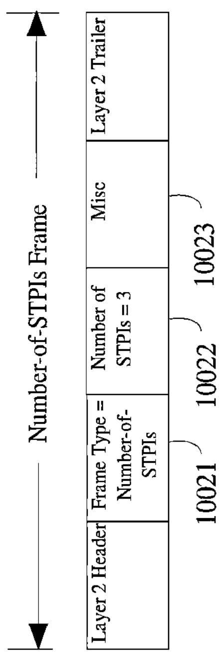


FIG. 10C

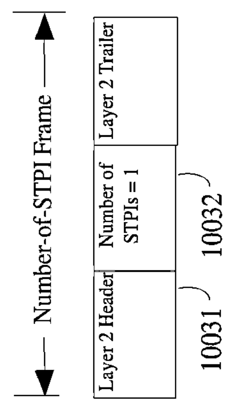


FIG. 10D

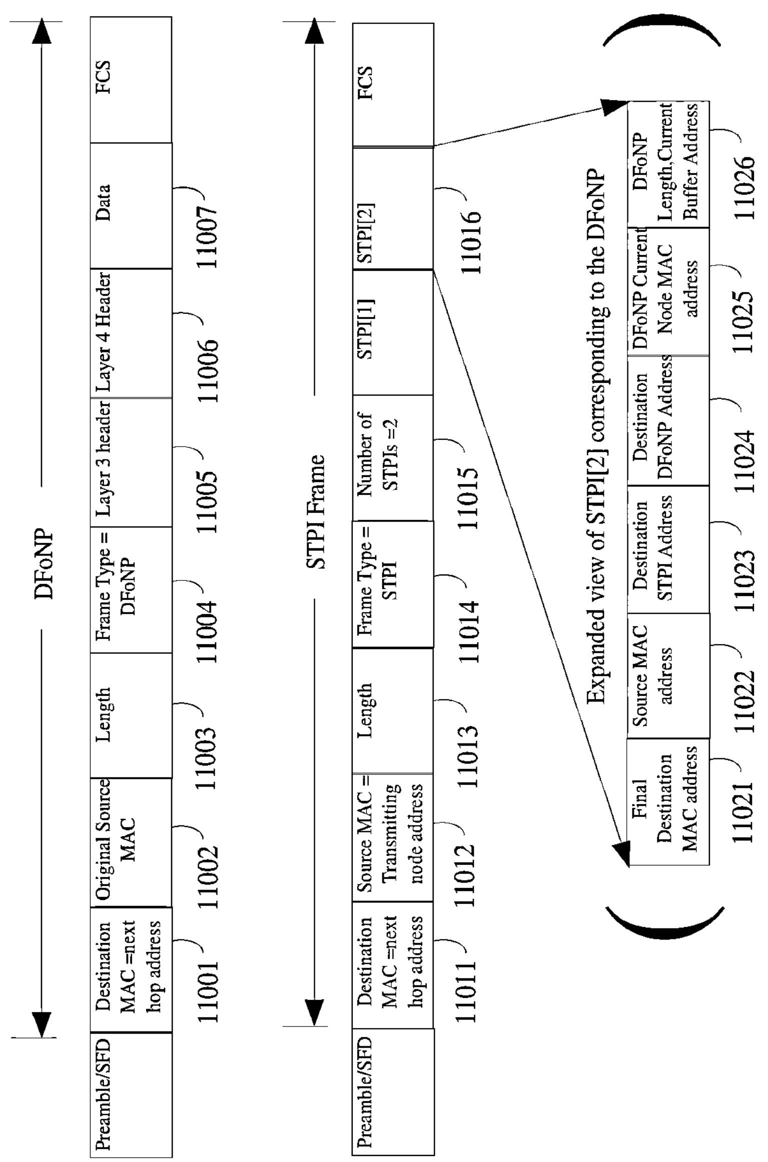

DFoNP
Preamble/SFD
Destination MAC =next hop address
Original Source MAC
Length
Frame Type = DFoNP
Layer 3 header
Layer 4 Header
Data
FCS
11001
11002
11003
11004
11005
11006
11007
STPI Frame
Preamble/SFD
Destination MAC =next hop address
Source MAC = Transmitting node address
Length
Frame Type = STPI
Number of STPIs =2
STPI[1]
STPI[2]
FCS
11011
11012
11013
11014
11015
11016
Expanded view of STPI[2] corresponding to the DFoNP
Final Destination MAC address
Source MAC address
Destination STPI Address
Destination DFoNP Address
DFoNP Current Node MAC address
DFoNP Length,Current Buffer Address
11021
11022
11023
11024
11025
11026

FIG. 11A

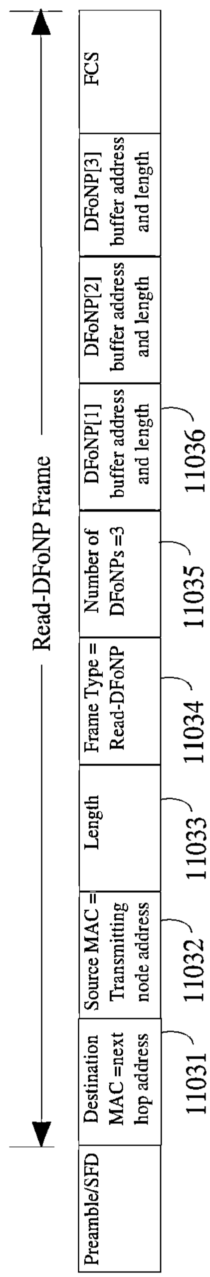

Read-DFoNP Frame
Preamble/SFD
Destination MAC =next hop address
Source MAC = Transmitting node address
Length
Frame Type = Read-DFoNP
Number of DFoNPs =3
DFoNP[1] buffer address and length
DFoNP[2] buffer address and length
DFoNP[3] buffer address and length
FCS
11031
11032
11033
11034
11035
11036

FIG. 11B

Start – 8 bits	Sequence number – 16 bits	Reserved 1 bit	Fmt – 2 bits = 10	Type – 5 bits = 01010	Reserved 1 bit	TC – 3 bits	Reserved – 4 bits	TD – 1 bit = 1
EP – 1 bit	Attr – 2 bits = from Memory Read Request	Reserved 2 bits	Length – 10 bits	Completor ID – 16 bits = ID of the root bridge	Compleion Status – 3 bits	BCM – 1 bit	Byte Count = remaining byte count – 12 bits	Requestor ID – ID of the node reading DFoNP
Tag – 8 bits from Memory Read Request	Reserved – 1 bit	Lower Address – 7 bits	Frame Type = DFoNP	Layer 3/4 Protocol Information	Data	ECRC	LCRC – 32 bits	End – 8 bits

12001 12002 12003

FIG. 12A

Start – 8 bits	Sequence number – 16 bits	Reserved 1 bit	Fmt – 2 bits = 10	Type – 5 bits = 01010	Reserved 1 bit	TC – 3 bits	Reserved – 4 bits	TD – 1 bit = 1
EP – 1 bit	Attr – 2 bits = from Memory Read Request	Reserved - 2 bits	Length – 10 bits = DFoNP size in DW	Completor ID – 16 bits = ID of the root bridge	Compleion Status – 3 bits	BCM – 1 bit	Byte Count = remaining byte count – 12 bits	Requestor ID – ID of the node reading DFoNP
Tag – 8 bits from Memory Read Request	Reserved – 1 bit	Lower Address – 7 bits	Frame Type = STPIs	Number of STPIs = 3	STPI[1], STPI[2], STPI[3]	ECRC	LCRC – 32 bits	End – 8 bits

12011 12012 12013

Expanded view of STPI[2]

Final Destination Node Identifier	Source Node Identifier	Destination STPI Address	Destination DFoNP Address	DFoNP Current Node Identifier	DFoNP Length,Current Buffer Address

12021 12022 12023 12024 12025 12026

FIG. 12B

Start – 8 bits	Sequence number – 16 bits	Reserved 1 bit	Fmt – 2 bits = 1x	Type – 5 bits = 00000	Reserved 1 bit	TC – 3 bits	Reserved – 4 bits	TD – 1 bit = 1
EP – 1 bit	Attr – 2 bits	Reserved 2 bits	Length – 10 bits	Requestor ID – 16 bits = ID of the switch	Tag – 8 bits	Last DW Byte Enable – 4 bits	First DW – Byte Enable – 4 bits	Address 30 or 62 bits
Reserved – 2 bits	Frame Type = DFoNP	Layer 3/4 Protocol Information	Data	ECRC	LCRC – 32 bits	End – 8 bits		

12031 12032 12033

FIG. 12C

Start – 8 bits	Sequence number – 16 bits	Reserved 1 bit	Fmt – 2 bits = 1x	Type – 5 bits = 00000	Reserved - 1 bit	TC – 3 bits	Reserved – 4 bits	TD – 1 bit = 1
EP – 1 bit	Attr – 2 bits	Reserved 2 bits	Length – 10 bits	Requestor ID – 16 bits = ID of the switch	Tag – 8 bits	Last DW Byte Enable – 4 bits	First DW – Byte Enable – 4 bits	Address 30 or 62 bits
Reserved – 2 bits	Frame Type = STPIs	Number of STPIs = 2	STPI[1], STPI[2]	ECRC	LCRC – 32 bits	End – 8 bits		

12041 12042 12043

Expanded view of STPI[2] corresponding to the DFoNP

Final Destination Node Identifier	Source Node Identifier	Misc	DFoNP Current Node Identifier	DFoNP Length,Current Buffer Address

