

# AS Design Issues

Bob Davis  
Network Appliance Inc.  
January 5, 2003

## **Introduction**

This paper is to discuss the issues surrounding the development of the PCI Express AS specification, here after referred to simply as AS. I will cover issues pertinent to the design and use of this powerful interconnect system with the goal of broadening the application space for AS.

Major issues will be discussed with current position, minimum acceptable position, desired position, and the rationale for the position or choice. I trust this approach will prove useful in the design decisions before us.

Balancing between: 1) Time-To-Market, and 2) Broadening the application space and developing needs, has been a major stress for all of us. Both solution spaces have their place, but I believe several compromises that have been proposed to push TTM do not need to be made. The life expectancy and ROI of the products designed to this standard will be greatly enhanced with a little more thought to broadening the application space and looking at projected requirements of the systems to be designed with the vast potential of this interconnect fabric.

## **General**

AS is being positioned as a fabric through which multiple PCI Express endpoints, and architecturally different endpoints, may be interconnected as the building blocks of the next generation systems. Each of the endpoints, or nodes, has its own address space dictated by the design of the node being an addressing domain unto itself. AS connects these nodes through an independent address domain, or name space, of its own and separate from those of the attached nodes.

AS Nodes exist in the AS fabric on a Peer-to-Peer basis. There is no imposed PCI style, or based, hierarchy attached to this configuration.

All efforts in designing the AS fabric should be limited to the development of the AS fabric and the transport of packets through that fabric from any node to one, or more, other nodes in a reliable, secure, manner. How endpoints treat the delivered packet, possibly after reassembly, is the endpoint's business, not the fabrics. Part of this development plan must include the Segmentation-And-Reassembly (SAR) process to permit the delivery of large blocks of data over a fabric with a restrictive, and dynamic, Maximum Transfer Unit (MTU), from one endpoint to one or more other endpoints. Completion of the SAR design will eliminate many of the contentious design issues.

We should distinguish between a Routed Payload in the fabric and the Endpoint Payload Packet. The Endpoint Packet may consist of one or more fabric Routed Payloads through SAR operations.

Endpoints and Nodes are equivalent throughout this discussion to be distinguished from a switch which may, or may not, have an attached node, or endpoint, for management or other functions.

## **Longevity**

PCI Express AS should be planned with a minimum of 10 year design life expectancy. If the history of the ISA and PCI architectures are any kind of a predictor, it will probably still be viable in 15 years.

Consider normal trends, with history, as indicators of future developments. Interconnect speeds still double every 18 months. Typical and Max Memory sizes increase requiring an additional bit of address every year to 18 months (12 bits in 1974 + 28 years = 40 bits now). Sun SunFire 15K is offering 512 GB of RAM now (39 bits) and IBM is just behind at 256 GB of RAM. This would lead to 48-50 bits of RAM address space per processing element requiring the capability of 64 bit addressing is needed in each endpoint. Alternate forms of storage will become available that will be significantly denser and less expensive over the life of this specification.

Processing elements will continue to grow, based on the Intel long term technology roadmap. Please review the Intel Labs technology roadmaps presented at the Spring 2002 IDF. Those roadmaps also show Intel

technology advancing to 35nm IC technology by 2009, and the greater probability of geometry induced faults and the need for better error detection in all circuit elements. Planning/allowing for future versions of the specification will help grow to places we have not thought about yet.

## ***Routing Header***

Routing Header must be designed to transport packets and sub-packets (SAR) through a FABRIC, of switches and compound switches, from one endpoint to one or more other endpoints in a reliable manner. The parameters for measuring the goodness of the header are:

1. Header Integrity
2. Ability to configure the **fabric**
3. Ability to stay viable for 10+ years
4. Ability to route packets through the **fabric** from any endpoint to one or more other endpoints
5. Ability to recover from broken **fabric** elements or links
6. Ability to handle **fabric** events
7. Ability to identify all elements in the **fabric**.
8. Ability to scale **FABRIC** in size and bandwidth without limit
9. Ability to grow with Technology
10. Ability to grow through version control for new services
11. Ability to measure **fabric** performance
12. Ability to optimize routing in **fabric** for performance
13. Ability to dynamically reconfigure to optimize data flow in **fabric** from pairs of endpoints
14. Ability to remain cost effective in the switch
15. Ability to work with older, less effective, lower cost, switches
16. Ability to provide secure links in the future

## ***Endpoint Payload Packet***

The Endpoint Payload is the ULP required packet and is the validated packet. Attributes of the Endpoint Payload Packet are:

1. Validated packet content including Payload Header
2. Source and Destination Address used for verification of correct path
3. Security Key validation available
4. Directory based Cache Coherency Protocol Available
5. End-to-End CRC required
6. Long Packets possible/probable
7. Transparent SAR functions
8. Encrypted Packets

## ***Routed/Routing Payload Packet***

Endpoint, or node, payload packets may use one or more Routing Payload Packets to complete the transfer of data from endpoint to endpoint. The routed payload is determined by the MTU of the fabric between the Source node and the Destination node. Endpoint application packets, of any size, are sent through the fabric with Routed Payload packets of the size determined by capabilities of the endpoints and the elements between the endpoints with the SAR functions.

Issues with the Routed Payload Packet include:

1. Performance measurement
2. Time Stamping
3. Routing optimization
4. Secure transmission
5. Congestion management
6. Multiple Pathing
7. Alternate Pathing for reliability
8. Continuous Upgrade

## Segmentation And Reassembly

This is the largest major hole in the current specification that has not been addressed in some manner. This hole must be filled prior to the 1.0 version of the Specification and possible the 0.9 version. Developing this SAR operation will relieve much of the pressure and confusion in the discussions of FABRIC traffic and Endpoint Payload. Endpoint payload packet is described above and may consist of one or more packets through the fabric. The endpoints are unaware of how the payload packet arrived; only that it has arrived and is ready for use by the endpoint. A potential model for the SAR packet may be borrowed from the IB Vol1 page 217.

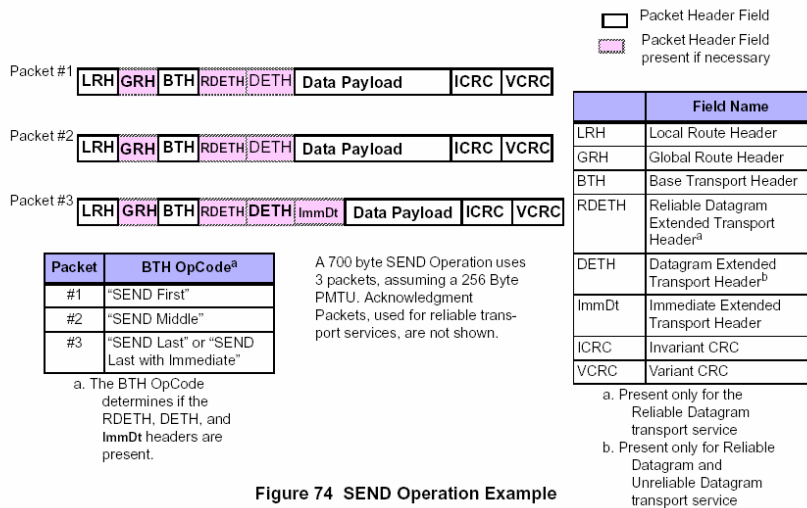


Figure 74 SEND Operation Example

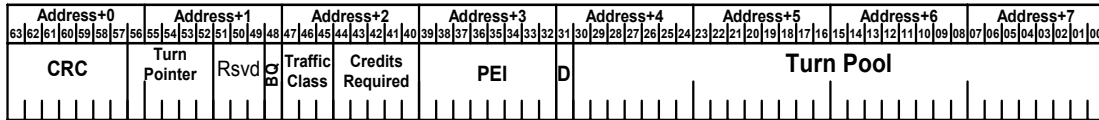
It appears the much of the longer header format is needed even in SAR packets. The addressability requirements remain. 82 bytes of overhead are described in the IB SAR header without any coherency protocol. A proposal for an SAR will be made.

## Issue/Concern – Routing Header Turnpool/Pointer

Routing packets between a source and a destination requires a route. This route is propagated through the various switches to reach an endpoint.

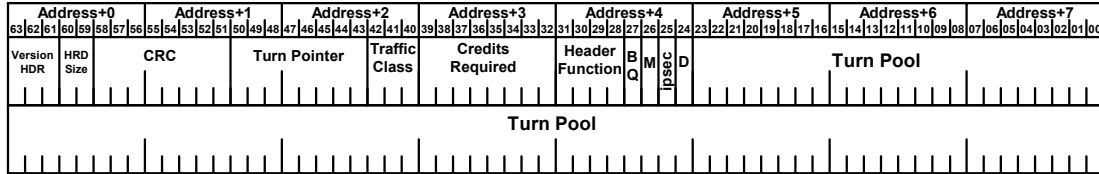
### Current Plan/Position

31 bit turnpool is the current plan of record.



### Minimum Required Position

Add another octlet of 64 bits to the turnpool.

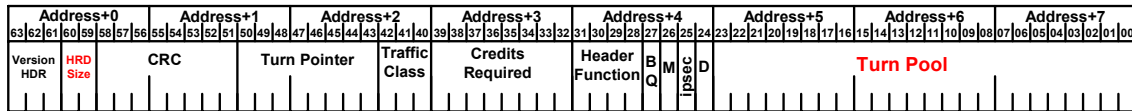


To make room for the additional fields for Version and Header Size, and to expand the CRC and Turn Pointer fields, the Routing Header is proposed to be modified as shown above.

### Desired Position

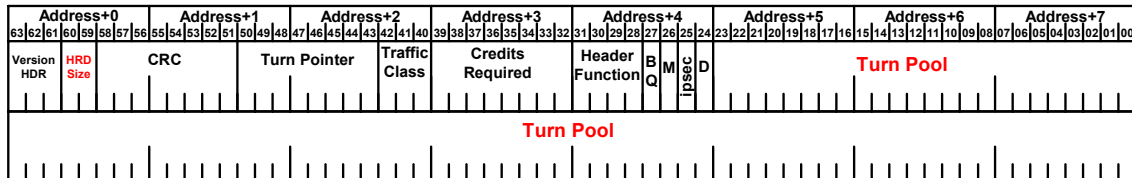
Add 2 extension lengths, one of an additional 64 bits octlet and one that adds 3 octlets of bits to the turnpool for those that need it. This would bring the total to 192 added bits plus the 31 in the initial turnpool for a combined total of 223 bits of turnpool, assuming no invasion of the current 31 bits of turnpool is made by other changes to the header. If the modified Quad Header proposal is adopted, the 31 bit initial turnpool is reduced to 24 bits and the total for one additional octlet is 88 bits and 216 bits with the 3 octlet addition. This would also require the enlargement of the turn pointer and the hop count fields to 8 bits. This additional turnpool bits, along with longer turn pointer provides for larger and more complex fabrics.

