FC Speed Issues

- This presentation describes some basic issues to consider when developing a speed roadmap or strategy for FC
- A more detailed document has been posted to the T11 web site (04-096v0)
- Most of this presentation was derived from that document
- The goal of this presentation is to ensure that the basic architectures relating to speed are understood
- A few recommendations are made (not official HP positions, however)

What do we mean by the word “speed” anyway?

- The term ‘speed’ refers to the rate of data transport that is possible in the physical layer of the FC variant
- Unfortunately, even with this simplification, one must be very careful not to compare apples with oranges because one may legitimately speak of speed in several different ways

What do we mean by the word “speed” anyway?

- Speed comparisons must deal with the different physical architectures that are specified
- For example:
  - Simplex vs duplex
  - Single port vs dual port
  - Four lane vs single lane
- And possibly other differences like:
  - Encoding schemes
  - Framing and signaling overhead
  - Workload properties
  - Speed / latency tradeoffs

What do we mean by the word “speed” anyway?

- It has been common practice within the FCIA for several years to report the aggregate theoretical maximum payload data rate for a duplex port as the port speed
- That scheme reports twice the simplex payload data rate, accounts for framing and signaling overhead and assumes balanced workloads for both sides of the port
- While not intended as a criticism of this FCIA practice, this presentation differs by using only the simplex port assumption to report speed numbers
- One may readily calculate the speed under different workload and device configuration conditions from the simplex speed

What do we mean by the word “speed” anyway?

- Speed comparisons, even for simplex links, are easy to misstate or misinterpret because there are several numbers that may be used to describe the speed of a simplex link:
  - Line rate: encoded bit rate on the line (Baud)
  - Bit level data rate: number of bits transmitted per second before encoding (bits / sec)
  - Average payload data rate: the number of Bytes transmitted per second in the payload (bytes / sec)
- The payload data rate measured in Bytes per second is the same with or without encoding – what goes in must come out (assuming no buffering)
  - Before encoding there are 8 bits per byte, after encoding the number of encoded bits per byte depends on the encoding scheme
- Benchmarks usually report the payload data rate in Bytes per second

What do we mean by the word “speed” anyway?

- There is some overhead in transmission that does not contribute to the payload data rate
- For example, inter-frame primitives, start of frame, headers, CRC, and end of frame are all overhead
- The percentage overhead depends on the relative size of the payload
  - Large payloads have relatively less overhead (but may impact latency)
  - In this presentation it is assumed that this average overhead is 6.25% for all links
- So a 1.0625 GBaud line rate with 8b10b encoding is assumed to produce 1.0 Gbits/sec of encoded payload data -- This supports 800 Mb/s of unencoded payload data that produces 100 MBytes/sec of payload data

8.5 Gig compared to 10 Gig

- An 8.50 GBaud line rate with 8b10b encoding is assumed to produce 8000 Mbits/sec of encoded payload data -- this supports 6400 Mb/s of unencoded payload data that produces 800 MBytes/sec of payload data
- A 10.51875 GBaud link with 64b66b encoding is assumed to produce 9900 Mbits/sec of encoded payload data -- this supports 9600 Mbits/sec of unencoded payload data that produces 1200 MBytes/sec of payload data
- So a proper comparison between 8.5 G and 10 G (as presently specified) is 800 MBytes/s compared to 1200 MBytes/s (a 50% difference – not a 15% difference)

Workload types may map to specific speeds

- Issues like workload, balance between ports of a duplex link, frame size, dual port etc. may map to a specific speed
- For example the traffic between switches of ISL's (inter switch links) may tend to be balanced between the duplex ports while the traffic on an HDD port may have a 4 to1 read to write ratio
- If 10G is most common in ISL's then 10 G maps to balanced workloads and 4 G maps to an unbalanced workload

Simplex speeds in current FC standards

Line rate (GBaud per lane)	Payload data rate (GBytes / s per simplex port)	Encoding scheme used	Physical architect-ture for simplex connection	Variant(s)	Standard
1.0625	0.1	8b10b	SSS	all	FC-PI
2.125	0.2	8b10b	SSS	all	FC-PI
4.25	0.4	8b10b	SSS	all	FC-PI optical FC-PI-2 optical and electrical
3.1875/ lane, 12.75/ 4 lanes	0.3 / lane, 1.2 / 4 lanes	8b10b	4LSP	4LSP variants only	10 GFC optical FC-PI-2 electrical
10.51875	1.2	64b66b	SSS	SSS optical variants only	10 GFC

Simplex speeds in future FC standards?