12051 12052 12053 12054 12055

FIG. 12D

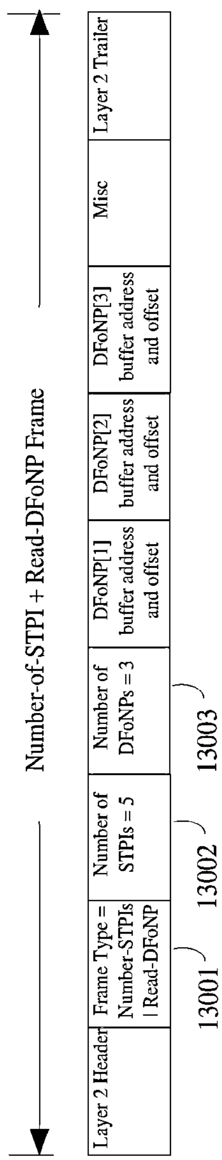


FIG. 13A

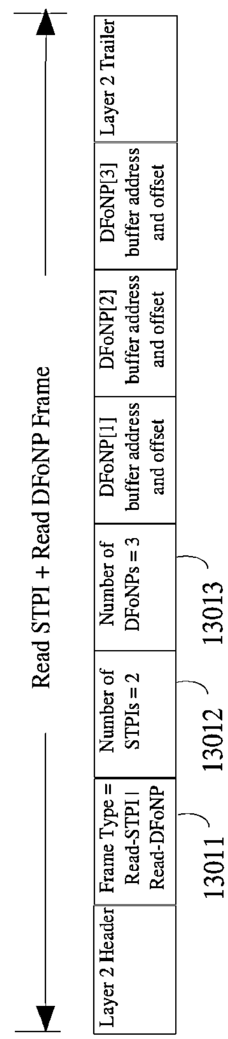


FIG. 13B

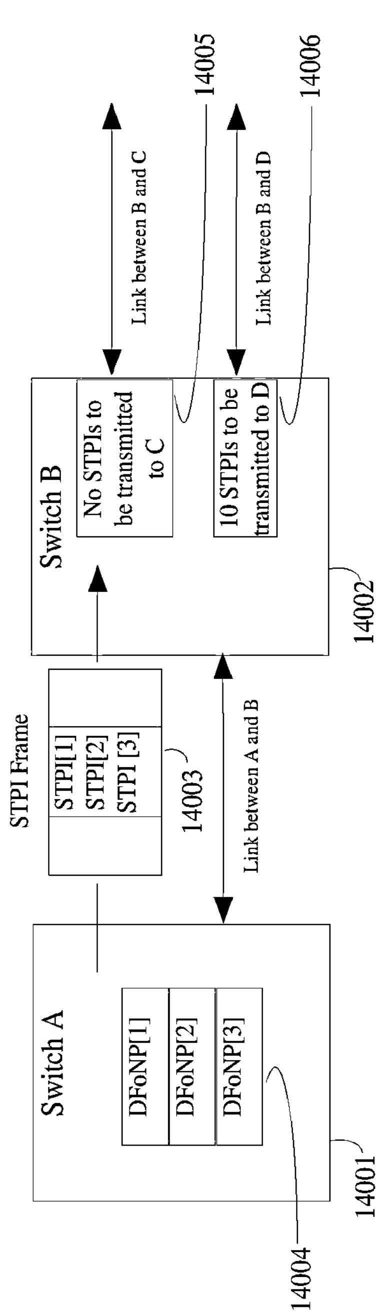


FIG. 14A

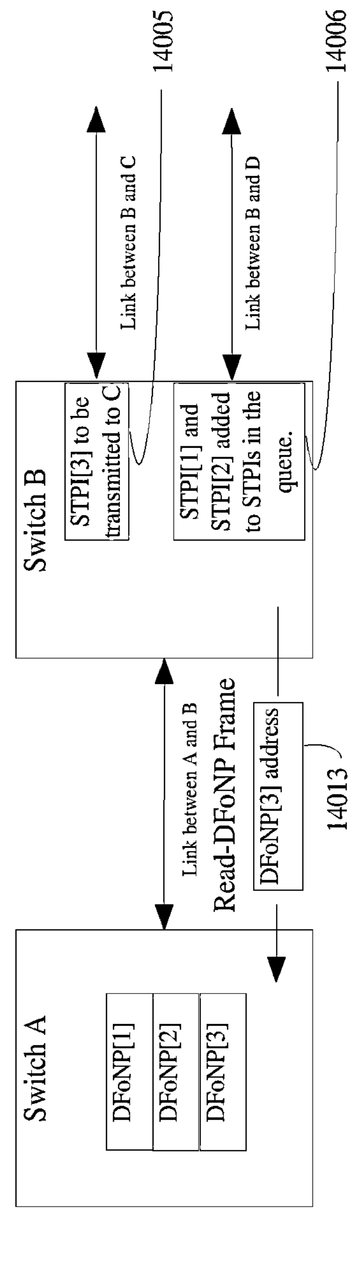


FIG. 14B

RDMA IN DATA CENTER APPLICATIONS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application titled "RDMA in Data centers" is a continuation of the pending application titled "Remote Direct Memory Access (RDMA) over Converged Ethernet with congestion reduction for Prevention of Loss of Data" _filed as U.S. application Ser. No. 18/991,816 on Dec. 23, 2024 which is a continuation of U.S. application Ser. No. 18/900,714 filed on Sep. 28, 2024 titled 'Congestion Elimination in Networks', which issued as U.S. Pat. No. 12,218,848 on Feb. 4, 2025 which is a continuation of U.S. application Ser. No. 18/648,425 filed on Apr. 28, 2024 titled "DMA in PCI Express Network Cluster" which issued as a U.S. Pat. No. 12,126,537 on Oct. 22, 2024, which is a continuation of U.S. application Ser. No. 18/600,441, filed on Mar. 8, 2024, entitled "PCI Express network Cluster" which is a continuation of U.S. application Ser. No. 18/201,779, filed on May 25, 2023, entitled "A System for Avoiding Layer 2 Network Congestion", now U.S. Pat. No. 11,956,154, issued on Apr. 9, 2024, which is a continuation of U.S. application Ser. No. 17/834,097, filed on Jun. 7, 2022, entitled "Delaying Layer 2 Frame Transmission", now U.S. Pat. No. 11,706,148, issued on Jul. 6, 2023, which is a continuation of U.S. application Ser. No. 17/062,594, filed on Oct. 4, 2020, entitled "Data link Frame Reordering", now U.S. Pat. No. 11,398,985, issued on Jul. 26, 2022, which is a continuation of U.S. application Ser. No. 16/132,427, filed on Sep. 16, 2018, entitled "Network Congestion and Packet Reordering", now U.S. Pat. No. 10,841,227, issued on Nov. 17, 2020 which is a continuation of U.S. application Ser. No. 15/268,729, filed on September 19, 2016, entitled "Networking using PCI Express", now U.S. Pat. No. 10,110,498, issued on Oct. 23, 2018, which is a divisional application of U.S. application Ser. No. 14/120,845, filed on July 1, 2014, entitled "Method for Congestion Avoidance", now U.S. Pat. No. 9,479,442, issued on Oct. 25, 2016, which is a continuation of U.S. application Ser. No. 13/385,155, filed on Feb. 6, 2012, entitled "Method for Identifying Next Hop", now U.S. Pat. No. 8,811,400 issued on August 19, 2014, which is a continuation of U.S. application Ser. No. 11/505,788, filed on Aug. 18, 2006, entitled "Creation and Transmission of Part of Protocol Information Corresponding to Network Packets or Data link Frames Separately", now U.S. Patent No. 8,139,574 issued on Mar. 20, 2012, all of which are incorporated herein by reference in their entirety.

BACKGROUND OF THE INVENTION

[0002] The present invention relates to efficient transfer of data link frame or network packets in a "custom" network. The network is "custom" as all switches and end nodes need to create or process data link frames or packets of special formats.

[0003] The OSI, or Open System Interconnection, model defines a networking framework for implementing protocols in seven layers. Most networking protocols do not implement all seven layers, but only a subset of layers. For example, TCP and IP protocol corresponds to layers 4 (TCP) and 3 (IP) respectively. Network packets contain protocol layer information corresponding to the packet. For example, a TCP/IP packet contains a header with both TCP and IP information corresponding to the packet.

[0004] The physical layer (layer 1) specifies how bits stream is created on a network medium and physical and electrical characteristics of the medium. The data link layer (layer 2) specifies framing, addressing and frame level error detection. For outgoing packets to the network, the datalink layer receives network packets from networking layer (layer 3) and creates datalink frames by adding data link (layer 2) protocol information and passes the frame to the physical layer. For incoming packets from network, data link layer receives data link frames from physical layer (layer 1), removes the data link (layer 2) protocol information and passes network packet to the networking layer. The network layer (layer 3) specifies network address and protocols for end to end delivery of packets.

[0005] Network packets contain protocol layer information corresponding to the packet. FIG. **1**A illustrates a network packet containing **01001** layer 1, **01002** layer 2, **01003** layer 3, **01004** layer 4 headers, **01005** Data and **01008** layer 1, **01007** layer 2, **01006** layer 3 trailers. FIG. **1**B illustrates a network packet with **01011** layer 1, **01012** layer 2 (data link), **01013** layer 3 (networking) and **01014** layer 4 (transport) headers and **01017** layer 1 and **01016** layer 2 trailers and **01015** Data. For each layer, the corresponding header and trailer (if present) together contain all the protocol information required to send the packet/frame to the the consumer of the data in a remote node.

[0006] For example, headers/trailers corresponding to a TCP/IP packet in a 10BaseT Ethernet LAN are:

- **[0007]** i) Physical layer header contains Start-of-Stream Delimiter
- **[0008]** ii) Data link layer header contains Preamble, Start-of-Frame Delimiter, Ethernet Addresses, Length/Type Field etc.
- **[0009]** iii) IP header contains Version, Length, IP Address etc.
- **[0010]** iv) TCP header contains Port Numbers, Window, Flags etc.
- **[0011]** v) Data link layer trailer contains 32 bit FCS
- **[0012]** vi) Physical layer trailer contains End-of-Stream Delimiter.

[0013] When parts of networks get congested and end nodes continue transmitting packets to congested parts of a networks, more and more switches can get congested. This can lead to switches dropping large number of packets, nodes retransmitting the dropped or lost packets and network slowing down.