Short Turn Pool Routing Header with header size field of 00b:



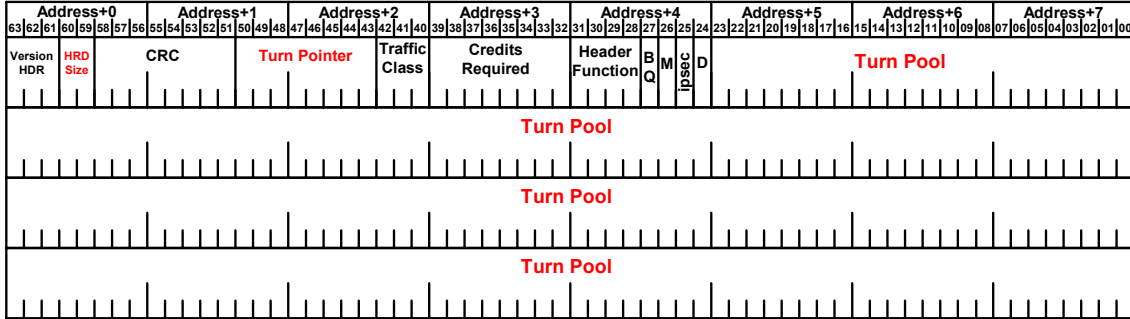
This Short Routing Header contains a Turn Pool of 24 bits.

Medium Turn Pool Routing Header with header size field of 01b:



This Medium Length proposed Turn Pool Routing Header contains 88 bits of Turn Pool.

Maximum Turn Pool Routing Header with header size field of 10b:



The longest turnpool proposed is 216 bits with three additional octlets of turn pool available. The 8 bit Credit Requirements field should use the similar type of multiplication in the high order bits that are used in the current scheme.

CreditRequirement(7,6) act as a multiplier on the low order 6 bits with:

1. CreditRequirement(7,6) = 00 being a multiplier of 1
2. CreditRequirement(7,6) = 01 being a multiplier of 4
3. CreditRequirement(7,6) = 10 being a multiplier of 16
4. CreditRequirement(7,6) = 11 being a multiplier of 64

In addition to this calculation a constant adder of 3 to account for 64 bytes of header, 64 bytes of addresses, command, time, key, CRC, and 64 byte for data alignment and future Routing Header or Packet Header expansion.

To move 4096 bytes of data CreditRequirement(7,0) of  $00111111_b + 3$  or  $64 + 3 = 67$ .

To move 65536 bytes of data CreditRequirement(7,0) would be  $10111111_b + 3 = 1024 + 3 = 1027$ .

A Routing Header with header size field of 11b has not yet been defined.

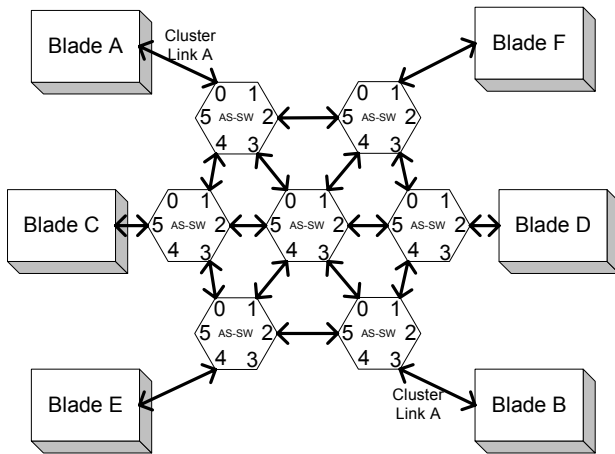
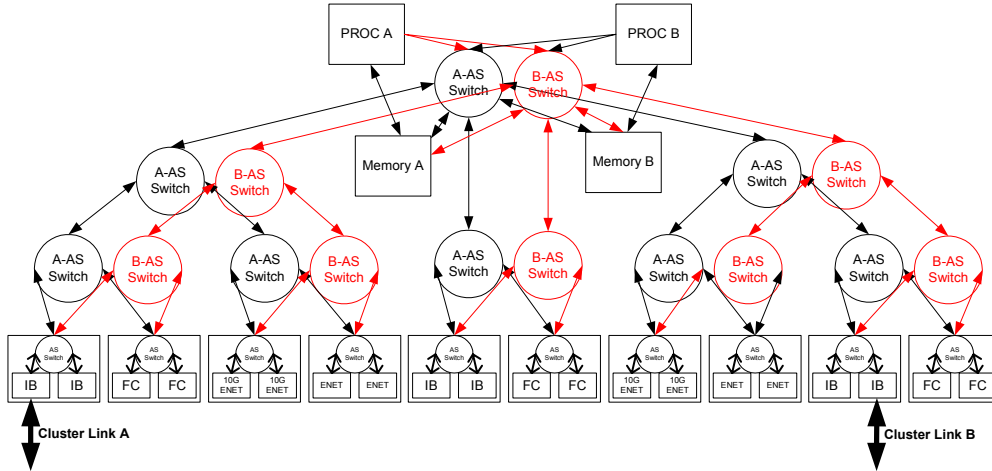
### Rational

With a strict address routed model the use of 31 bits of address would result in the potential of  $2^{31}$  devices in the system which would of course be sufficient. With address routing, the shape of the network is not a factor in being able to reach any of the nodes. Source routing, the current plan of record, uses bits in a different pattern with 8 port switches using 3 bits of routing turn pool per switch limiting the reach to 10 switch hops plus 1 two position switch before exhaustion of the turnpool. A strictly tree structured system could theoretically reach  $2^{31}$  devices from the head of the tree and the last leaf of the tree could reach back to the root node. However the all leaves in the tree may reach only a small section of the overall fabric based on their positioning in the bottom of the tree and can only reach 8192 devices if 6 bit switches are used or 65536 devices if 1 bit switches were used. Non-tree structured system design with a peer-to-peer connection model will yield a much smaller span than  $2^{31}$  nodes in the system.

If a 1 bit switch is used and connected with one port going to the next switch and one to an endpoint the maximum number of devices is 31. If this were used to emulate a common Fibre Channel configuration used by the storage industry, we could not achieve our current configurations of up to 56 disks attached to a standard 1 Gb/s FC loop. FC drives are actually dual attached with 2 independent FC loops going to each drive. New designs would use a different approach using higher port count switch and not have this limitation.

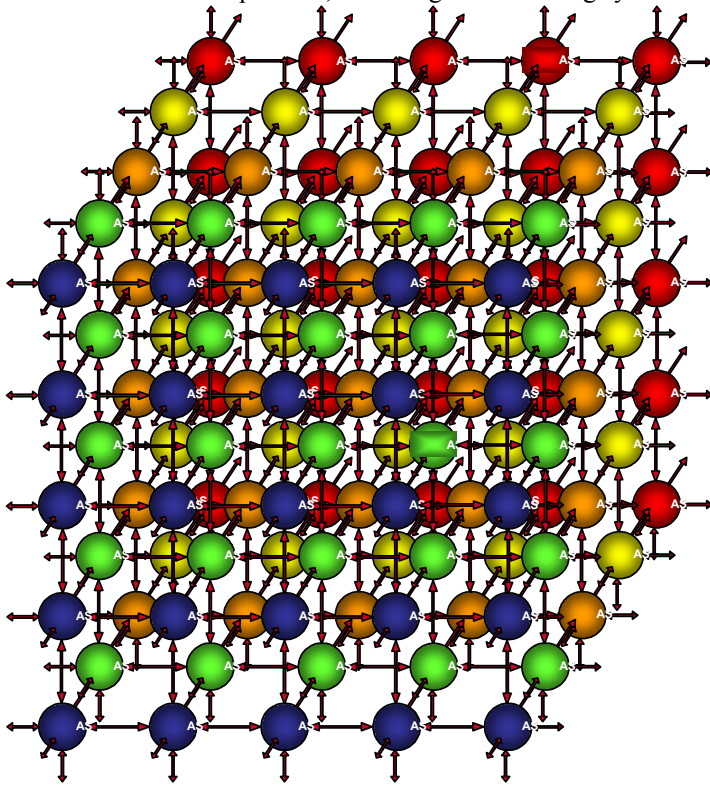
In a High Availability environment, the ability to take over the I/O of a failed cluster partner requires more hops and turnpool than currently allowed for. For instance if several of the filer head configurations shown in the first drawing below were connected with the matrix in the second drawing, for Blade A to take over the all of the I/O of Blade B be would require up to 18 hops and more than 31 bits of turnpool, assuming that all the switches use 3 bits of turnpool and all connections are not shown.

## Current PCI Express AS Specification Issues



Higher order matrix/mesh routing using 7 port switches limits the reach to 10 hops which fails to reach elements from corner to corner in a 5 x 5 x 5 matrix of nodes/switches. The applications shown below

would be a video composition, switching and rendering system.



Blue = Film Scanner or  
Video Feeds  
Green = Video Proc  
Orange = Frame buffer  
Yellow = Frame buffer  
Red = Storage system  
Nodes

All inputs go to storage  
and to Video Processor

36 bits of route  
needed from  
lower left to  
upper right

## Issue/Concern – AS Address/Name Space

AS Fabric is its own Address Domain or naming space. Every element of an attached PCI or Non-PCI Domain must be uniquely identified by the it's node address in the AS Domain and it's separate additional address of the device in the PCI, or Non-PCI, attached node address Domain.

How are Nodes/Endpoint and Switches identified in the AS Fabric. Each device, Node, Switch, Endpoint must be identified to be addressable from any other node.

Additional concern is the identification of ports on a switch or endpoint.

We must do one of:

1. Add a field for the port identifier such that the port is uniquely defined as {Device:Port}. Current plans include switches with up to 64 ports. This would make the default choice for the management port FF<sub>h</sub>.
2. Assign a EUI-64 name/address to each port separately. This would consume EUI64's for each port and for the device itself and possibly for its management function.
3. Assign and EUI-64 to every device in the fabric and assume a sub address space of 64 bits for the contents of an endpoint or switch, possibly through a translation table, and to identify the ports on a switch when needed in switches.

### Current Plan/Position

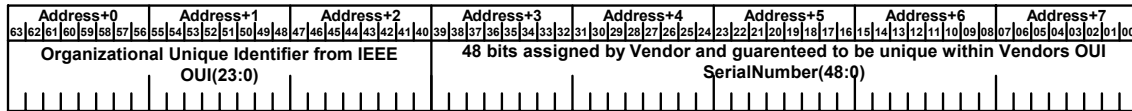
The Route is the identifier for any device in the fabric.

### Minimum Required Position

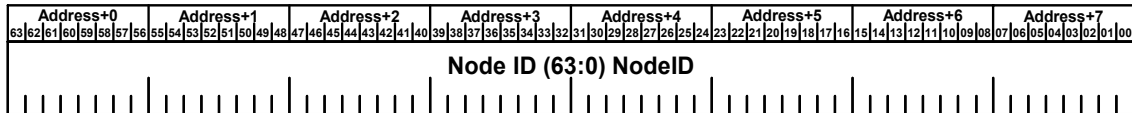
Each node/switch/endpoint has a unique address that is describable to any other node/switch/endpoint in the fabric.