Line rate (GBaud per lane)	Payload data rate (GBytes / s per simplex port)	Encoding scheme used	Physical architect-ture for simplex connection	Variant(s)	Standard
6.375	0.6	8b10b	SSS	all	???
8.5	0.8	8b10b	SSS	all	FC-PI-4,5
10.51875	1.2	64b66b	SSS	all	FC-PI-3
12.75	1.2	8b10b	SSS	all	FC-PI-3
17	1.6	8b10b	SSS	all	TBD
34	3.2	8b10b	SSS	all	TBD
42.075*	4.8	64b66b	SSS	all	TBD
68	6.4	8b10b	SSS	all	TBD
84.15*	9.6	64b66b	SSS	all	TBD
136	12.8	8b10b	SSS	all	TBD

* wrong number in 04-096v0 these are correct

Other important considerations for speed strategies

- Agility of ports
- Implementation mode assumptions
- Backward and forward compatibility and technology leverage
- Value added proposition
- Need for a higher speed
- Sequence of speed introductions (same SAN, back end)
- Risks of having more speeds
- How much interoperability is really desirable?
- Error recovery scaling and costs
- Transition process to introduce a new speed

Physical issues related to speed

- Link Length
- Connectors
  - SSS variants
  - 4LSP variants
- Removable PMD modules
- Agility
- Cable size issues
- Test equipment
- Testing methods

Additional physical architecture considerations

- Device physical port architecture (e.g. dual port)
- Port density issues
- Powered ports (for external circuits)
- Safety issues (e.g. multiple optical fibers per link)
- EMI issues

What is the need for a higher speed?

The following list offers some plausible reasons:

- Need to keep ahead of competing technologies in the marketing arena
- Higher bandwidth due to workload changes
- Higher bandwidth to keep up with quality of service demands
- Higher bandwidth to enable SAN's with more levels of switches and for SAN to SAN connectivity
- Higher speed to keep ahead of the HDD media transfer rate (especially for loops)
- Higher bandwidth to keep the FC links from being a gate to servers, controllers
- Requirement for the lowest possible latency
- Backup/restore time

- At the moment the first item in the list is sufficient to make the effort worthwhile.

Transition strategy

- Both SAN and back end configurations are built from multiple port to port connections (usually duplex) between devices
- For any port to port connection to be able to function, the ports on both sides of the link must be compatible with each other
- This somewhat obvious fact has significant ramifications for the transition strategy
- If ports are not agile (i.e. operate only at a single speed) then the situation is pretty simple:
  - The devices on both sides of the link must both support the new speed and be available at the same time in order to form the link
  - In other words, a fundamental requirement exists that the market timing for the devices on both sides of the link be in sync with each other.

Market Synchronization

This synchronization consists of several dimensions:

- Development maturity schedule including all features for the devices, not just the speed
- Volume availability schedule
- Cost/price maturity schedule
- Infrastructure availability schedule (training, test equipment, cables, installation procedures, verification and debugging tools, etc.)

Market Synchronization for ISL's

- One important connection in a SAN fabric that automatically has these synchronization properties is the inter switch link or ISL between identical fabric switches from the same supplier.
- Since an ISL is between two identical ports from theh same supplier, market synchronization is delivered along with the switches
- Virtually every other connection in an FC system has different devices on each side of the link
- These connections require the synchronization to be developed via close cooperation between a number of suppliers and users
- So the essence of a successful transition strategy requires that the ports on both sides of every link in the FC system come together at the same time and be compatible with each other at all times

Assumptions about the properties of FC system components

- Being intrinsically multiple port devices, fabric switches may be capable of any speed, physical architecture, or protocol on different switch ports
- HDD controllers are also intrinsically multiple heterogeneous port devices on both the SAN and the back side
- Servers only connect to the in-band FC world through HBA's but may have multiple HBA's and any number of other ports available on the server -- Servers are therefore effectively multiple heterogeneous port devices
- Some multiple port devices (fabric switches, servers, HDD controllers, JBOD enclosures, etc) may connect out of band to the FC management world via an Ethernet port on the device.