[0014] U.S. Pat. No. 6,917,620 specifies a method and apparatus for a switch that separates the data portion and the header portion. This method has a disadvantage that overhead and logic for separating the data portion and the header portion and then combining the header portion and the data portion before transmission is required. This method also can not consolidate headers from more than one packet for transmission to the next node or delay packet arrival if the destination path of the packet is congested and therefore, can not avoid congestion.

[0015] According to claim 1)(c) of U.S. Pat. No. 5,140,582, the header portion of a packet is decoded prior to the receipt of full packet to determine the destination node. This invention can help with faster processing of the packet within a switch. This method can not consolidate headers from more than one packet for transmission to the next node or delay packet arrival if the destination path of the packet is congested and therefore, can not avoid congestion.

[0016] U.S. Pat. No. 6,032,190 specifies an apparatus and method of separating the header portion of an incoming packet and keeping the header portion in a set of registers and combining the header portion with the data portion before transmitting the packet. This method has a disadvantage that overhead and logic for separating the data portion and the header portion is required. This method can not consolidate headers from more than one packet for transmission to the next-node or delay packet arrival if the destination path of the packet is congested and therefore, can not avoid congestion.

[0017] U.S. Pat. No. 6,408,001 improves transport efficiency by identifying plurality of packets having common destination node, transmitting at least one control message, assigning label to these packets and removing part or all of header. This method has a disadvantage that switches need to identify messages with common destination node and additional logic to remove header and add label. This method can not delay packet arrival if the destination path of the packet is congested and therefore, can not avoid congestion.

BRIEF SUMMARY OF THE INVENTION

[0018] It is the object of the present invention to create and transmit part of protocol information separately from the Datalink Frame or Network Packet (DFoNP) containing data. The Separately Transmitted Protocol Information is referred to as STPI. Network congestion can be reduced or avoided using STPI.

[0019] According to the invention, there should be at least one DFoNP which contains the data and rest of the protocol information not contained in STPI, corresponding to each STPI. Preferably, there will be only one DFoNP corresponding to each STPI. The STPI and DFoNP together contain all the protocol information required to send the packet/frame to the the consumer of the data in a remote node.

[0020] The creation of STPI and DFoNP is done by the originator of the frame or packet such as an operating system in an end node. The format (contents and location of each information in a frame or packet) of the frame or packet containing STPI and DFoNP should be recognized by the final destination of the frame or packet. The format of STPI and DFoNP should also be recognized by switches in the network. So preferably, all STPIs and DFoNP in a given network should be of fixed formats.

[0021] Preferably, one or more STPIs are transmitted in a datalink frame or a network packet. The datalink frame containing STPIs is referred to as STPI Frame. The network packet containing STPIs is referred to as STPI packet. The switches in this case should be capable of extracting each STPI in an incoming STPI Frame or STPI packet and forwarding it to the next node in a different STPI Frame or STPI Packet. The switches can add each STPI from an incoming STPI Frame or STPI Packet into an STPI Frame or STPI Packet it creates. Preferably, the layer 2 address in the datalink frame containing multiple STPIs will be the next hop node address.

[0022] Optionally, STPI Frame or STPI Packet contains number of STPIs or length of the STPI frame. Optionally, STPI Frame or STPI Packet contains the offset or position of STPIs in the STPI frame—this is required only if STPIs supported by the network are not of fixed length.

[0023] Optionally, STPI Frame or STPI Packet does not contain the number of STPIs and switches in the network are capable of identifying the number of STPIs from length of the frame as they are of fixed length.

[0024] Preferably, some protocol information contained in STPI may not be contained in the corresponding DFoNP. But protocol information contained in STPI and the corresponding DFoNP need not be mutually exclusive. In this method, the switches obtain both STPI and the corresponding DFoNP before the STPI and the corresponding DFoNP are forwarded. Optionally, STPI need not be forwarded to end node if sufficient protocol information is contained in the corresponding DFONP.

[0025] The proposed invention can be employed for data, control and/or RDMA packets in a network.

[0026] The proposed method allows switches to read the more than one STPI, and then delay obtaining the corresponding DFoNP. The DFoNP may be read or forwarded in a different order compared to the order in which STPI are read or forwarded. This method allows switches to optimize resources and packet/frame forwarding efficiency.

[0027] STPI contain temporary information such as current node or port number of the node containing the corresponding DFoNP. STPI also contains an address of a buffer containing the corresponding DFoNP or an offset in a buffer where the corresponding DFoNP is stored or an index of the corresponding DFoNP in an array. These information help in associating STPI to the corresponding DFoNP. The exact information contained in STPI whether it is an address or an offset or an index or a combination of these is implementation specific.

[0028] Optionally, STPI may contain originating node identifier and a sequence number. Such information can help in reporting errors when STPI or corresponding DFoNP are corrupted or lost.

[0029] Optionally, STPI may contain other vendor specific or DFoNP related miscellaneous information.

[0030] Optionally, DFoNP may contain some information that help in associating itself with corresponding STPI, such as originating node identifier and a sequence number. Preferably, DFoNP sequence number is same as the sequence number of the corresponding STPI.

[0031] Optionally, DFoNP may contain other vendor specific miscellaneous information.

[0032] The originating node creating an STPI by creating and initializing one or more data structures. Preferably, there is only one data structure containing STPI.

[0033] A switch receiving both frame containing STPI and the DFoNP before forwarding a frame containing STPI or DFoNP to the next switch or node.

[0034] Preferably, a switch receiving frame containing STPI before reading the corresponding DFONP.

[0035] A switch can delay transmitting or reading DFoNP after the corresponding STPI is transmitted or received, allowing the switch to optimize its resource usage and improve efficiency.

[0036] A switch can read DFONPs corresponding to a switch port with minimum outbound traffic, ahead of other DFONPs, thereby improving link efficiency.

[0037] The switch modifying temporary information in STPI such as node number or port number corresponding to the node containing corresponding DFoNP and buffer pointer or index or offset for the corresponding DFoNP, when the DFoNP is transmitted to another node.

[0038] If the DFoNP and STPI is forwarded to another subnet, layer 2 information in STPI and DFoNP should be updated to be compatible with the subnet to which it is forwarded (for example, in an IP network when a packet moves from Ethernet to ATM, layer 2 protocol information will have to be modified to be made compatible with ATM network).

[0039] If STPI contains a multicast or broadcast destination address, the switch transmitting both the DFONPs and the STPI to all next hop nodes identified by the address.

[0040] A switch can delay reading or forwarding the DFoNP after the corresponding STPI is received or forwarded, and vice versa.

[0041] A switch may or may not receive or transmit DFONPs in the same order as the corresponding STPIs are received or transmitted from a switch port.

[0042] Optionally, a switch may receive or transmit one or more DFoNP in one frame.

[0043] For networks that support layer 5/6/7 (example OSI networks), STPI optionally containing part of or all of layer 5/6/7 information. Preferably, no layer 5/6/7 information may be contained in STPI.

BRIEF DESCRIPTION OF THE SEVERAL VIEWS OF THE DRAWINGS

[0044] FIG. 1A illustrates a network packet containing layer 1, layer 2, layer 3, layer 4 headers, Data and layer 1, layer 2, layer 3 trailers.

[0045] FIG. 1B illustrates a network packet with layer 1, layer 2 (datalink), layer 3 (networking) and layer 4 (transport) headers and layer 1 and layer 2 trailers and Data.

[0046] FIG. 2A illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0047] FIG. 2B illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0048] FIG. 2C illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0049] FIG. 2D illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0050] FIG. 2E illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0051] FIG. 2F illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0052] FIG. 2G illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0053] FIG. 2H illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0054] FIG. 2I illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0055] FIG. 2J illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0056] FIG. 2K illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0057] FIG. 2L illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0058] FIG. 2M illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0059] FIG. 2N illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs.

[0060] FIG. 3A illustrates Switch/Node A containing an STPI and the corresponding DFoNP to be transmitted to the Switch/Node B.

[0061] FIG. 3B illustrates the Switch/Node A sending an STPI frame containing the STPI.

[0062] FIG. 3C illustrates the Switch/Node B deciding to fetch the DFoNP corresponding to the STPI and sending Read-DFoNP Frame to the Switch/Node A containing the Read-DFoNP request for the DFONP.

[0063] FIG. 3D illustrates the Switch/Node A responding to the Read-DFoNP request for the DFoNP by sending the DFONP.

[0064] FIG. 3E illustrates the STPI being updated with the identifier of the Switch/Node B and the location of the DFoNP in the Switch/Node B.

[0065] FIG. 4A illustrates Switch/Node A containing an STPI and the corresponding DFoNP to be transmitted to the Destination Node B.

[0066] FIG. 4B illustrates the Switch/Node A transmitting an STPI Frame containing the STPI to the Switch/Node B.

[0067] FIG. 4C illustrates the Switch/Node A transmitting the DFoNP to the Switch/Node B.

[0068] FIG. 4D illustrates the Switch/Node B updating the STPI with the location of the DFoNP in the Switch/Node B.

[0069] FIG. 5A illustrates Switch/Node A containing an STPI and the corresponding DFoNP to be transmitted to the Switch/Node B.

[0070] FIG. 5B illustrates Switch/Node A transmitting a frame containing the STPI to the Switch/Node B.

[0071] FIG. 5C illustrates the Switch/Node B deciding to fetch the DFoNP corresponding to the STPI and sending Read-DFoNP Frame to the Switch/Node A containing DFoNP request for the DFONP.

[0072] FIG. 5D illustrates the Switch/Node A responding to the Read-DFoNP request by transmitting the DFONP.

[0073] FIG. 5E illustrates the STPI being updated with identifier of Switch/Node B and the location of the corresponding DFoNP in the Switch/Node B.

[0074] FIG. 6A illustrates Switch/Node A containing an STPI and the corresponding DFoNP to be transmitted to the Switch/Node B.

[0075] FIG. 6B illustrates the Switch/Node A responding by sending an STPI frame containing all STPIs to be transmitted to the Switch/Node B.

[0076] FIG. 6C illustrates the Switch/Node A transmitting the DFoNP corresponding to the STPI to the Switch/Node B.