### Desired Position

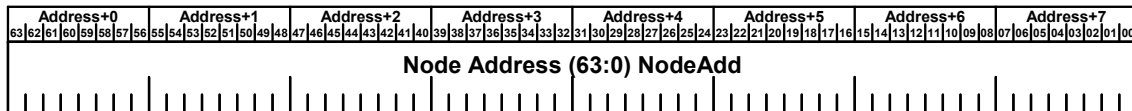
Each endpoint/node/switch in the AS Fabric has a EUI-64 identifier consisting of an Organizationally Unique Identifier and a vendor assigned serial number or other identifier that is unique within the vendor's OUI. This is the extended version of the Ethernet addressing system and uses the same OUI.



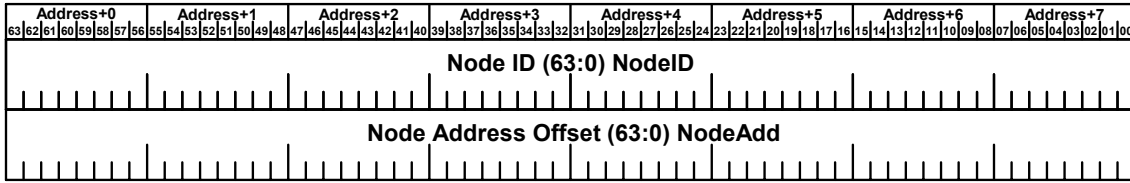
This becomes the NodeID(63:0) as:



Addressing within the nodes addressing domain may be of any length. Most of the expected nodes for the AS fabric will be PCI, or PCI like, with a 64 bit address space. Assuming the use of the memory space, this will be shown as Node Address Space, Node Address Offset (from 0) or endpoint address domain:



PCI contains several address spaces that would require at least 66 bits to fully describe with the addition of Memory Space, I/O Space, Message/Event Space and Configuration Space. An approach using the general protocol encapsulation layer which handle this problem by adding the additional bits in the command field. Endpoints can reach any other endpoints address by creating a 128 concatenation of the ASnodeEUI64 and the 64 bit address in the endpoints address space. Bridge hardware will look up the path and forward the access.



This addressing scheme maintains symmetry between the switches and endpoints.

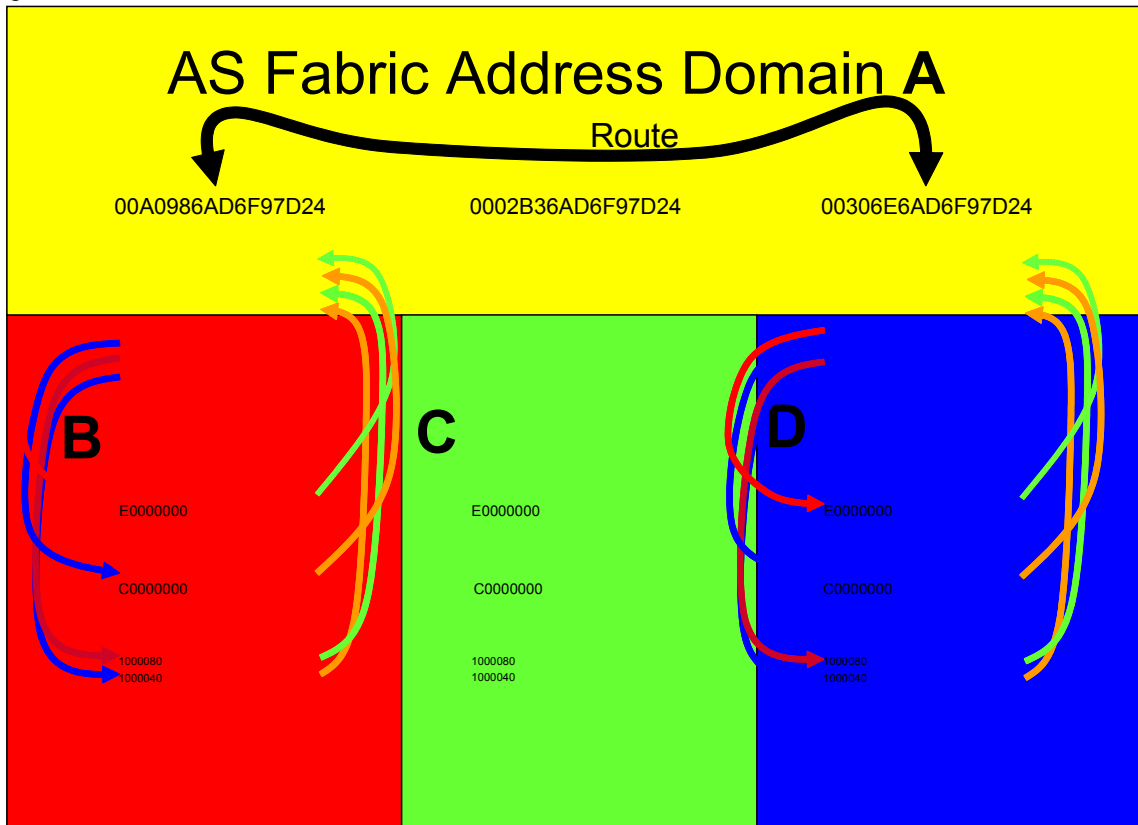
**Rational**

The design of the AS Fabric is to tie together many endpoints with a switching fabric. Each of those endpoints has an addressing domain that is local to that nodes domain. For example, each PCI Express or PCI-X endpoint has a complete 64 bit address space that may be used in any manner the designer of that endpoint desires. If we are connecting 100 different PCI-X endpoints, we must assume that each has utilized the space with that endpoints domain in an identical manner: i.e. each has an Ethernet controller at 64 bit address F000000000000006<sub>h</sub> and memory from 0001000000000000<sub>h</sub> to 00010000FFFFFFFF<sub>h</sub>. It is clearly necessary to differentiate between these various common address domains. Adding the AS Address Domain solves this problem.

Using the EUI-64 unique identifier for the bridge node provides a name space for the AS Fabric. Each element has an address. The route is the path from one address to another address. Passing an AS Address from one node of the fabric to another node of the fabric allow the second node to look up its path to that address. The message is sent by the path from a source address to a destination address.

The switches only involvement in this process is to respond as needed for configuration and for identifying itself in and fabric event.

All disk systems have a worldwide name of 64 bits assigned to each node. Please see the SAS, FC, ATAPI7 and SATA specifications for details. These are all derived from the EUI-64 specification



This shows the function of the bridge in mapping multiple PCI spaces with identical address configurations being mapped to each other in the bridge. Each of the PCI Express or PCIX nodes write to a local address

which generates the packet with proper addresses and Routing Header to get the write to the planned location.

## Issue/Concern – Source and Destination Address

Each packet must include the source and destination address to validate the packet.

### Current Plan/Position

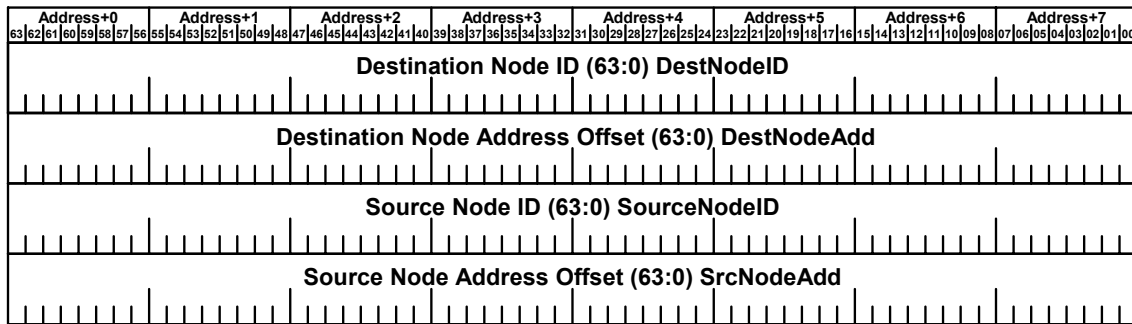
The Route is sufficient.

### Minimum Required Position

Destination and Source Addresses independent of Route

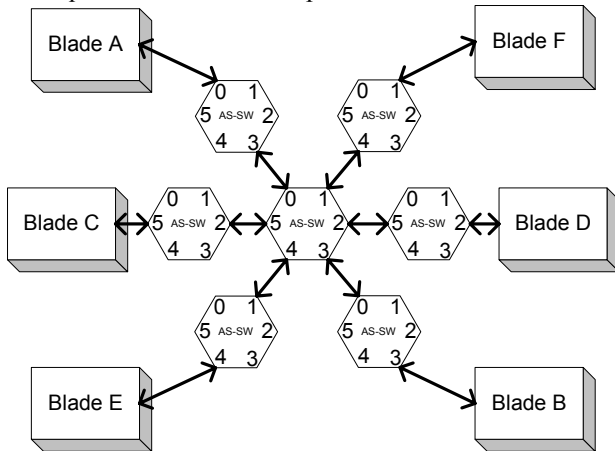
### Desired Position

Source and Destination address included in packet to validate transfer. This is to be the SourceNodeEUI64:SourceNodeAddress and DestinationNodeEUI64:DestinationNodeAddress.



### Rational

The requirement is the complete address of the source of the data and the destination of the data is delivered with the data packet to the destination. This still assumes that the Route Header is using source based routing per the current plan with the possible addition of additional turn pool bits. Translation tables will probably be address to convert the addresses present into the real addresses in each endpoint. Currently, AS endpoint identification is only by the Routing information from the source. This is inadequate for validation of packets.



This example demonstrates the problem;  
 Route for Blade A to Blade B = 2,2,2  
 Route for Blade C to Blade D = 2,2,2

## Current PCI Express AS Specification Issues

Route for Blade E to Blade F = 2,2,2

Route for Blade B to Blade A = 2,2,2

Route for Blade D to Blade C = 2,2,2

Route for Blade F to Blade E = 2,2,2

Any error in the central switch cannot be detected in the received packet based on the route. A packet from Blade A to Blade B that is misrouted to Blade D will still have a valid Route and Blade D would perform the operation destined for Blade B. This is an Undetected Error.

In this example if Blade B sends an event to Blade A and Blade A directs Blade E to handle the problem, how does Blade A tell Blade E how to access Blade B to service the Event?

Many ULP's will be used with AS, many of them need the extended addressing capability.

While older interconnects used 16 bits for the nodeID and 48 bits as addressing in that node, PCI Express and PCI-X and PCI-X2.0 all have a 64 bit address within the node preventing the copying of earlier formats.

It should also be noted that one of the severe limits of SCI, and its derivatives, is the shortened address space in the node, which will most likely be fixed in a future revision of that standard. At the time that SCI was created, 1986, this address space was sufficient to the purpose. Time has passed and the address space within the nodes is inadequate.