- HBA'a and HDD's are generally single or dual port devices.

- Back end interconnect components such as loop switches and PBC's are multiple port devices but usually have the same properties on all ports due to the single chip implementations often used.

MHP vs SP components

- The distinction between multiple heterogeneous port (MHP) devices, multiple homogenous port devices, and single (SP) or dual port devices can be very important when considering the transition strategy to a new speed
- MHP devices have a basic degree of freedom that is not available to SP or dual port devices
  - MHP devices can offer ports that are capable of the new speed but are not backwards compatible
  - These ports are used when devices capable of the new speed become available
  - This degree of freedom allows MHP devices to ship with some independence from the schedules of other devices
- For SP devices or multiple homogeneous devices that are not backwards compatible there is simply no market for the devices until the other end becomes available
- For single port, non-backward compatible devices one can paint a picture where it is impossible to introduce a new speed because both sides are waiting for the other and neither will proceed until the other is in place
- The situation with MHP devices is not all wonderful either since significant complexity is required within the MHP devices to support multiple different ports and one would like to use all the ports in the MHP device
  - Having the new speed capability on some ports effectively raises the price for the old speed ports
- MHP devices can probably survive a speed transition without backward compatibility but it is likely that SP, dual port, and multiple homogeneous port devices may not.

Starting Point for a Transition to a New Speed

- Creating a transition strategy to a new speed is a whole lot easier if there is a known starting point
- For HDD's up to 4 G the speeds are known, the connectors are known, the encoding is known – the biggest unknown is whether point to point with loop switches (or back end fabric like switches) or just plain loops will be dominant at 4 G – not a big deal either way
- For SAN's the picture is not nearly so clear especially for the 10 G installations
- The 4 G SAN was put on the FCIA roadmap largely because of the presently more or less universal 2 G SAN infrastructure – not because of a 4 G HDD speed choice
- Key question: what will be the next more or less universal SAN infrastructure? 4 G or some version of 10 G

And what about that 10 G infrastructure?

- The 10 G infrastructure, what there is today, seems to be significantly fragmented
- Among the fragments are single serial stream (SSS), 4 lane serial/parallel (4LSP) for both electrical and optical
- Within optical there are MM optical (several variants), and SM optical (several variants)
- The physical packaging available spans several types
- And the cabling may or may not be diverse
- And the ports that need 10 G first are not clearly specified by FCIA (recall ISL's can be a special case where the infrastructure comes almost for free with the switches that have the ISL ports)

So…..

- We should not try to base a speed strategy on commonality between SAN fabrics and back end applications – problems are too different, schedule requirements are too different and connections between the back end and the SAN fabrics are thru controllers or servers that allow/demand this separation
- The first order of business for 10 G based SAN infrastructure is not to define a speed strategy but rather to define no more than three variants that are on the FCIA roadmap: one SM optical, one MM optical, and one electrical
- This will define a much better platform to develop a credible SAN speed strategy beyond today
- The speed strategy for HBA's should be based on:
  - the issues facing a SP/dual port device
  - the capabilities of servers
  - the actual established SAN infrastructure that emerges after 2 G
- The speed strategy for switches and controllers should be based on the issues facing MHP devices and on the actual established SAN infrastructure that emerges after 2G
- The speed strategy for HDD's and single chip back end components such as loop switches should be based on the issues facing a SP/dual port device that uses a 4 G electrical infrastructure as the starting point

“Conclusions?”

- For HDD’s and single chip SP/dual port back end devices, immediately establish the supported variant as 8.5 G electrical single serial stream 8b10b encoding and based on backward compatibility with present intra-enclosure 4 G electrical (including connectors, encoding, interconnect requirements etc.)
- Clearly define the roadmap for 4 G and 10 G SAN infrastructures including variants allowed, order of introduction into the SAN (ISL, HBA, SAN to SAN, SAN appliance, core switch, edge switch etc), schedule, and projected volumes for each
- Do not attempt a SAN speed strategy beyond 4 / 10 G until the above item is completed
- Recognize the fundamental differences between MHP and SP devices in the transition to a new speed
- Look for leverage only from the established FC base and be very careful to determine if beneficial leverage actually exists