[0077] FIG. 6D illustrates the STPI being updated with identifier of the Switch/Node B and the location of the corresponding DFoNP in the Switch/Node B.

[0078] FIG. 7A illustrates Switch/Node A containing an STPI and the corresponding DFoNP to be transmitted to the Destination End Node B.

[0079] FIG. **7**B illustrates Switch/Node A transmitting the DFoNP to the Destination End Node B and updating the STPI with the location (DMA address) of the DFoNP in the Destination End Node B.

[0080] FIG. **7**C illustrates Switch/Node A transmitting the STPI in an STPI frame to the Destination End Node B.

[0081] FIG. **7**D illustrates that both STPI and DFoNP are received by End Node B.

[0082] FIG. **8**A illustrates a Read-STPI frame with Frame Type "Read-STPI" and "Number of STPIs" set to 3.

[0083] FIG. **8**B illustrates a Read-STPI frame in a network where explicit frame type specification is not required.

[0084] FIG. **8**C illustrates a Read-STPI frame in a network without layer 1 headers or trailers.

[0085] FIG. **8**D illustrates a Read-STPI frame in a network without layer 1 headers or trailers.

[0086] FIG. **9**A illustrates a Read-DFoNP frame with Frame Type "Read-DFoNP" and "Number of Read-DFoNP requests" set to 2.

[0087] FIG. **9**B illustrates a Read-DFoNP frame in a network where explicit frame type specification is not required.

[0088] FIG. **9**C illustrates Read-DFoNP frame in a network without layer 1 headers or trailers.

[0089] FIG. **9**D illustrates a Read-DFoNP frame in a network without layer 1 headers or trailers.

[0090] FIG. **10**A illustrates a Number-of-STPIs frame with Frame Type "Number-of-STPIs" and "Number of STPIs" set to 3.

[0091] FIG. **10**B illustrates Number-of-STPIs frame in a network where explicit frame type specification is not required.

[0092] FIG. **10**C illustrates Number-of-STPIs frame in a network without layer 1 headers or trailers.

[0093] FIG. **10**D illustrates a Number-of-STPIs frame in a network without layer 1 headers or trailers.

[0094] FIG. **11**A illustrates an example of DFoNP and STPI frames which can be used with Ethernet.

[0095] FIG. **11**B illustrates Read-DFoNP frame which can be used with Ethernet.

[0096] FIG. **12**A illustrates format of PCI Express Read Completion containing DFoNP, from a root bridge in response to a Memory Read request from a switch.

[0097] FIG. **12**B illustrates format of PCI Express Read Completion containing STPIs, from a root bridge in response to a Memory Read request from a switch.

[0098] FIG. **12**C illustrates a PCI Express Memory Write transaction containing DFoNP, from a switch to a root bridge.

[0099] FIG. **12**D illustrates a PCI Express Memory Write transaction containing STPIs, from a switch to a root bridge.

[0100] FIG. **13**A illustrates a frame containing both Number-of-STPIs message and Read-DFoNP requests.

[0101] FIG. **13**B illustrates a frame containing both Read-STPI request and Read-DFoNP requests.

[0102] FIG. **14**A illustrates Switch A has 3 DFONPs to be transmitted to Switch B.

[0103] FIG. **14**B illustrates the switch identifying that STPI[1] and STPI[2] received are for node D and adding STPI[1] and STPI[2] to the queue for the node D.

DETAILED DESCRIPTION OF THE INVENTION

[0104] There are a very large number of design options with network component designers with respect to the format of DFoNP, STPI and STPI frame/packet. FIG. **2**A, FIG. **2**B, FIG. **2**C, FIG. **2**D, FIG. **2**E, FIG. **2**F, FIG. **2**G, FIG. **2**H, FIG. **2**I, FIG. **2**J, FIG. **2**K, FIG. **2**L, FIG. **2**M and FIG. **2**N illustrate some examples of different formats in which the STPI and the corresponding DFoNP can be created adhering to this invention. The layer 2, layer 3, and layer 4 information that may be present in the DFoNP and STPI may or may not be mutually exclusive and is dependent on specific format or formats of STPI and DFoNP supported by switches and endnodes. Each network will employ only few STPI/DFoNP formats (preferably, as few as 1-3), one each for a subtype of a packet or a frame. Preferably, a network may employ only one format for STPI and one format for DFoNP to reduce complexity in switches and endnodes. STPI should have enough information for the switch to find the port for the next hop.

[0105] i) FIG. **2**A illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. All layer 2 **02021 02024** (including Destination Node Address used for routing), layer 3 **02022** and layer 4 **02023** information are in STPI and the DFoNP contains no layer 3 and 4 information. DFoNP contains minimal layer 2 **02001 02004** information mandated by datalink layer (an example of optional layer 2 information is the VLAN tag in Ethernet). Frame Type in the frame gives the type of frame, DFoNP **02002**, STPI **02012**, etc. All data **02003** are in DFoNP. Three STPIs **02013** are sent in a STPI Frame. The destination address **02011** of the STPI Frame is the next hop switch or node address. In this example, 3rd STPI **02014** in the STPI Frame corresponds to the DFoNP shown. The STPI contains the length **02026** of the corresponding DFoNP and the current node number **02025** and current buffer address **02026** containing the corresponding DFoNP. When the DFoNP is transmitted to the next node the node number **02025** and buffer address **02026** in the corresponding STPI are updated.

[0106] ii) FIG. **2**B illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. The frames in this network do not have layer 2 trailer. All layer 2 **02051** (includes destination node address for routing), RDMA address **02051** for STPI in the destination node, RDMA address **02054** for DFoNP in the destination node, layer 3 **02052** and layer 4 **02053** information are in STPI. The DFoNP contains no layer 3 and 4 information. In this network, layer 2 **02031 02041** contains frame type and hence, no additional field for frame type is present. DFoNP contains layer 2 header **02031** with next hop node address. STPI contains the node number **02055** and an index **02056** to the array containing the corresponding DFoNP. When DFoNP is transmitted to the next node, the node number **02055** and the index **02056** in the corresponding STPI are updated. STPI also contains Source Node Number **02057** (the node number of the node which created the STPI) and STPI sequence number **02058**. The STPI **02042 02043** is the only STPI in the STPI Frame.

[0107] iii) FIG. **2**C illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. All layer 3 **02081** and layer 4 **02082** information are in STPI and the DFoNP contains all **02061** layer 2 information. In this network, switches use **02081** layer 3 address to find next hop port. So **02071** layer 2 of STPI Frame does not have next hop node address. Frame Type in the frame gives the type of frame, DFoNP **02062**, STPI **02072**, etc. There are 2 STPIs **02073** in the STPI Packet and the first STPI **02074** corresponds to DFoNP. STPI contains the DFoNP Current Node Port Number **02083** corresponding to the node containing DFoNP and an offset **02084** in a buffer to the current location of the corresponding DFoNP. The port number **02083** is the port number on the switch containing STPI. When DFoNP is transmitted to the next node, the port number **02083** and offset **02084** in the corresponding STPI are updated. The port number **02083** is also updated-when STPI is transmitted to the next node. STPI also contains Source Node Number **02085** and a sequence number **02086**.

[0108] iv) FIG. **2**D illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. STPI contains layer 3 **02112**, layer 4 **02113**, and part of layer 2 **02111** protocol information (including route to the destination), RDMA address **02111** for STPI in the destination node. DFoNP contains data **02093**, part of layer 2 protocol information **02091 02096** and RDMA address **02091** for the DFoNP in the destination node. STPI contains **02115** DFoNP length and the port number **02114** and the buffer address **02115** to the location of the corresponding to DFONP. When DFoNP is transmitted to the next node, the port number **02114** is reset (as DFONP is in the same node) and buffer address **02115** in the corresponding STPI are updated. DFoNP Port number **02114** is also updated when STPI is transmitted to the next node. Both STPI and DFoNP contains originating node number **02116 02094** and STPI sequence number **02117 02095**. The address in the datalink header **02101** of the STPI Frame is the final destination node address in the subnet indicating all STPIs in the STPI Frame are to the same final destination and switching can be done using STPI Frame address. Frame Type in the frame gives the type of frame, DFoNP **02092**, STPI **02102**, etc. STPI Frame does not contain the number of STPIs as STPIs are of fixed length and the number of STPIs can be derived from the length of STPI frame. The first STPI **02103** in the frame corresponds to the DFoNP shown.

[0109] v) FIG. **2**E illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. STPI contains part of **02143** layer 2 (Layer 2 in STPI contains destination address used for routing), RDMA address **02143** for STPI in the destination node, **02144** part of layer 3 information and all of **02145** layer 4 information. The DFoNP contains **02121** layer 2 protocol information, RDMA address **02121** for DFoNP in the destination node and **02123** part of layer 3 information. Frame Type in the frame gives the type of frame, DFoNP **02122**, STPI **02132**, etc. STPI corresponding to the DFoNP shown is the first STPI **02133** in the STPI Frame. STPI contains the current node number **02146** and index **02147** to the location of the corresponding to DFoNP. When DFoNP is transmitted to the next node, the node number **02146** and index **02147** in the corresponding STPI are updated. STPI also contains Source Node Number **02141**, STPI Sequence Number **02142** and miscellaneous **02148** information. The layer 2 header **02131** of the STPI frame contains next hop node address.

[0110] vi) FIG. **2**F illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. The network medium does not support layer 1 header or trailer. STPI contains part of layer 2 **02173** (including destination node identifier used for routing) and part of layer 3 **02174** protocol information. DFoNP contains layers 2 **02151**, part of layer 3 **02153** and all of layer 4 **02154** protocol information. STPI contains the buffer address **02175** and an index **02175** in the buffer to the location of the corresponding to DFoNP. When DFoNP is transmitted to the next node, buffer address **02175** and offset **02175** in the corresponding STPI are updated. STPI also contains Source Node Number **02171**, STPI sequence number **02172** and miscellaneous **02176** information. Frame Type in the frame gives the type of frame, DFoNP **02152**, STPI **02162**, etc. The STPI Frame contains length **02163** of STPIs and since STPIs of this network are of fixed length, the position of the STPIs in the frame can be determined by the switch. Expanded view of the second STPI **02164** in the STPI frame is shown. The layer 2 header **02161** of the STPI frame contains next hop node address.