## Issue/Concern – Improved Packet Header Format

Improvements needed to accommodate current and projected system.

### Current Plan/Position

None in AS Ver 0.8. A SLS proposal has be forwarded

### Minimum Required Position

Modification to include source and destination address – 128 bit version.

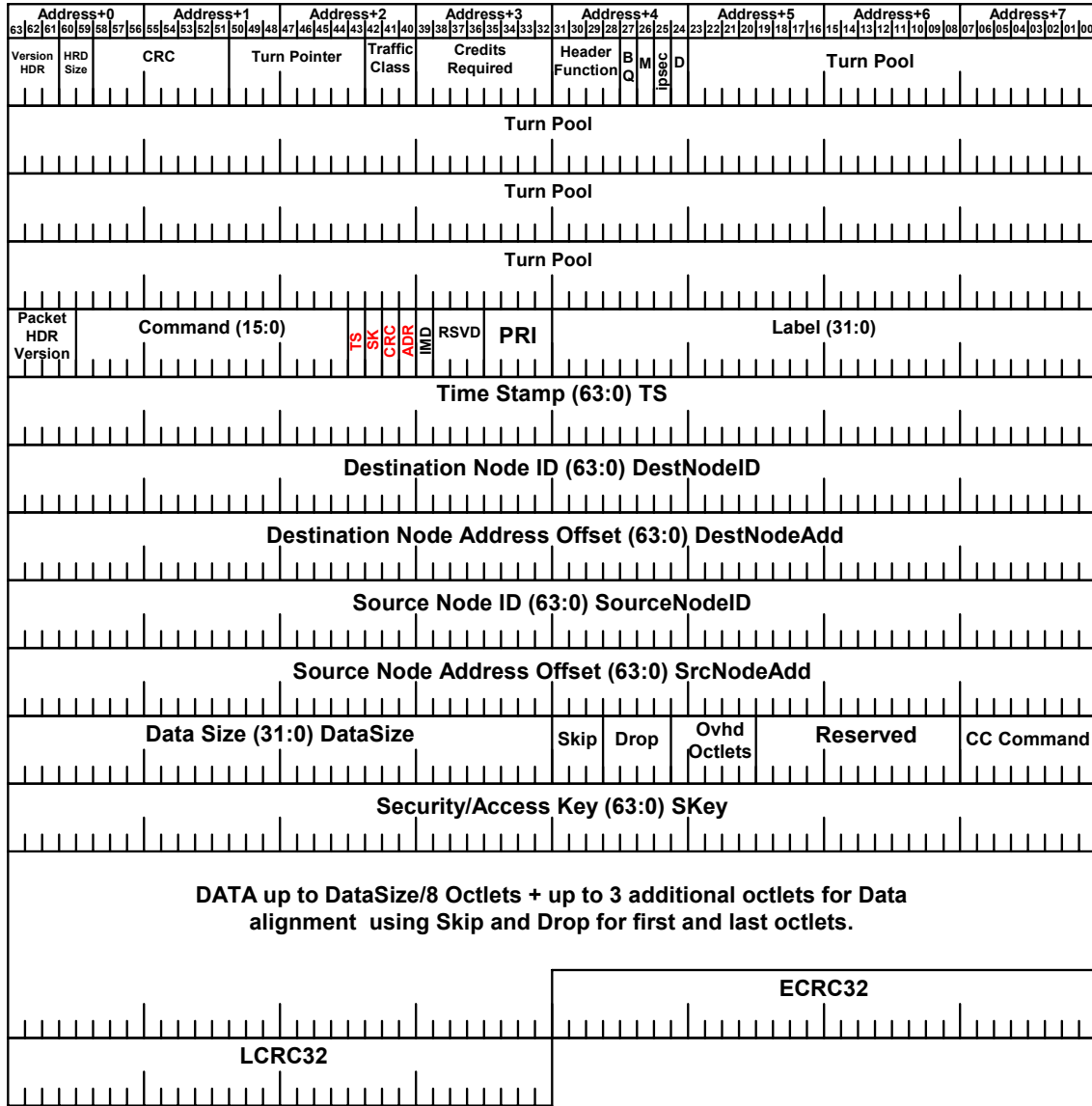
### Desired Position

Adopt the following proposal. This proposal has optional elements that are included as needed.

The shortest version that would transfer immediate data is:

Address+0		Address+1				Address+2				Address+3				Address+4				Address+5				Address+6				Address+7																																					
63	62	61	60	59	58	57	56	55	54	53	52	51	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00
Version HDR		HRD Size		CRC				Turn Pointer				Traffic Class		Credits Required				Header Function		B	M	IPSEC		D	Turn Pool																																						
Packet HDR		Command (15:0)															TS		SIX		CRC		ADR		IMD		RESERVED				Data(31:0)																																
Version																																																															

With the IMB bit set data is included in the Command octlet, data is included in the lower quadlet. Using the longest proposed Routing Header and the option control bits in the Command octlet, additional information can be included to provide increased security and flexibility.



Each of the potential options will be discussed in response to the remaining issues.

### ***Rational***

Greatly improve flexibility in this proposed format with the retained ability to send very short packets with immediate data.

## Issue/Concern –SAR (Segmentation And Reassembly)

Must be defined.

### Current Plan/Position

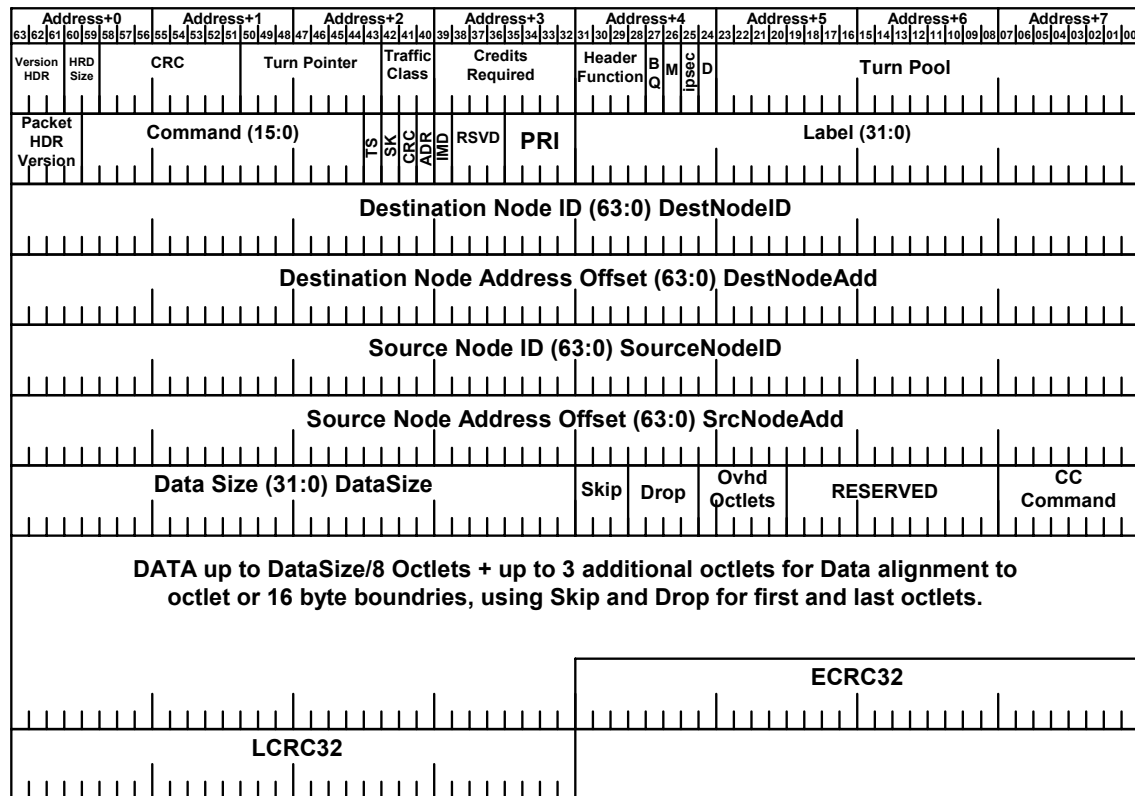
None stated, may be assumed but not specified.

### Minimum Required Position

SAR policy to meet the needs of the data flow.

### Desired Position

Adopt the following proposal to put SAR frames in the command elements.



This general mechanism can use the DestNodeAdd and the SrcNodeAdd in two ways:

1. As pointers defining the queue pair participating in this operation.
2. As pointers, through a Protection Translation Table, if one exists, or directly to the appropriate memory locations in the respective endpoints. This would require that the addresses change to properly deposit the data in the proper place.

The Label(31:1) is incremented with each succeeding data packet.

With the addition of the time stamp, this set of SAR transactions can be encrypted for transfer.

The Command field guides the setup and progress for the SAR operation

1. Command Field (15:0) = 01xx<sub>h</sub> for first transfer in a SAR operation
2. Command Field (15:0) = 02xx<sub>h</sub> for middle transfers in a SAR operation
3. Command Field (15:0) = 03xx<sub>h</sub> for last transfer in a SAR operation
4. Command Field (15:0) = 04xx<sub>h</sub> for setup of a SAR operation where the addresses and length are for the overall data block to be SAR'ed

***Rational***

Required to get large blocks of data through the system in smaller packets. This proposal has 64 bytes of overhead compared to the 82 of IB. Adding time and key octets would bring the total to 80 bytes.

## **Issue/Concern – Arbitration Resolution**

Finding the Fabric Master uniquely and determining paths from any endpoint to all other endpoints.

### ***Current Plan/Position***

Use the broadcast mechanism to establish Fabric Master and Alternate Fabric Master.

Use a timed approach to bias the selection in favor a particular master.

Limit the size of the fabric to 31 hops from a master.

### ***Minimum Required Position***

Use the probing method to determine the fabric and the masters. Enlarge the turnpool to allow the discovery of larger fabrics.

### ***Desired Position***

Use probing method only to determine the primary fabric master (configuration master) and secondary master. This will also provide all possible paths from any endpoint to all other endpoints by nodeID.

There is still a limit but it is expanded by the use of the long turnpool header. In complex matrix systems, this limit may be reached but alternate paths to the same nodeID should be able to resolve the problem.

Using the probe method is fully deterministic. It can start at any time, possibly periodically, and transfer of mastership negotiated.

### ***Rational***

The broadcast mechanism works if sufficient time is allowed to reach all participating nodes. The problem is in determining the all inclusive timeout, it is not deterministic.

In a large fabric the endpoints for one master candidate may be within the 31 limit but they may not be for another master candidate. This severely constrains the layout of the fabric. This may result in 2 master candidates each presuming that it has won the election within the area prior to pruning per the current design. This is a bad and unstable condition.

Fabric will be added onto in a haphazard manner without regard to the maximum distance to a particular fabric master.