[0111] vii) FIG. **2**G illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. The network medium does not support layer 1 header or trailer. STPI contains part of layer 2 **02203** (including destination node address for routing), part of layer 3 **02204** and part of layer 4 **02202** protocol information. DFoNP contains layer 2 **02181**, part of layer 3 **02183** and part of layer 4 **02184** protocol information. STPI contains the current node number **02205**, an index to a buffer **02206** and an offset **02206** in the buffer to the location of the corresponding DFoNP. When DFoNP is transmitted to the next node, the node number **02205**, the index **02206** and the offset **02206** in the corresponding STPI are updated. STPI also contains the Source Node Number **02201** and miscellaneous **02207** information. Frame Type in the frame gives the type of frame, DFoNP **02182**, STPI **02192**, etc. The STPI Frame contains length **02193** of STPIs and since STPIs of this example are of fixed length, the position of the STPIs in the frame can be determined by the switch. Expanded view of the second STPI **02194** in the frame is shown. The layer 2 header **02191** of the STPI frame contains next hop node address.

[0112] viii) FIG. **2**H illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. STPI contains part of layer 2 **02233** (including destination node address for routing) and all of layer 3 **02234** protocol information. The DFoNP contains layer 2 **02211** and layer 4 **02213** protocol information. STPI contains the length **02235** of the corresponding DFoNP and the current node identifier **02235**, buffer address **02236** and an offset **02236** in a buffer to the location of the corresponding DFoNP. When DFoNP is transmitted to the next node, the Current Node identifier **02235**, buffer address

02236 and the offset **02236** in the corresponding STPI are updated. STPI also contains Source Node Number **02231** and STPI Sequence Number **02232**. Frame Type in the frame gives the type of frame, DFoNP **02212**, STPI **02222**, etc. The STPI Frame in this example is allowed to have only one STPI **02223**. The layer 2 header **02221** of the STPI frame contains next hop node address. Expanded view of the STPI is shown.

[0113] ix) FIG. **2**I illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. The network supports protocol layers 5, 6 and 7 in addition to lower layers. STPI contains **02263** layer 2 and **02264** layer 3 information. The DFoNP contains minimal layer 2 **02241** protocol information allowed by the datalink layer, layer 4, layer 5, layer 6, and layer 7 **02243** protocol information. STPI contains the current node number **02265**, a buffer address **02266** in the node and an offset **02266** in the buffer to the location of the corresponding DFoNP. When DFoNP is transmitted to the next node, the node number **02265**, the buffer address **02266** and the offset **02266** in the corresponding STPI are updated. STPI also contains Source Node Number **02261** and STPI sequence number **02262**. Frame Type in the frame gives the type of frame, DFoNP **02242**, STPI **02252**, etc. The STPI Frame in this example is allowed to have only one STPI **02253** and **02251** layer 2 of the STPI frame contains address of the destination node in the subnet which is used for routing the STPI frame. Expanded view of the STPI is shown.

[0114] x) FIG. **2**J illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. STPI contains layer 2 **02293** protocol information (including destination node address for routing). The DFoNP contains **02271** part of layer 2 and all of layer 3 and layer 4 **02273** protocol information. Frame Type in the frame gives the type of frame, DFoNP **02272**, STPI **02282**, etc. The STPI[1] **02284** is the only STPI **02283** in the STPI Frame. STPI contains the current node number **02294** and the buffer address **02295** in the node to the location of the corresponding DFoNP. When DFoNP is transmitted to the next node, the node number **02294** and the buffer address **02295** in the corresponding STPI are updated. STPI also contains Source Node Number **02291** and STPI Sequence Number **02292**. DFoNP contains Source Node Number **02274** and a DFoNP sequence number **02275** which is different from STPI sequence number. The layer 2 header **02281** of the STPI frame contains next hop node address. Expanded view of the STPI[1] is shown.

[0115] xi) FIG. **2**K illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. STPI contains layer 2 **02323** information (including destination node address for routing). The DFoNP contains minimal layer 2 **02301** mandated by datalink layer of the subnetwork and all of layer 3 and 4 **02302** information. The DFoNP contains control data **02303** such as requests to open a file in addition to data **02303**. In this network, layer 2 **02301** **02311** protocol information contains frame type and hence, no additional field for frame type is present. The STPI[1] **02313** is the only STPI **02312** in the STPI Frame. STPI contains the length **02324** of the corresponding DFoNP and the node number **02324** and the buffer address **02325** in the node to the location of the corresponding to DFoNP. When DFoNP is transmitted to the next node, the node number **02324** and buffer address **02325** in STPI are updated. STPI also contains the Source Node Number **02321** and STPI sequence number **02322**. DFoNP contains Source Node Number **02304** and a DFoNP Sequence Number **02305** which is different from STPI sequence number. Expanded view of STPI[1] is shown.

[0116] xii) FIG. **2**L illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. STPI contains layer 2 **02354** (including destination node address for routing) and layer 3 information **02353** and part of layer 5/6/7 **02357** protocol information. The DFoNP contains minimal layer 2 Header **02331** mandated by datalink layer of the subnet, layer 4 **02333** and part of layer 5/6/7 **02334** protocol information. The DFoNP contains control data **02335** such as requests to open a file in addition to data **02335**. Frame Type in the frame gives the type of frame, DFoNP **02332**, STPI **02342**, etc. The STPI[1] **02344** is the only STPI **02343** in the STPI Frame. STPI contains the node number **02355** and buffer address **02356** in the node to the location of the corresponding DFoNP. When DFoNP is transmitted to the next node, the node number **02355** and buffer address **02356** in the corresponding STPI are updated. STPI also contains the Source Node Number **02351** and STPI sequence number **02352**. The layer 2 header **02341** of the STPI frame contains next hop node address. Expanded view of the STPI[1] **02344** is shown.

[0117] xiii) FIG. **2**M illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. STPI contains layer 2 **02386** (including destination node identifier used for routing), layer 3 **02385** and layers 5/6/7 **02387** protocol information. The DFoNP contains layers 2 **02361**, layer 3 **02363** and layer 4 **02364** protocol information. Frame Type in the frame gives the type of frame, DFoNP **02362**, STPI **02372**, etc. STPI frame contains two STPIs **02373** and expanded view of the 2nd STPI (STPI[2]) **02376** is shown. The STPI frame contains offsets **02374** to all STPIs in the frame. The network in this example supports more than one length for STPIs. STPI[1] offset **02374** gives the location of the first STPI (STPI[1] **02375**) in the STPI frame. STPI[2] offset **02374** gives the location of the second STPI in the STPI frame. Offsets in this example are with respect to beginning of the frame. STPI contains the node number **02381** and buffer address **02382** in the node to the location of the corresponding DFONP. When DFoNP is transmitted to the next node, the node number **02381** and buffer address **02382** in STPI are updated. STPI also contains Source Node Number **02383** and STPI sequence number **02384**. The layer 2 header **02371** of the STPI frame contains next hop node address.

[0118] xiv) FIG. **2**N illustrates example formats for DFoNP, the corresponding STPI and an STPI frame which contain STPIs. STPI contains layer 2 **02415** (including destination node identifier used for routing) protocol information. The DFoNP contains layer 2 **02391**, layers 3 **02394** and layer 4 **02394** protocol information. Frame Type in the frame gives the type of frame, Read Completion **02392** for DFoNP frame and Write **02402** for STPI frame. The STPI Frame contains the length of write **02403** (which is the length of STPI[1] **02404** and STPI[2] **02405**) and address **02403** for the write. DFoNP contains Read Requester ID **02393** (Identifier) and a tag **02393** to identify the read request. DFoNP also contains address **02393** from which the layer 3/4 headers and the data **02395** is read and the length **02393** of the the read. The STPI Frame contains two STPIs and expanded view of the 2nd STPI (STPI[2]) **02405** is shown. STPI contains the node number **02411** and buffer address **02412** in the node to the location of the corresponding DFoNP and the length of the DFoNP **02416**. These information are used to read the corresponding DFoNP. When DFoNP is transmitted to the next node, the node number **02411** and buffer address **02412** in STPI are updated. STPI also contains Source Node Number **02413**, STPI Sequence Number **02414** and Miscellaneous **02416** information. The layer 2 header **02401** of the STPI frame contains next hop node address.

[0119] Below five options for transferring STPI and the corresponding DFoNP from one node to another, are described. One of the first 4 methods can be used for transferring STPI and the corresponding DFoNP from the originating node or a switch to another switch or end node. The fifth method can be used for transferring STPI and the corresponding DFoNP to a destination end node:

[0120] i) FIG. **3**A, FIG. **3**B, FIG. **3**C, FIG. **3**D and FIG. **3**E illustrate one of the options that could be used in a given network for transmitting STPI and DFoNP to the next hop node. In this option a switch/node responds to Read-STPI request by transmitting STPIs. The switch/node receiving STPIs sends Read-DFoNP requests using the information contained in STPIs to fetch the corresponding DFONPs. A frame containing a Read-STPI request is called Read-STPI Frame. A frame containing Read-DFoNP requests is called Read-DFoNP Frame. In FIG. **3**A, Switch/Node A **03001** contains an STPI **03003** and the corresponding DFoNP **03004** to be transmitted to the Switch/Node B **03002**. In FIG. **3**A, the Switch/Node B transmits Read-STPI Frame **03005** to the Switch/Node A giving the maximum number of STPIs that can be transmitted. The maximum number of STPIs **03005** are 5 in the example. In FIG. **3**B, the Switch/Node A responds by sending an STPI frame **03011** containing the STPI **03003** (the STPI frame in this example can contain up to 5 STPIs). In FIG. **3**C, the Switch/Node B decides to fetch the DFoNP corresponding to the STPI **03003** and sends Read-DFoNP Frame **03021** to the Switch/Node A containing the Read-DFoNP request for the DFoNP **03004**. The Read-DFoNP request contains the location (a location could be a buffer address or an offset in a buffer or an index or a combination of addresses, offsets or indexes) of the DFoNP **03004** in the Switch/Node A. The location of the DFoNP to be used in Read-DFoNP request will be present or can be derived from the contents of the corresponding STPI **03003**. In FIG. **3**D, the Switch/Node A responds to the Read-DFoNP request for the DFoNP by sending the DFoNP **03004**. In FIG. **3**E, the STPI **03003** is updated with the identifier of the Switch/Node B and the location of the DFoNP **03004** in the Switch/Node B.

[0121] ii) FIG. **4**A, FIG. **4**B, FIG. **4**C and FIG. **4**D illustrate another option for transmitting STPI and the corresponding DFoNP to the next hop node. In this option, a switch/node transmits STPIs followed by DFONPs corresponding to the STPIs transmitted. In FIG. **4**A Switch/Node A **04001** contains an STPI **04003** and the corresponding DFoNP **04004** to be transmitted to the Destination Node B **04002**. In FIG. **4**B, the Switch/Node A transmits an STPI Frame **04011** containing the STPI **04003** to the Switch/Node B. In FIG. **4**C, the Switch/Node A transmits the DFoNP **04004** to the Switch/Node B. In FIG. **4**D, the Switch/Node B updates the STPI **04003** with the location of the DFoNP **04004** in the Switch/Node B.

[0122] iii) FIG. **5**A, FIG. **5**B, FIG. **5**C, FIG. **5**D and FIG. **5**E illustrate another option for transmitting STPI and the corresponding DFoNP to the next hop node. In this option a switch/node transmits STPIs and the switch/node receiving STPIs sends Read-DFoNP requests using information contained in STPIs to fetch the corresponding DFONPs. In FIG. **5**A Switch/Node A **05001** contains an STPI **05003** and the corresponding DFoNP **05004** to be transmitted to the Switch/Node B **05002**. In FIG. **5**B Switch/Node A transmits a frame **05011** containing the STPI to the Switch/Node B. In FIG. **5**C, the Switch/Node B decides to fetch the DFoNP corresponding to the STPI and sends Read-DFoNP Frame **05021** to the Switch/Node A containing DFoNP request for the DFoNP **05004**. The DFoNP request contains the location of the DFoNP **05004**. The location of the DFoNP used in the Read-DFoNP request will be present or can be derived from the contents of the corresponding STPI **05003**. In FIG. **5**D, the Switch/Node A responds to the Read-DFoNP request by transmitting the DFoNP **05004**. In FIG. **5**E, the STPI **05003** is updated with identifier of Switch/Node B and the location of the corresponding DFoNP **05004** in the Switch/Node B.

[0123] iv) FIG. **6**A, FIG. **6**B, FIG. **6**C and FIG. **6**D illustrate another option for transmitting STPI and DFoNP to the next hop node. In this option a switch/node responds to Read-STPI request by transmitting STPIs followed by the corresponding DFONPs. In FIG. **6**A Switch/Node A **06001** contains an STPI **06003** and the corresponding DFoNP **06004** to be transmitted to the Switch/Node B **06002**. The Switch/Node B transmits Read-STPI Frame **06005** to the Switch/Node A giving the maximum number of STPIs that can be transmitted. The maximum number of STPIs **06005** is 0 in the example indicating that all STPIs can be transmitted. In FIG. **6**B, the Switch/Node A responds by sending an STPI frame **06011** containing all STPIs to be transmitted to the Switch/Node B. In FIG. **6**C, the Switch/Node A transmits the DFoNP **06004** corresponding to the STPI to the

[0124] Switch/Node B. In FIG. **6**D, the STPI **06003** is updated with identifier of the

[0125] Switch/Node B and the location of the corresponding DFoNP **06004** in the Switch/Node B.

[0126] v) FIG. **7**A, FIG. **7**B, FIG. **7**C, and FIG. **7**D illustrate an option which can be used for transmitting DFoNP and optionally the corresponding STPI from a switch/node to a destination node: In this option DFoNP is transmitted to the destination node and then optionally, the corresponding STPI is transmitted. In FIG. **7**A, Switch/Node A **07001** contains an STPI **07003** and the corresponding DFoNP **07004** to be transmitted to the Destination End Node B **07002**. In FIG. **7**B, Switch/Node A transmits the DFoNP **07004** to the Destination End Node B and updates the STPI **07003** with the location (DMA address) of the DFoNP in the Destination End Node B. In FIG. **7**C, Switch/Node A transmits the STPI in an STPI frame **07021** to the Destination End Node B. In FIG. **7**D, both STPI **07003** and DFoNP **07004** are received by End Node B.

[0127] A switch can employ one of the STPI and DFoNP transfer options (strategies) listed above, for each port. Both ports on a point-to-point link must agree to the same frame transmitting option. All ports on a link or bus must follow the same frame transmitting option. Preferably, a network employs only one of the four STPI/DFoNP transfer options listed in FIG. **3**A to FIG. **3**E, FIG. **4**A to FIG. **4**D, FIG. **5**A to FIG. **5**E, FIG. **6**A to FIG. **6**D. Preferably, a network also employs the STPI/DFoNP transfer option listed in FIG. **7**A to FIG. **7**D. For the option corresponding to FIG. **7**A to FIG. **7**D, updating STPI with address (location) of DFoNP in the end node is optional.

[0128] If DFoNPs do not contain information (such as originating node identifier, DFoNP identifier, DFoNP address in previous node, etc.) that allow a DFoNP to be mapped to the corresponding STPI, then the DFONPs must be transmitted in the same order as requested in Read-DFoNP frame/s with design options listed in FIG. **3**A to FIG. **3**E and FIG. **5**A to FIG. **5**E. With design options listed in FIG. **4**A to FIG. **4**D and FIG. **6**A to FIG. **6**D, if DFONPs do not contain information that allow the DFoNP to be mapped to the corresponding STPI, DFONPs must be transmitted in the same order as the corresponding STPIs are transmitted. This will allow switches to identify STPI corresponding to an DFoNP that is received.

[0129] There are a very large number of design options with network component designers with respect to the format of Read-STPI request and Read-STPI Frames containing Read-STPI request. FIG. **8**A, FIG. **8**B, FIG. **8**C and FIG. **8**D illustrate some examples of different formats in which the Read-STPI Frames can be created adhering to this invention. Preferably a given network employs only one format (design option) for Read-STPI request to keep the design of switches and end nodes simple.

[0130] i) FIG. **8**A illustrates a Read-STPI frame with Frame Type "Read-STPI" **08001** and "Number of STPIs" **08002** set to 3. The frame also contains Miscellaneous **08003** field.

[0131] ii) FIG. **8**B illustrates a Read-STPI frame in a network where explicit frame type specification is not required. The frame specifies an address **08011** for read (the location of the STPIs) in the node receiving the Read-STPI Frame. The frame also provides the length **08012** for read. The address where STPIs are stored can be dynamically configured on the switch for each node/switch it is connected to.

[0132] iii) FIG. **8**C illustrates a Read-STPI frame in a network without layer 1 headers or trailers. Frame Type **08021** is "Read-STPI". The "Number of STPIs" **08022** is 0 indicating permission to transmit an STPI Frame with as many STPIs for the node transmitting Read-STPI Frame as possible, from the node receiving the Read-STPI Frame. The frame also contains a Miscellaneous **08023** field.

[0133] iv) FIG. **8**D illustrates a Read-STPI frame in a network without layer 1 headers or trailers. Layer 2 header **08031** contains Frame Type (Read-STPI). The "Number of STPIs" **08032** is −1 indicating permission to transmit all STPIs for the node transmitting Read-STPI Frame, from the node receiving the Read-STPI Frame.

[0134] A Read-DFoNP Frame contains one or more Read-DFoNP requests and each Read-DFoNP request contains the location of the requested DFoNP. There are a very large number of design options with network component designers with respect to the format of Read-DFoNP requests and Read-DFoNP Frames containing Read-DFoNP requests. FIG. **9**A, FIG. **9**B, FIG. **9**C and FIG. **9**D illustrate some examples of different formats in which the Read-DFoNP Frame can be created adhering to this invention. Preferably, a given network employs only one format (design option) for Read-DFoNP request to keep the design of switches and end nodes simple.

[0135] i) FIG. **9**A, illustrates a Read-DFoNP frame with Frame Type **09001** "Read-DFoNP" and "Number of Read-DFoNP requests" **09002** set to 2. The DFoNP[1] **09003** and DFoNP[2] **09004** buffer addresses provide the location of the DFONPs in the node receiving the Read-DFoNP Frame. The frame also contains Miscellaneous **09005** field.

[0136] ii) FIG. **9**B illustrates a Read-DFoNP frame in a network where explicit frame type specification is not required. Frame specifies an address **09011** for read (the location of the DFoNP) in the node receiving the Read-DFoNP Frame. The frame also provides the length **09012** for read.

[0137] iii) FIG. **9**C illustrates Read-DFoNP frame in a network without layer 1 headers or trailers. Frame Type **09021** is "Read-DFoNP", the "Number of Read-DFoNP requests" **09022** is 3. Each Read-DFoNP request contains a buffer address and an offset. The DFoNP[1] **09023**, DFONP[2] **09024** and DFoNP[3] **09025** buffer addresses and offsets provide the location of the DFONPs in the node receiving the Read-DFoNP Frame.

[0138] iv) FIG. **9**D illustrates a Read-DFONP frame in a network without layer 1 headers or trailers. Frame Type (Read-DFoNP) is contained in layer 2 header **09031**. Only one Read-DFoNP request **09032** is allowed in the frame and the the Read-DFoNP request gives the index of the DFoNP to be read.