Using timeouts to favor a given self selected master candidate shall presume to be duplicated by all other master candidates in the system. Hence all will assume they have won the election and start to dictate different configuration request in parallel. Most likely Fatal.

All endpoints or nodes in the fabric must probe all other points in the fabric to establish one or more paths from itself to every other endpoint in the fabric. This will include switches as the servicing of events from the switches will require the servicing node to know the path from itself to the switch by name.

This is a peer-to-peer fabric with all potential masters being equal. Once all of the endpoints are located, each of the endpoints will service and configure the endpoints that it expects to be in its sub-fabric as part of the larger fabric. There may be as many sub-fabrics as there are pairs of nodes.

Each sub-fabric may continuously optimize itself based on congestion mechanism results. This is also where the time part of the packets becomes important again.

Overall fabric time is the responsibility of the fabric master, but may be delegated to another node, usually one with outside time sync capability.

The elected fabric master shall be responsible for maintaining the time, possibly through delegation, and be a target for events, including interrupts from all the endpoints and switches, and may assign all other duties, such as node configuration and interrupt response, to other nodes in the system, by setting up the multicast event tables.

## Issue/Concern – Version Numbers

### **Current Plan/Position**

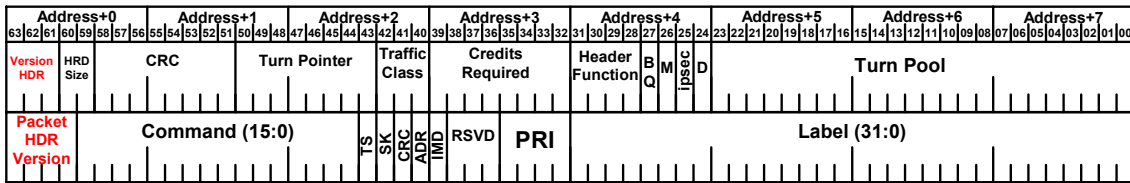
None – There is no way to upgrade later without the ability to determine the version number for proper interpretation.

### **Minimum Required Position**

2 bit field to allow for 4 versions of the specification in both the route header and the payload packet

### **Desired Position**

3 bit field to allow for up to 9 versions of the specification for the route header and a 4 bit version field in the payload packet.



### **Rational**

A version number is needed in the routing header and the routed packet to support later changes to the specification and make them backward compatible.

## Issue/Concern – Optional Time Stamp and Time Synchronization

Proposal to add an optional time stamp into the packet format. This time stamp, TS, should be in a standard format, internationally recognized such as IEEE Std 1588-2002 or IETF rfc2030.

The purpose is to have a method of time in the fabric for use in measuring performance and to date messages for discarding.

### Current Plan/Position

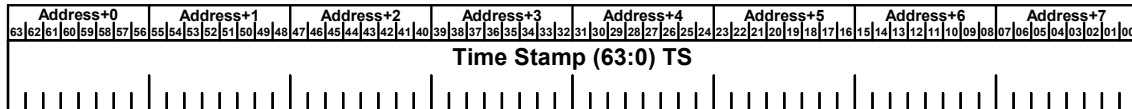
None

### Minimum Required Position

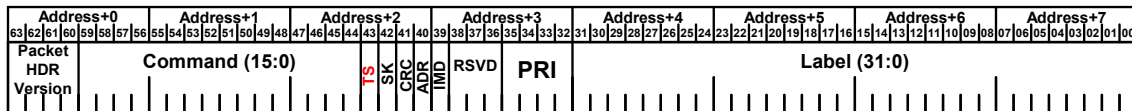
Add a time object into the packet as an optional element.

### Desired Position

Add time to the packed format in either the IEEE Std 1588-2002 Precision Time Protocol, PTP, format.



IEEE Std 1588-2002 defines the high order quadlet, TS(63:32), as the time in seconds from the PTP epoch of 0 hours 1 January 1970, and the low order quadlet, TS(31:0) as the time in nanoseconds since the last second. Granularity is 1 nanosecond.



Bit 43 in the command octlet is set to 1 if the time stamp is present and 0 if it is not present. The time reference point in the transfer shall be the start of the first symbol for the Packet Header. The time stamp octlet is positioned directly after the command octlet to be use also as an initial variable for an ipsec encrypted packet. There is a current specification for PTP for USB and PCI.

### Rational

Please see IEEE Std 1588 – 2002 “IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems.” I believe the 1588 format is preferred as this standard also includes the algorithms to maintain a sense of time in a networked environment (fabrics) such as PCI Express AS. The other option for the time stamp is the IETF rfc2030. For rfc2030 the high order quadlet, TS(63:32), is in seconds from the NTP epoch of 0 hours, 1 January 1900 and the low order quadlet, TS(31:0) is in 1/2<sup>32</sup> partial seconds. Granularity is 232.83 picoseconds. The conversion between NTP and GPS to PTP is covered in the 1588 document. Time is used for timeout purposes, for heartbeats, for stale packet disposal and for keeping of performance statistics on the fabric. Fabric delays can be measured accurately and the fabric tuned for better performance with this help. A sense of time is required throughout the fabric to be able to measure performance, timeout events, understand heartbeats and halt runaway processes. Timeouts based on time of day are the other method of stopping swarms of packets circulating in the fabric. A short time to live can remove the packet from the system.

## **Issue/Concern – Scalability**

Reconsider built in limits that greatly reduce the scalability of the fabric.

### ***Current Plan/Position***

Limits such as the turnpool, the turnpool counter, the hop counter and others limit the scalability of the Fabric.

### ***Minimum Required Position***

Adopt the multiple routing header protocol proposals using optionally longer Turn Pools.

### ***Desired Position***

Adopt the multiple size routing header proposal.

Adopt a version field to support future redefinition of the specification.

Remove other arbitrary constraints.

Look forward, with admittedly imperfect eyes, to changes in requirements and technologies over the next 10 years. Use some of the Intel Labs IDF presentations as a good starting point.

Look at the Carnegie Mellon University research work in advanced projects <http://www-2.cs.cmu.edu/afs/cs.cmu.edu/Web/Groups/systems/projects.html> .

IBM development work [http://www.almaden.ibm.com/StorageSystems/autonomic\\_storage/index.shtml](http://www.almaden.ibm.com/StorageSystems/autonomic_storage/index.shtml) .

### ***Rational***

Requirements will change for the AS fabric. We need to be able to meet the new needs without completely inventing a new Fabric.

It is much more cost effective to provide for growth now than try to fix a broken protocol in the near future. Changes will occur that none of us can predict at this point in time. Design in a little leeway to be able to adopt the new requirements when they occur.

Changing technology will support much larger protocols than we currently support. The speed of interconnects will also increase on a regular basis.

## **Issue/Concern – Routing Header Hop Count**

This is the limit on the maximum distance, through switches, that can be accommodated in a system.

### ***Current Plan/Position***

Maximum Hop Count is 15

### ***Minimum Required Position***

Maximum of 64

### ***Desired Position***

Maximum of 256

### ***Rational***

Larger hop count field us needed to be able to support multicast over longer distances in the system.

The 4 bit field for the hop count was arbitrary choice based on consideration of a few smaller systems and an attempt to fit into a smaller header. If the additional 4 bits needs to come out of the initial turnpool space, this is acceptable as long as the availability of the longer routing header is present.

## **Issue/Concern – A Secure Transmission Capability**

Adding ipsec style security protocols in the AS specification for future use. Wire Security is one of the strongest requests from customers for use in protecting their data while passing through the system. It is becoming a demand for Storage, Financial, Industrial, Data Processing, and Data Base environments. As AS will be a prime mover in the coming years, I hope, adding the capability of securing the packets while they pass through the fabric is highly attractive and desirable.

Source Routing has several major advantages in providing a more secure environment.

1. The Packet that can be intercepted is directed by a turn pool and not an address. There may be many turn pool sets that will get traffic from Node A to Node B. By randomly choosing the path, no intercept will know he has all the information.
2. Source Routing means that all that needs to be exposed is the Turn Pool and the turn pointer.
3. The overhead cost is 1 bit in the Routing Header.
4. The turnpool can be used by the receiving node to point to the decryption key. Many different turn pool indexed can point to the same key.
5. All information, source, destination, length, data, and ECRC are in the encrypted block.
6. Small changes to the Packet Payload optimize it for encryption with block level ciphers.
7. Time Stamp can act as a Initiation Variable to prevent the cipher text changing even if the same set of data is resent many times.
8. The LCRC is not within the encrypted payload and can still be checked as normal

Virtually no change or addition burden is placed on the switches.

### ***Current Plan/Position***

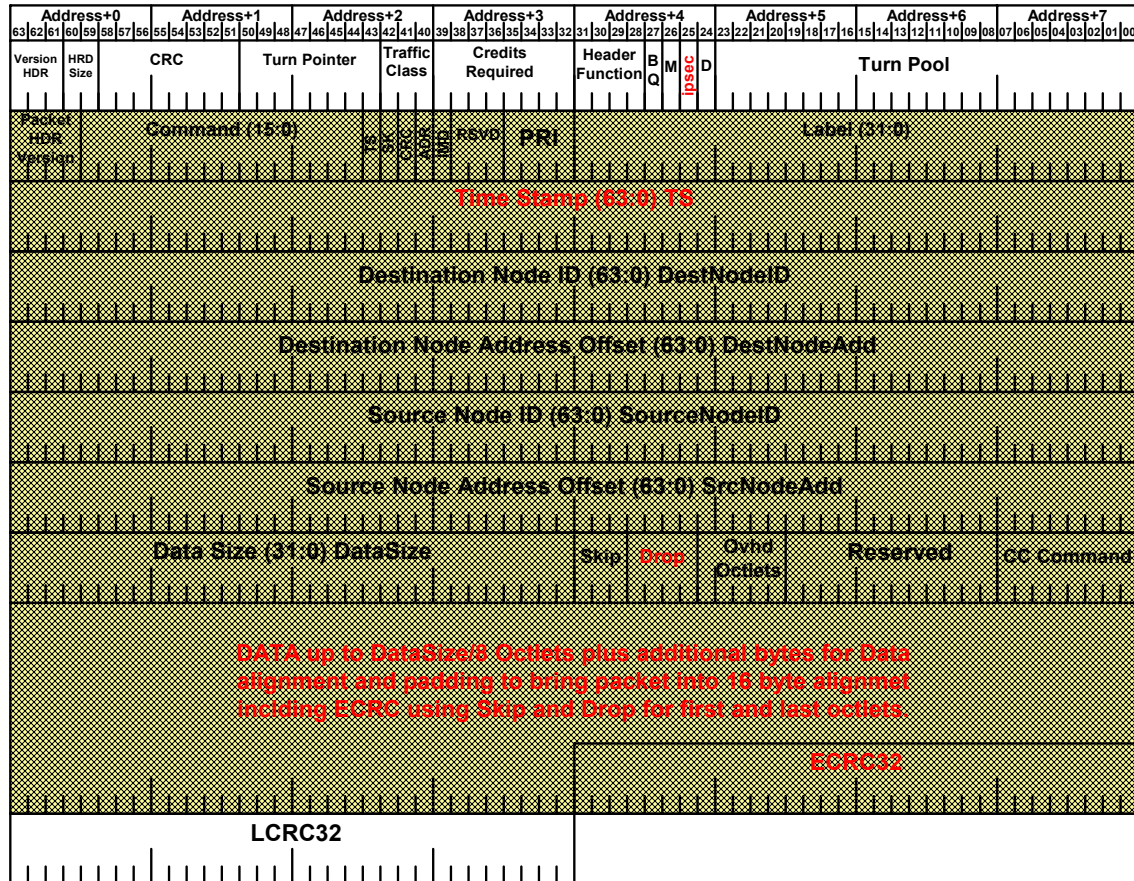
None current, considered, or defined.