[0139] Optionally, a switch or node can send the number of STPIs available for transmission to the next hop node or switch. There are a very large number of design options with network component designers with respect to the format of Number-of-STPIs message and Number-of-STPIs Frames containing Number-of-STPIs message. FIG. **10**A, FIG. **10**B, FIG. **10**C and FIG. **10**D illustrate some examples of different formats in which the Number-of-STPIs Frame can be created adhering to this invention. Preferably a given network employs only one format for Number-of-STPI message to keep the design of switches and end nodes simple.

[0140] i) FIG. **10**A, illustrates a Number-of-STPIs frame with Frame Type **10001** "Number-of-STPIs" and

"Number of STPIs" **10002** set to 3. The frame also contains a Miscellaneous **10003** field.

[0141] ii) FIG. **10**B illustrates Number-of-STPIs frame in a network where explicit frame type specification is not required. Frame specifies an address **10011** to the location where value of Number of STPIs will be written and the length **10012** of the field to be written. The next field contains data (Number of STPIs) **10013** for the write, which is 2.

[0142] iii) FIG. **10**C illustrates Number-of-STPIs frame in a network without layer 1 headers or trailers. Frame Type **10021** is "Number-of-STPIs". The "Number of STPIs" **10022** is 3. The frame also contains a Miscellaneous **10023** field.

[0143] iv) FIG. **10**D illustrates a Number-of-STPIs frame in a network without layer 1 headers or trailers. Layer 2 header **10031** contains Frame Type (Number-of-STPIs). The "Number of STPIs" **10032** is **1**.

[0144] The network described in this invention can be connected to an I/O card (in a server or embedded system) or to a PCI bus.

[0145] i) The switch corresponding to this invention can be connected to an Ethernet card.

[0146] a) A recommended frame format for use with Ethernet is as follows:

[0147] 1) Ethernet header contains destination MAC: The network can use next hop MAC address in the STPI/DFoNP/Read-STPI/Read-DFoNP/Number-of-STPIs frame.

[0148] 2) Ethernet header contains source MAC address: A DFoNP frame can contain the MAC address of the originating node in this field. All other types of frames (STPI, Read-STPI, Read-DFoNP, Number-of-STPI) can contain MAC address of the node transmitting the frame in this field.

[0149] 3) The Ethernet header contains length field as per Ethernet Protocol standard.

[0150] 4) The first byte of the data field contains the "Frame-Type": one bit each for STPI, DFONP, Read-STPI, Read-DFoNP and Number-of-STPIs.

[0151] 5) Each STPI will contain the final destination MAC address. Optionally, each STPI can also contain source MAC address of the the originating node of the STPI.

[0152] 6) The formats specified examples such as FIG. **2**A, FIG. **2**C etc., can be used with Ethernet.

[0153] 7) The Ethernet trailer contains FCS for the frame.

[0154] b) FIG. **11**A illustrates an example of DFoNP and STPI frames which can be used with Ethernet. FIG. **11**B illustrates Read-DFoNP frame which can be used with Ethernet.

[0155] 1) Destination MAC address **11001** in DFoNP frame is the MAC address corresponding to the port or node (next hop node) receiving the frame. If switches are designed to ignore Destination MAC address in a DFoNP frame, the final destination node MAC address could be used in the Destination MAC address field.

[0156] 2) Source MAC address **11002** in the DFoNP frame is the MAC address of the node that created the DFONP.

[0157] 3) The length field **11003** provides the length as per Ethernet Protocol standard.

[0158] 4) The first field in the data portion of Ethernet Frame is Frame Type **11004** and Frame Type of DFoNP frame is DFoNP (DFoNP bit is set).

[0159] 5) The DFoNP contains layer 3 **11005**, layer 4 **11006** protocol information and data **11007**.

[0160] 6) Destination MAC address **11011** in the STPI frame is the MAC address corresponding to the port or node (next hop node) receiving the frame.

[0161] 7) Source MAC address **11012** in the STPI frame is the MAC address corresponding to the port transmitting the frame.

[0162] 8) The length field **11013** provides the length as per Ethernet Protocol standard.

[0163] 9) The first field in the data portion of the Ethernet Frame is Frame Type **11014** and Frame Type of STPI frame is STPI (STPI bit is set).

[0164] 10) The STPI frame in this example contains 2 STPIs **11015**.

[0165] 11) Expanded view of the second STPI **11016** is shown.

[0166] 12) Each STPI contains the Final Destination MAC address **11021** for the STPI and the corresponding DFoNP. Switches can use this address for routing.

[0167] 13) The STPI contains the Source MAC Address **11022** of the Ethernet port through which the STPI entered the Ethernet LAN.

[0168] 14) STPI contains "Destination STPI Address" **11023** which is the address to be used for RDMA Writing the STPI in the destination node.

[0169] 15) STPI contains "Destination DFoNP Address" **11024** which is the address to be used for RDMA Writing the corresponding DFoNP in the destination node.

[0170] 16) The STPI contains the MAC address of the node containing DFoNP **11025**, buffer address **11026** of the DFoNP in this node and length **11026** of the DFoNP. These fields are used to create Read-DFoNP request.

[0171] 17) After an STPI an STPI frame is received, the next hop node can initiate read for the corresponding DFoNP. FIG. **11**B illustrates a Read-DFoNP frame containing 3 Read-DFoNP requests.

[0172] 18) The destination MAC address **11031** in the Read-DFoNP frame is the "DFoNP Current Node MAC address" **11025** from the STPI.

[0173] 19) The source MAC address **11032** in the Read-DFoNP frame is the MAC address corresponding to the port transmitting the Read-DFoNP Frame.

[0174] 20) The length field **11033** provides the length as per Ethernet Protocol standard.

[0175] 21) The first field in the data portion of the Ethernet Frame is Frame Type **11034** and Frame Type of Read-DFoNP frame is "Read-DFoNP" ("Read-DFoNP" bit is set).

[0176] 22) The Number of DFONPs **11035** being requested from the node receiving Read-DFoNP frame is 3 in this example.

[0177] 23) The DFoNP buffer address **11036** and the length **11036** of DFoNP in Read-STPI frame are from DFoNP Current Buffer Address **11026** and DFoNP Length **11026** fields in STPI.

[0178] ii) If the switch corresponding to this invention is connected to a PCI bus, it behaves like an end node. The switch will use PCI transactions to communicate with the server.

[0179] a) The host (in turn the PCI root bridge) can use PCI memory write transaction to transfer STPIs to a switch corresponding to this invention OR the switch can use PCI memory read transaction to read STPIs. The host can use PCI memory write transaction to write the address of the memory location holding STPIs which the switch can use for PCI Memory Read transaction.

[0180] b) The switch can use PCI read transaction to read each DFoNP using the buffer address contained in the corresponding STPI.

[0181] c) The host (in turn the PCI root bridge) can optionally use PCI write transaction to write the number of STPIs to a switch corresponding to this invention.

[0182] d) The switch can use PCI memory write to write DFONPs and STPIs to the memory of the destination node.

[0183] e) FIG. **12**A, FIG. **12**B, FIG. **12**C and FIG. **12**D illustrate an example of transaction formats which can be used within PCI Express TM (PCI Express TM is a trade mark of PCI-SIG) transactions for transferring STPIs and DFONPs from root bridge to a switch corresponding to this invention and vice versa.

[0184] 1) Example in FIG. **12**A illustrates format of PCI Express Read Completion containing DFoNP, from a root bridge in response to a Memory Read request from a switch. The first field of PCI Express Read Completion data provides the Frame Type **12001** which is DFoNP. The rest of the Read Completion data is layer 3/4 protocol information **12002** and Data **12003** being transmitted to the remote node.

[0185] 2) Example in FIG. **12**B illustrates format of PCI Express Read Completion containing STPIs, from a root bridge in response to a Memory Read request from a switch. The first field of data provides the Frame Type **12011** which is STPI. The second field in data is "Number of STPIs" **12012** which is 3 followed by three STPIs **12013**. Each STPI contains "Final Destination Node Identifier" **12021** which is used by switches for routing, Source Node Identifier **12022** which is the identifier of the node that created the STPI, "Destination STPI Address" **12023** to be used for RDMA Writing STPI in the destination, "Destination DFoNP Address" **12024** to be used for RDMA Writing the corresponding DFoNP in the destination, "DFoNP Current Node ID" **12025**, DFoNP Length and DFoNP Current Address **12026** to be used for reading DFoNP from the node where it is currently stored. The DFoNP Length field **12026** is also used for RDMAing DFoNP to the memory of the destination node.

[0186] 3) Example in FIG. **12**C illustrates a PCI Express Memory Write transaction containing DFoNP, from a switch to a root bridge. The first field of PCI Express Memory Write transaction data provides the Frame Type **12031** which is DFoNP. The rest of the Read Completion data is layer 3/4 information **12032** and Data **12033** that arrived from the remote node.

[0187] 4) Example in FIG. **12**D illustrates a PCI Express Memory Write transaction containing STPIs, from a switch to a root bridge. The first field of PCI Express Memory Write data provides the Frame Type **12041** which is STPI. The second field in the data is "Number of STPIs" **12042** which is 2 followed by two STPIs **12043**. Each STPI contains "Final Destination Node Identifier" **12051** which is used by switch for routing, Source Node Identifier **12052** which is the identifier of the node that created the STPI, a miscellaneous field **12053**, "DFoNP Current Node Identifier" **12054**, DFoNP Current Buffer Address **12055** and DFoNP Length **12055** to be used for reading DFoNP from the node where it is currently stored. The DFoNP Length field **12055** is also used for doing PCI Express Memory Write transaction to the root bridge (DMAing DFoNP to the memory of the destination node). The DFoNP and STPI are DMAed into read buffers provided by the destination node.

[0188] When destination address contained in an STPI is a Multi-cast and Broadcast address, both STPI and DFoNP are transmitted to all next hop nodes identified by the Multi-cast or Broadcast address.

[0189] When STPI or DFoNP frames are corrupted or lost, switches and nodes may employ retransmission of the corrupted or lost frame. The retransmission policy and error recovery are link (example PCI) and vendor specific.

[0190] Some networks allow more than one type of content to be present in the same frame. The types of contents are STPI, DFoNP, Read-STPI request, Read-DFoNP request and Number-of-STPIs message.

[0191] i) FIG. **13**A illustrates a frame containing both Number-of-STPIs message and Read-DFoNP requests. The Frame Type **13001** is a bit-OR of "Number-of-STPIs" and "Read-DFoNP". The "Number of STPIs" **13002** is 5 indicating that there are 5 STPIs available to be transmitted to the receiving node. The "Number of DFONPs" **13003** is 3 and the receiving node is expected to respond to the request by transmitting the three DFONPs requested.

[0192] ii) FIG. **13**B illustrates a frame containing both Read-STPI request and Read-DFoNP requests. The Frame Type **13011** is a bit-OR of "Read-STPI" and "Read-DFoNP". The "Number of STPIs" field **13012** is 2 and the "Number of DFONPs" field **13013** is 3. The node receiving the frame is expected to respond with two STPIs and the three requested DFONPs.

[0193] FIG. **14**A and FIG. **14**B illustrate an example of reading DFONPs in a different order compared to the order in which STPIs are received. In FIG. **14**A, Switch A **14001** has 3 DFONPs **14004** to be transmitted to Switch B **14002**. The Switch A forwards 3 STPIs corresponding to the

DFONPs in an STPI frame **14003** to Switch B. The Switch B has **10** STPIs in its queue **14006** for its link to node D. The switch B has no STPIs in its queue **14005** for its link to node C. In FIG. **14**B, the switch identifies that STPI[1] and STPI[2] received are for node D and adds STPI[1] and STPI[2] to the queue **14006** for the node D. The Switch B delays reading DFoNP[1] and DFoNP[2] since there are a large of STPIs already queued for the node D. The Switch B identifies that STPI[3] received is for the node C and queues STPI[3] to the queue **14005** for the node C. The Switch B sends Read-DFoNP Frame **14013** to the Switch A with DFoNP[3] address.

[0194] If STPI contains a priority or QoS field, a switch can use it for controlling the order in which DFONPs are read. Similarly, a priority or QoS field in STPI or DFoNP could be used by switches or nodes to control the order in which STPIs are transmitted to the next node.

[0195] A network corresponding to this invention could be used to connect a server or servers to storage devices (such as disks, disk arrays, JBODs, Storage Tapes, DVD drives etc.). iSCSI and iSER (ISCSI Extensions for RDMA) are examples in which SCSI commands and SCSI data are transmitted using networks technologies used for server interconnect.

Advantages

[0196] A switch can delay receiving DFoNP for paths which are already congested.

[0197] A switch can read DFoNP corresponding to a lightly loaded link ahead of other DFONPs and transmit STPI and DFoNP more quickly to the lightly loaded link improving link efficiency.

[0198] A switch can delay reading DFONPs based on QoS or priority field in STPI.

[0199] A switch can optimize switch resources, memory and frame/packet queues as congestions are minimized by delaying DFONPs for ports which are already congested.

[0200] The switch can ensure higher throughput on all links by rearranging order in which DFONPs are read.

What is claimed is:

1. A network system, comprising a plurality of interconnected network nodes in a data center;

the plurality of interconnected network nodes in the data center comprising:

a plurality of network switches, distributed within the network system, wherein each of the plurality of network switches are configured for interconnecting network nodes in the network and for forwarding data to the connected network nodes;

a plurality of connected end nodes comprising at least a processor and at least a data memory interconnected by the network switches, wherein the end nodes are enabled to be source nodes or destination nodes for data in the network;

the plurality of network switches comprising at least a first network switch at a first network node connecting to a second network switch at a second network node;

a first end node at a third network node connecting to the first network switch and a second end node at a fourth network node connected to the second network switch;

the first end node configured as a source node for a data stored in a sorce memory and the second end node configured as a destination node for the data to be received and stored in a destination memory;

the data from the source memory at the source node configured to be transferred over the network and stored in the destination memory at the destination node;

the plurality of network switches enabled to use per flow congestion control (PFC) to avoid data loss of priority data due to congestion at a loaded queue indicative of congestion at any of the plurality of network switches in the network including the first and the second network switches;

wherein the first and the second network switches are configured to implement a method for data transfer from the source memory at the third network node and the destination memory at the fourth network node; the method comprising:

the first network switch receiving a source address and a destination address for a data stored in the source memory in the third network node which is the source nod;

the first network switch retrieving the data from the source address in the source memory using Remote Direct Memory Access;

the first network switch forwarding the received destination address and data to the second switch over the connected network;

the second switch forwarding the data to the fourth network node, which is the destination node and writing the data to the destination memory in the fourth network node using remote direct memory access.

2. The network system of claim **1**, wherein PFC comprise sending by a switch at a node a request to the prior node to delay or pause transmission of data to a loaded queue at the switch, indicative of a congestion at the queue at the node; and

the prior node responding to the request by delaying or pausing for a time the transmission of data to be stored in the loaded queue, thereby reducing the congestion at the loaded queue and avoiding a loss of data.

3. The system of claim **2**, wherein the plurality of network switches configured for interconnecting network nodes in the network and for forwarding data to the connected network nodes carry data flows of differing priorities over the same network and wherein different queues at each port of a switch are configured to store and forward the data flows of differing priorities.

4. The system of claim **3**, wherein more than one queue is used to store data based on differing priorities associated with the data; and

multiple different data priorities are enabled to be stored in a single queue.

5. The system of claim **2**, wherein Remote Direct Memory Access enable bypassing of the processor and directly accessing the source memory and the destination memory at end nodes for data retrieval and storage.

6. The system of claim **5**, wherein bypassing the processors at the end nodes reduce the latency associated with data retrieval, data transfer and data storage within the network system.

7. A network system of a plurality of interconnected network nodes for enabling remote direct memory access (RDMA) in a data center;

the plurality of interconnected network nodes in the data center comprising:

a plurality of network switches, distributed within the network system at the plurality of network nodes:

a plurality of end nodes, each end node comprising at least a processor and at least a data memory interconnected by the network switches;

wherein the end nodes are configured to act as source nodes and destination nodes for data;

the end nodes comprising at least a processor and at least a memory configurable as a source memory or a destination memory for data:

wherein each of the plurality of network switches is configured for interconnecting the network nodes in the network for forwarding data;

wherein each of the plurality of network switches is also configured for retrieving data from the source memory at the source nodes and storing data to a destination memory at the destination nodes connected to the network switch using the remote direct memory access (RDMA) technology;

the plurality of network switches comprising at least a first network switch at a first node interconnecting to a first end node, that is configured as a first source node, at a second network node; and connecting to a second end node, that is configured as a first destination node for the data from the first source node, at a third network node;

wherein a first memory in the first source node is configured as a first source memory and a second memory in the first destination node is configured as a first destination memory for a data being transferred;

the first network switch using RDMA for collecting the data being transferred from the first source memory and using RDMA to store the data being transferred in the first destination memory at the first destination node.

8. The network system of claim **7**, wherein a plurality of network switches, distributed within the network at the plurality of network nodes are enabled to use per flow congestion control (PFC) to avoid data loss of priority data due to congestion at a loaded queue at the network switch.

9. The network system of claim **7**, wherein the first network switch is configured to implement a method for data transfer between the source memory at the second network node and the destination memory at the third network node, the method comprising:

receiving by the first network switch a first address in the first source memory in the first source node of a data to be transferred to a second address in the destination memory;

the first network switch retrieving the data to be using RDMA from the first address in the source memory; and

the first network switch writing the data to be transferred to the destination memory at the second address, using RDMA.

10. The network system for enabling RDMA in a data center of claim **8**, wherein PFC comprise sending by a switch at a node a request to the prior node to delay or pause transmission of data link frames to a loaded queue indicative of a congestion at the queue in the switch at the node; and

the prior node responding to the request by delaying or pausing for a time the transmission of datalink frames to be stored in the loaded queue.

11. The network system for enabling RDMA in a data center of claim **10**, wherein the network is configured to carry data flows of differing priorities over the same network.

12. The network system for enabling remote memory access in a data center of claim **10**, wherein different queues at a port of each of the plurality of network switches are configured to store and forward data flows of differing priorities; and

wherein data of different priorities are combinable to store in a queue.

13. The network system for enabling remote memory access in a data center of claim **7**, wherein RDMA enable bypassing of the processor and directly accessing the memory for data storage and retrieval.

14. The network system for enabling RDMA in a data center of claim **13**, wherein use of RDMA enabling bypassing the processors, reduce the latency associated with data retrieval, and data storage.

15. A network system comprising a plurality of network nodes comprising network switches for forwarding network packets, the network system implementing a method for congestion reduction, said method comprising:

a. identifying congestion in a first network switch when it occurs;

b. said first network switch creating requests which are sent to at least one second network switch to reduce congestion;

c. said at least one second network switch delaying forwarding network packets which would cause additional congestion at said first network switch while forwarding other network packets which would not cause congestion;

d. continuing said delaying forwarding network packets which would cause additional congestion until congestion at said congestion first network switch subsides.

16. The network system of claim **15**, wherein a first plurality of network switches in said plurality of network nodes implementing said method to delay forwarding network packets which would cause additional congestion in a second plurality of network switches to reduce and remove congestion at said second plurality of network switches.

17. A network system comprising at least one first originating node generating network packets and a plurality of network nodes comprising network switches for forwarding network packets, the network system implementing a method for congestion reduction, said method comprising:

a. identifying congestion in a network switch when it occurs;

b. said network switch creating requests which are sent to said at least one first originating node generating network packets which caused congestion;

c. said at least one first originating node delaying creating or forwarding network packets which would cause additional congestion at said network switch while forwarding other network packets which will not cause congestion;

d. said method being used until congestion at said network switch subsides.

18. The network system of claim **17**, wherein a plurality of originating nodes generating network packets and a first plurality of network switches in said plurality of network nodes cooperating by implementing said method to delay creating or forwarding network packets which would cause additional congestion in a second plurality of network switches to reduce and remove congestion at said second plurality of network switches.

* * * * *