### ***Minimum Required Position***

A place holder is needed to provide for future ipsec style security on the transmission of data through the fabric.

## Desired Position

In appreciation for the need for a secure link please consider this proposal:



This example is a Short Routing Header and a full packet header. The ipsec bit is set to one in the Routing header. Using AES as the standard method of encryption, a 128 bit blocking factor is desired. The Drop field allows the dropping up to 15 bytes to support the additional bytes needed to fill needed for 16 byte alignment on the encryption packet.

The encrypted part is from the first bit of the Packet HDR Version through the ECRC as shown in the shaded area of the attached drawing.

Time Stamp is located in the second 64 bits of the encrypted packet to guarantee a change in the first 128 bit encryption block.

The Security/Access Key may or may not be included in the packet. This is again controlled by the command octlet.

This encrypted packet can also be multicast using the multicast table as the index in the receiving units to decode the packet.

The current draft-ietf-ipsec-ciph-aes-cbc-04.txt looks like a valid starting point for an ipsec definition using the AES specification. This proposal makes use of the TimeStamp as the Initialization Vector. Another IV may be chosen as an independent feature and transparent to the fabric.

## Rational

Well with in the life of this standard additional security will be needed on the wire protocol. As it is now defined, AS will have a reach of about 15 meters on a InfiniBand style wire connection, which will allow exposure to possible hostile environment.

This is a low cost approach to adding a shortly required feature.

## Issue/Concern – Coherence Protocol

Coherency Protocol is needed in modern fabrics with large memory spaces. The protocol must be flexible and designed for inter-domain service. Separate coherency mechanism will be defined in each node based on the design of that node. This protocol is concerned with the coherency across the fabric.

### Current Plan/Position

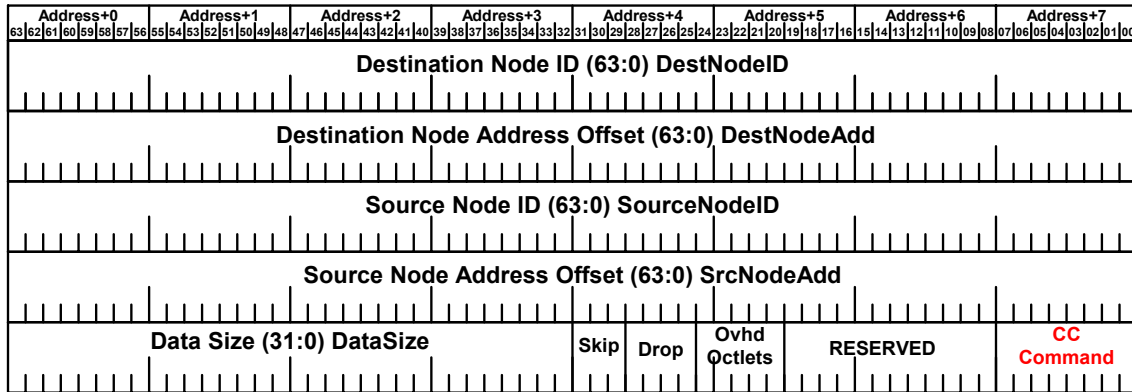
None

### Minimum Required Position

Incorporation of the CC Command field as demonstrated

### Desired Position

Add the field to the header to allow coherency commands to propagate from node to node.



The CC Command field CCC(8:0) is required for future command for the coherency protocol.

### Rational

A directory based Coherency Protocol was developed to be used in large fabric/network configurations where a snoopy protocol is not viable.

This mechanism has been shown to be correct and valid. Many different types of coherency may be implemented over this common mechanism.

The protocol follows the work of many is the SCI definitions and has been updated for use in this specification.

## **Issue/Concern – Common Protocol Send**

Suggest one, and only one, protocol for sending traffic through the fabric with common security treatment.

### ***Current Plan/Position***

Several different routing headers in use.

### ***Minimum Required Position***

Single header set with only options being the path routing pool length.

### ***Desired Position***

Single header set with only options being the path routing pool length.

The PEI-4 would revert to being a recognized address in the switch or any node that is being configured.

The PEI-4 function would replace the Configuration Space requests with a Request to a configuration address block of  $256 * 2^{20}$  bytes, a portion of node space with a base address of FFFF FFFF F000 0000<sub>16</sub>. Registers defined by in the AS specification are located within node register space as are node-dependent registers defined by the applicable node architecture.

### ***Rational***

Simplicity and Security.

This would put the complexities in the packet where it can be protected and secured.

Currently the configuration protocol has no security features associated with it. Any node can reconfigure any other node, or switch, simply by knowing the path to that node or switch.

This is a LARGE security hole. With needed protection on all other transactions through the fabric, this security hole negates the other protections. A minimum protection would be the recognition of the authorized configurer for the node and a proper key to permit the configuration.

As an example consider a disk system as a node. For normal operation we validate the origin of a data transfer and validate that this node is the designated target prior to allowing the read or write on the disk. This hole would allow a device to reconfigure the disk to negate all the data protection attributes.

## **Issue/Concern – Native Mode Operation**

This AS specification should be capable of being used in a Native Mode operation with endpoint of similar or different architectures.

### ***Current Plan/Position***

AS was conceived primarily as a PCI Express Base node connector.

### ***Minimum Required Position***

The AS Specification to be complete and not require PCI Express endpoints for operation.

### ***Desired Position***

The AS specification should be for a generalized fabric design to interconnect large groups of node of different characteristics, for PCI-X, PCI Express, HT, RapidIO etc. I believe that clean operation with multiple PCI Express and PCI-X interfaces will force all the issues to generalize the fabric.

The Packet and Routing Headers presented here will accomplish this task.

### ***Rational***

The future will include many different endpoint designs. The longevity of the switches and endpoints will be supported best without the reliance on a specific substructure and specific endpoints.

Considering the AS fabric as a native interconnect fabric/bus to solve the larger fabric problems is the best use of time and effort. Best ROI.

## Issue/Concern – Packet Command Field

The descriptor used for the decoding of the packet information.

### Current Plan/Position

PEI's are currently defined in the Routing header.

### Minimum Required Position

A command field that is defined within the ECRC coverage.

### Desired Position

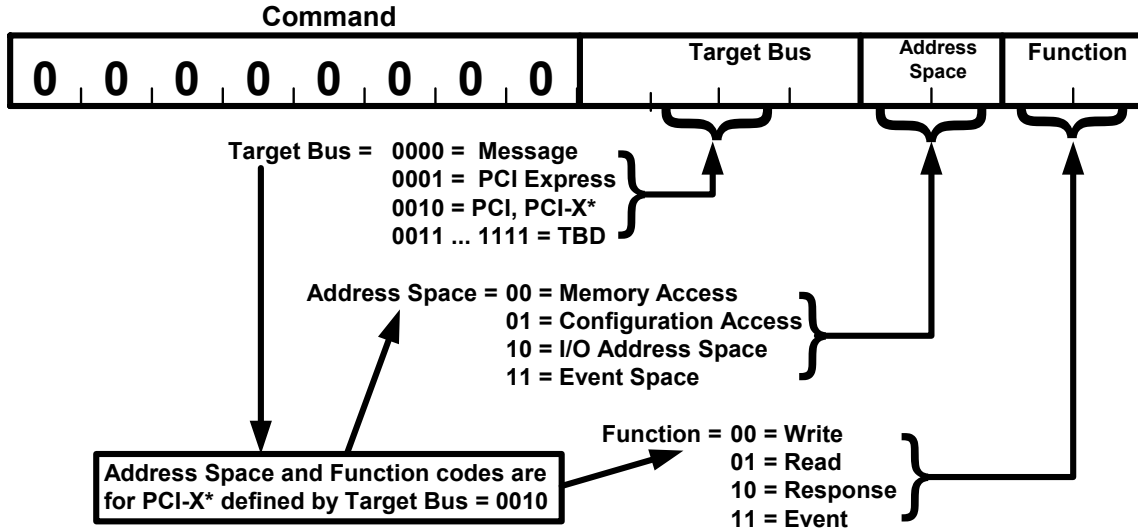
The proposed Command field is:

Address+0		Address+1				Address+2				Address+3				Address+4				Address+5				Address+6				Address+7																																				
63 62 61 60 59 58 57 56 55 54 53 52 51 50 49 48 47 46 45 44 43 42 41 40		39 38 37 36 35 34 33 32				31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16				15 14 13 12 11 10 09 08				07 06 05 04 03 02 01 00																																																
Packet HDR Version	Command (15:0)															TS	SK	CRC	ADR	IMD	RSVD	PRI	Label (31:0)																																							

1. Bits 63:60 -> Version number for the packet header – This version = 0001<sub>b</sub>.
2. Bits 59:44 -> Command(15:0).
3. Bit 43 -> TS, if set to 1 the time stamp immediately follow the command octlet. 0 = no time stamp.
4. Bit 42 -> SK, if set to 1 the security/access key is present otherwise it is not present.
5. Bit 41 -> CRC, if set to 1 the End to End CRC (ECRC) is added following the data and before the link CRC (LCRC).
6. Bit 40 -> ADR, if set to 1 the address block is included following the TimeStamp TS, if present; otherwise it follows the Command octlet. If set to 0, the address field is absent.
7. Bit 39 -> IMD, if set to 1 the Label(31:0) field in bit positions 31:0 contain immediate data in place of a Label. If this bit is 0 the bits 31:0 are a label field Label(31:0).
8. Bit 38:36 -> Reserved for future use.
9. Bits 35:32 -> PRIORITY, specifies the priority of the packet at the target endpoint. The priority field in the Routing Header, Traffic Class, specifies the priority through the fabric.
10. Bits 31:0 -> Label(31:0) to handle the labeling of packets on this source – destination, or multicast group. This labeling keeps with current IETF design philosophy. If the IMD bit is set this is a data doublet.

Two Commands are here proposed:

Command 0 = is defined to work with target buses of the PCI class:



Command 1 -> 4 = SAR Command.

**Rational**

Once the packet is delivered and CRC checked, the command field can be interpreted and the packet dissected into its individual components. This can only be done when the packet is known to be complete and correct.

## **Issue/Concern – Routing Header CRC**

Reliability of the Routing Header is the primary reason for requesting better coverage of the CRC with the additional field and extended turnpool. The possibility of UnDetected Error (UDE) in the routing header is the primary issue.

### ***Current Plan/Position***

7 bits of CRC that includes the entire header except the turnpool pointer. There will not be any errors in the header that are undetected.

### ***Minimum Required Position***

Minimum of 8 bits. To cover the header as it is extended to 256 bits.

### ***Desired Position***

Sufficiently robust for single bit ECC and possibly FEC for the 256 bit header.

### ***Rational***

Integrity of the Header is more important than the loss of bits from the initial turnpool. The issue is probability of Undetected Error in the routing.

The header does not have the 32 bit CRC of the data packet. The header is covered between nodes with the LCRC of 32 bits. The problem occurs in the switches. Each switch routes the header data through multiple multiplexers/switches and probably a minimum of 2 FIFO's and possibly many more in a compound store and forward switch. The header data will be serialized and de-serialized at least one at each switch. The header may go through 31 to 200 switches between Endpoint A and Endpoint B.

Technology is decreasing the size of all the device geometries to 32nm by 2009 according to Intel IDF presentation with test cases down to 15nm. This diminished geometry increases the probability of error in the handling of the header. Additional design precautions will be taken inside the switch design, and we should also take additional precautions outside of the switches to give the best chance of success.

Address phase on PCI-X2.0 has ECC on the address and Data. The probability of an UNDETECTED Error rate for PCI-X2.0 in a large system is  $\sim 3.37E-14$  per year. We should attempt to match the level of UDE in the AS Header. The Payload is better protected with the end to end CRC32 in addition to the LCRC.

If the PEI or Command function is moved, as proposed, to the payload part of the routed packet, there is less concern as the use of the data is covered by the much stronger CRC.

## Issue/Concern – Routing Header Multicast header indicator

How is a multicast detected and routed.

### Current Plan/Position

A separate routing header with a PEI indicates that the header is a multicast header with the real PEI inserted into the header later.

### Desired Position

A bit in the Header that indicates the multicast instead of a separate PEI.

Address+0		Address+1				Address+2				Address+3				Address+4				Address+5				Address+6				Address+7																																					
63	62	61	60	59	58	57	56	55	54	53	52	51	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00
Version HDR	HRD Size	CRC				Hop Count				Traffic Class	Credits Required				Header Function	B	Q	M	U	S	D	Turn Pool/Multicast Index																																									

Header size can remain a type 00 as more bits for the Multicast Index is probably not needed.

### Rational

With multicast there is not turnpool to be considered, only the address into the multicast routing table. Protection of this table entry should be raised. An error in propagating the table can cause the sending of the multicast packet partially to the desired location and partially to unintended nodes that do not know what to do with the data.

Multicast requires all the Data packet header information including the keys and other protection. A single bit in the routing header is efficient use of a bit to solve the problem.

## Issue/Concern – Route Header Credit Length

This is the amount of credit available for the link or the path, or the endpoint limit?

### Current Plan/Position

Maximum credit is 65 units of 64 bytes each. This is a maximum transfer of 4160 bytes.

### Minimum Required Position

A credit length of 67 units of 64 bytes each. This is a maximum transfer of 4352 bytes to support the added overhead.

### Desired Position

An 8 bit field of units of 64 bytes with a multiplication factor defined in the two high order bits. This would allow the growth of the packet frames and will track changes switch development.

Address+0 63 62 61 60 59 58 57 56		Address+1 55 54 53 52 51 50 49 48 47 46 45 44 43 42 41 40		Address+2 39 38 37 36 35 34 33 32		Address+3 31 30 29 28 27 26 25 24		Address+4 23 22 21 20 19 18 17 16		Address+5 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00				
Version HDR	HRD Size	CRC		Turn Pointer		Traffic Class	Credits Required		Header Function	B Q	M I P S E C	D	Turn Pool	

This 8 bit Credit Requirements field should use the same type of multiplication in the high order bits that are used in the current scheme. CreditRequirement(7,6) act as a multiplier on the low order 6 bits with:

5. CreditRequirement(7,6) = 00 being a multiplier of 1
6. CreditRequirement(7,6) = 01 being a multiplier of 4
7. CreditRequirement(7,6) = 10 being a multiplier of 16
8. CreditRequirement(7,6) = 11 being a multiplier of 64

In addition to this calculation a constant adder of 3 to account for 64 bytes of header, 64 bytes of addresses, command, time, key, CRC, and 64 byte for data alignment and future Routing Header or Packet Header expansion.

To move 4096 bytes of data CreditRequirement(7,0) of  $00111111_b + 3$  or  $64 + 3 = 67$ .

To move 65536 bytes of data CreditRequirement(7,0) would be  $10111111_b + 3 = 1024 + 3 = 1027$ .

### Rational

An artificial limit based on what assumptions, PCI? Why? A larger credit limit register will support changing technology.

According to the switch manufacturers on the committee, the initial switches will support 256 byte packets at most, with 512 bytes being a possibility in a year or so. The upper bound in the current spec is not reachable at this time. As time and technology move forward we will need to increase this number. I do not expect that the 4096 max data packet will be so forever. Plan for the future and for the other formats that we will most likely need to support.

## Issue/Concern – Transfer Length

Is this the length of the current transfer or the length of the entire block to be transferred?

### Current Plan/Position

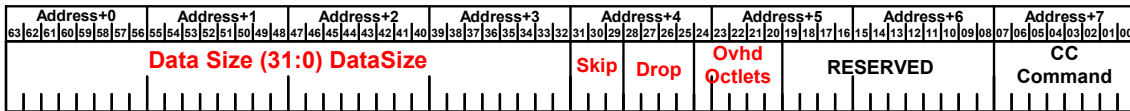
Maximum of 4096 bytes plus 64 bytes of overhead.

### Minimum Required Position

Maximum of 4096 bytes plus at least 192 bytes of overhead.

### Desired Position

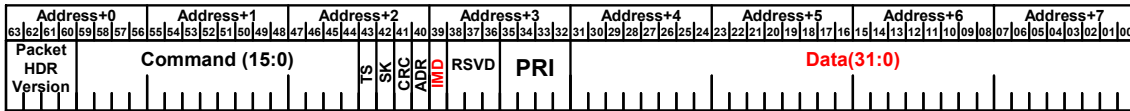
This be scalable to  $2^{32}$  for future and be in bytes.



Where:

1. DataSize(31:0) -> is the size of the valid data being transferred.
2. Skip -> is in range 0 ... 7 to indicate the offset to the first valid byte in the first octlet.
3. Drop -> is in range 0 ... 15 to indicate the number of trailing bytes are not part of the data transfer but needed for octlet alignment. The reason for up to 15 is too capable of adjusting the data block to be in units of 16 bytes, 128 bits to match encryption block sizes.
4. Ovhd Octlets -> Overhead octlets, the number of octlets, eight byte blocks, of overhead data is included in transfer. This will include routing header overhead and data packet overhead.

Very short data items, 4 bytes or less can be treated as immediate data and included on the command octlet if the tag field is not needed. The IMD bit is set to one to indicate the inclusion of immediate data which is always treated as a 32 bit quantity.



Immediate data forms the smallest packet in the fabric assuming the other optional features of the packet are not needed.

### Rational

This limit is artificial and based on limits specified in the PCI Express documentation. This limit does not need to be tied to the PCI formula created many years ago. The long field would allow the current limits and future limits to be accommodated.

SAR has not yet been defined and will make use of this limit.

This is negotiated as part of the MTU arbitration. There is no reason to limit this by restricting the field length. Take bit out of the initial turnpool space if needed.

## **Issue/Concern – Data Alignment/Byte Lanes/Endianness**

The Fabric does not have any Byte Lanes only the Endpoints do. There is however a need for consistent ability to know the order of data being transferred over the link. The fabric should be of a consistent endianness. This should be Big Endian as supported by IB, RapidIO, SCI, RIO and others. All descriptions should be shifted to Octlet based descriptions.

### ***Current Plan/Position***

Current position is defining 8 starting byte lanes and 8 trailing bytes lanes with individual enables. Descriptions are 32 bit (Dword in Intel) Quadlet (international) oriented.

### ***Minimum Required Position***

Change descriptions to Octlet (8 byte) based nomenclature. Maintain Big Endian byte ordering.

### ***Desired Position***

Byte Lanes should be removed. Use 3 bit skip and 4 bit drop descriptor to define placement of useable data in the transfer block octlet based. This could be inferred by the starting address and byte count. Minimum transfer is 1 Octlet. Maintain Big Endian byte ordering.

### ***Rational***

AS transfers are concerned with movement of data from one endpoint to one or more other endpoints. Each of the endpoints may chose to treat the data in a different manner. Endianness is in the endpoints point of view and any required conversions should be done at the endpoint. Byte Lanes are a PCI function and should not be considered in AS. Each PCI translator will adjust the received data to match the endpoint requirement. PCI and other busses may have one byte, doublet, quadlet, or octlet bus width. AS has no inherent knowledge of the bus structure of the endpoints and should not. It is necessary to detect, and properly handle, non-octlet aligned data. This selection of valid bytes is easily accomplished with the addition of a skip and drop field requiring a total of 7 bits of information. The alternative is to extract the information based on the starting address and the length. Octlet based description is helpful in current and future 64 bit register operation.

## Issue/Concern – Terminology

Adopt Normal International adopted terminology for the specification for octet, doublet, Octlet, Quadlet, etc. Word and Dword can be misinterpreted.

### **Current Plan/Position**

Minimal descriptions in document

### **Minimum Required Position**

Align nomenclature with internationally accepted standards.

### **Desired Position**

Add some of the definitions include above and seek others that describe the draft using widely used terms. Choose ISO/IEC naming conventions wherever possible.

### **Rational**

Less work later in the standards process and clearer presentations now.

## Issue/Concern – Security Key Length

Keys to restrict access and grant permissions to data structures and nodes.

### **Current Plan/Position**

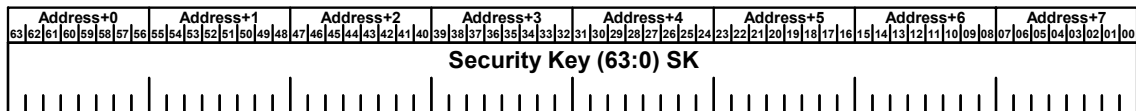
One dedicated 16 bit key. This may be expanded to 32 bits by consuming the current source address bits.

### **Minimum Required Position**

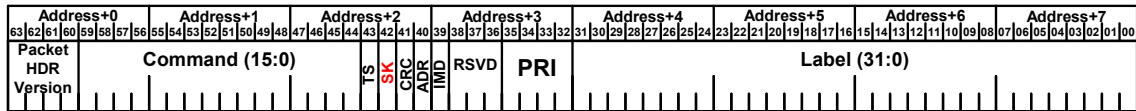
At least two 32 bit keys per the InfiniBand model

### **Desired Position**

A 64 bit key string.



This could be broken into 2 quadlet keys if required, or the high order quadlet ignored in some applications. This should be octlet aligned and controlled by the SK key in the command word.



### **Rational**

Keep the formats 64 bit (octlet) aligned  
 Current and new processors handle 64 bit data naturally.  
 More secure than 32 bit keys in the growing climate of increased security. A 64 bit number is more difficult to get to the wrong person. This is also more difficult for a rogue node to accidentally get a correct 64 bit key than a 32 of 16 bit key.

## Issue/Concern – Optional CRC

This is the PACKET CRC.

### ***Current Plan/Position***

Optional and defined by the packet.

### ***Minimum Required Position***

Optional ECRC controlled by a bit in the command octlet.

Address+0 63 62 61 60 59 58 57 56		Address+1 55 54 53 52 51 50 49 48 47 46 45 44 43 42 41 40			Address+2 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16				Address+3 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00			Address+4 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00		Address+5 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00			Address+6 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00			Address+7 15 14 13 12 11 10 09 08 07 06 05 04 03 02 01 00																	
Packet HDR Version	Command (15:0)															TS	SK	CRC	ADR	IMD	RSVD	PRI	Label (31:0)														

### ***Desired Position***

Mandatory use of the CRC on all packets. This CRC will include the packet header overhead and the data.

### ***Rational***

Reliability, Reliability, Reliability. This will check the command block and all of the addressing information in the packet overhead prior to the decoding of the information. If the CRC is broken the packet is discarded. Currently there is no required way to know if the end to end packet is still intact other than each of the links thinks that there was no error in that link. This will significantly reduce the Undetected Error probability.

## Issue/Concern – Packet Tag Field Length

A Packet Header issue. How is this handled differently between switches and endpoints?

### Current Plan/Position

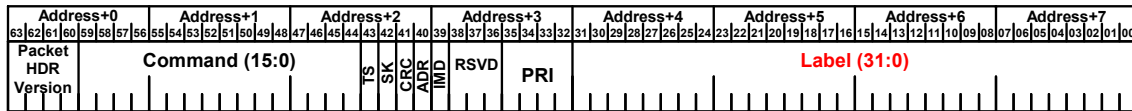
Current routing header has a transaction tag field of 5 bits.

### Minimum Required Position

16 bit field is marginal.

### Desired Position

32 bit packet tag field.



### Rational

The current limit of 31 outstanding packets is insufficient, unless this definition only applies from one switch to the NEXT switch.

The CURRENT turn pool allows a max distance along a path to be 31 hops through 1 bit switches. If the end to end tag field is to cover the outstanding packets it will need to be enlarged to at least twice that number. ACKs will take 63 hops (outstanding packets) to get to the endpoint and to return along the same path. The counter will have cycled several times before the Ack gets back. Delays in internal buffers in the switches will aggravate this problem.

Larger fabrics with more store and forward buffers will extend that need for a larger number of outstanding packets by an order of magnitude.

For IB an iWARP etc, the tag field is 32 bits.

Additionally, a receiving endpoint may wish to send back a message with a tag number that it wishes to acknowledge all packets up to and including that Packet. This will save a lot of traffic in the Fabric.

If this is a constantly incrementing number, the number must be large enough to prevent the misinterpretation due to recycled numbers. The rollover problem needs to be considered and properly handled.

## Issue/Concern – Route Header Multicast Fields

This is the fields defining the multicast index and multicast source.

### Current Plan/Position

Currently 12 bits of index into the multicast table, and 12 bits to identify the Source of the multicast.

### Minimum Required Position

Remove the source of the multicast and expand the Multicast Index.

### Desired Position

Use the available space in the route header for the Index – 24 bits. The source can be found in the delivered packet to the precision needed.

Address+0		Address+1				Address+2				Address+3				Address+4				Address+5				Address+6				Address+7																																					
63	62	61	60	59	58	57	56	55	54	53	52	51	50	49	48	47	46	45	44	43	42	41	40	39	38	37	36	35	34	33	32	31	30	29	28	27	26	25	24	23	22	21	20	19	18	17	16	15	14	13	12	11	10	09	08	07	06	05	04	03	02	01	00
Version HDR		HRD Size		CRC				Turn Pointer				Traffic Class		Credits Required				Header Function		B	M	D	Multicast Index																																								

When Routing Header Bit 26 is set to 1 the Routing Header Octlet bits 23:0 become the Multicast Index. The size of the Multicast Index table is vendor dependent, one is requires.

### Rational

Any pointer to the source must be the nodeID + offset. A hash of this number is not unique.

If the source is needed, it can be found in the addressing in the packet.

The 12 bit Source identifier is only useful is it is an index into a table of 64 bit nodeID and the 64 bit offset/internal node address.

Depending on the size of the fabric, a 12 bit index may not be sufficient for all of the multicast paths that may be established from the many attached nodes in this fabric.

Limitations in the implementation of the individual Switches will be noted and accommodated in the configuration information. The current plan is to setup for up to 4096 table entries per switch but implement 1. This same rule could also apply to the 24 bit index.

## **Issue/Concern – EVENT/Interrupt Multicasts**

Interrupts are a subset of events and must have a multicast capability

### ***Current Plan/Position***

None in current specification

### ***Minimum Required Position***

Interrupt event messages will support multicast using the same basic mechanism as all other fabric traffic.

### ***Desired Position***

Unicast and Multicast capability on all Events, including interrupts. Use the proposed Routing Header with the multicast bit.

### ***Rational***

Events, including interrupts need to be directed at a class of targets including the fabric master, the alternate fabric master and the designated interrupt/event handler(s).

The event generation source does not know which master is alive and which node may service the particular needs required by the event/interrupt in the system. A multicast will include all the nodes that are required to know about the event.

Events may be of such significance that there are effectively broadcast with the multicast mechanism to all nodes, or group of nodes.

A multicast event may alert all nodes to a specific problem or statement.

Time.

Events also need to identify the source of the event by the inclusion of the nodeID (eui-64) in the message. This allows a third party to be instructed to service the event.

## Issue/Concern – Packet Endpoint Priority

What priority systems are needed in AS.

### Current Plan/Position

Priority is currently handled by traffic class and VC.

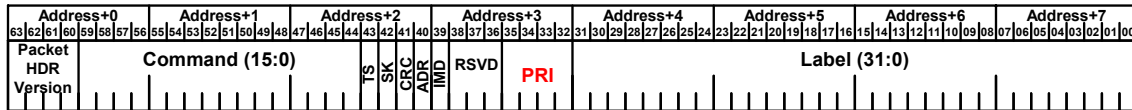
### Minimum Required Position

A priority system is needed in the fabric and in the endpoint packet system. The current TC and VC system is designed for operation in the switched environment of the center of the fabric.

We also need to design a priority that can be used by the endpoint in the endpoint’s environment.

### Desired Position

A priority field was added to the Command octlet in this proposal.



A 4 bit Priority field was added to the Command octlet at Bits 35:32 to provide a 16 level priority system for the delivered packets. This can be use as Virtual Channel indices or as direct priority.

### Rational

In addition to routing the traffic through the fabric and avoiding deadlock, we need a system to point to a priority in the delivered packet. The delivered packet is not interpreted by the switches for priority and the endpoints will have the Routing header stripped prior to interpretation of the data in the packet. Priority fields are needed in both.

## **Issue/Concern – Command Field Growth**

Leave room for the future.

### ***Current Plan/Position***

Use of the PEI as the command field.

### ***Minimum Required Position***

A command field of a minimum of 16 bits to encode multiple options in the packet header.

### ***Desired Position***

Please see the attached proposal for General Packet Header.

### ***Rational***

There are many definitions of packet encapsulations to come and we need to leave room for them. The proposed header command field defines only one of many possible command sets. Each set may have multiple options to use the remaining 12 bits in that command set.

## **Issue/Concern – PEI-8**

Routing through the fabric will require additional addressing for identification of valid receipt of a packet from another PCI Express node.

### ***Current Plan/Position***

Not yet well defined.

### ***Minimum Required Position***

Translation through a bridge, hopefully minimal, to convert the PCI Express packet to an AS fabric packet and then through a bridge again at the correct destination node to strip the AS specific packet wrapper.

### ***Desired Position***

Use the included proposal for packet header to transport the PCI Express packet from the source node to the destination node with validation and security.

### ***Rational***

PCI Express packets are well defined and NO knowledge of the AS addressing and transport Domain. These packets will need a wrapper to carry them through the AS fabric from a source to destination. I do not see any way that this can happen without the discovery and control provided by a bridge at each end. The PCI Express will need to use the bridge to configure the wrapper for delivery of a packet to the correct node.

## **Issue/Concern – R.A.S.**

Not yet spelled out.

### ***Current Plan/Position***

Not defined.

### ***Minimum Required Position***

Good measurement capabilities.

### ***Desired Position***

Good measurement capabilities.

### ***Rational***

RAS is one of the most important properties in modern fabrics. We should look at the AS specification from this point of view.

## **Issue/Concern – Heartbeat**

Different types of heart beat are needed in the fabric. Some of these will need participation from the switches.

### ***Current Plan/Position***

None established, but the problem is recognized.

### ***Minimum Required Position***

Heartbeats may be ordinary packets with unicast and multicast capability.

### ***Desired Position***

Heartbeats with timestamps are available. This also means that timestamps and a sense of time in the fabric are needed. This should be an Event which is just another send.

### ***Rational***

We need to know what died and what connections are no longer valid.

This will start the reconfiguration of routes through discovery to get from one node to another node by a new route from one nodeID to another nodeID.

## **Issue/Concern – Hot Plug**

How is a hot plug event handled?

### ***Current Plan/Position***

Being developed

### ***Minimum Required Position***

This needs to be non-disruptive to members of the fabric that do use the element being added or subtracted from the fabric. The issue of missing drivers and loss of control elements will probably be uncovered by heartbeat interactions.

### ***Desired Position***

Defined method of smoothly handling hot plug events in a continuously available system.

### ***Rational***

Hot plug events will be a regular part of the fabric operation.

## **Issue/Concern – Performance Monitoring**

We need built-in mechanisms to monitor performance and to support the tuning of the fabric for best aggregate performance based on priorities

### ***Current Plan/Position***

None defined

### ***Minimum Required Position***

Include time and other monitor conditions to include the end to end packet delays, queuing delays, hot spots in the fabric.

### ***Desired Position***

Include the time protocols and other measurement parameters.

### ***Rational***

This fabric will require continuous monitoring and optimization for best performance. Each fabric change thru additions to the fabric, broken links, and changing masters will require re-optimizations of the fabric.

## **Issue/Concern – Error Recovery**

How and from what.....

### ***Current Plan/Position***

Not well defined

### ***Minimum Required Position***

In an always alive system a method of fabric recovery and continued operation is required.

### ***Desired Position***

Continuous run systems require soft recovery from outages with workaround that leaves critical elements fully functional.

### ***Rational***

Continuous systems need to be continuous.