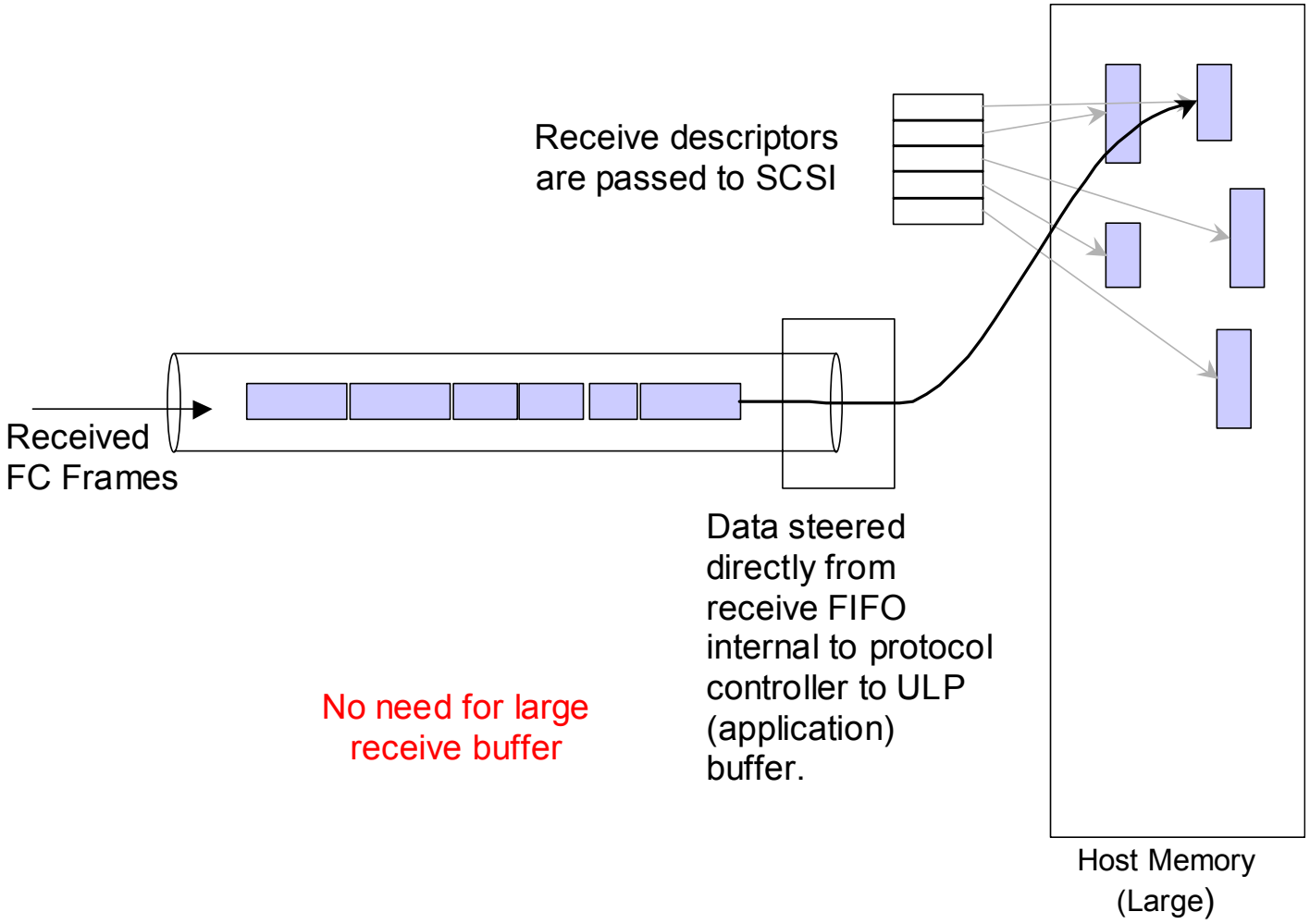
Matt Wakeley
27 June, 2001

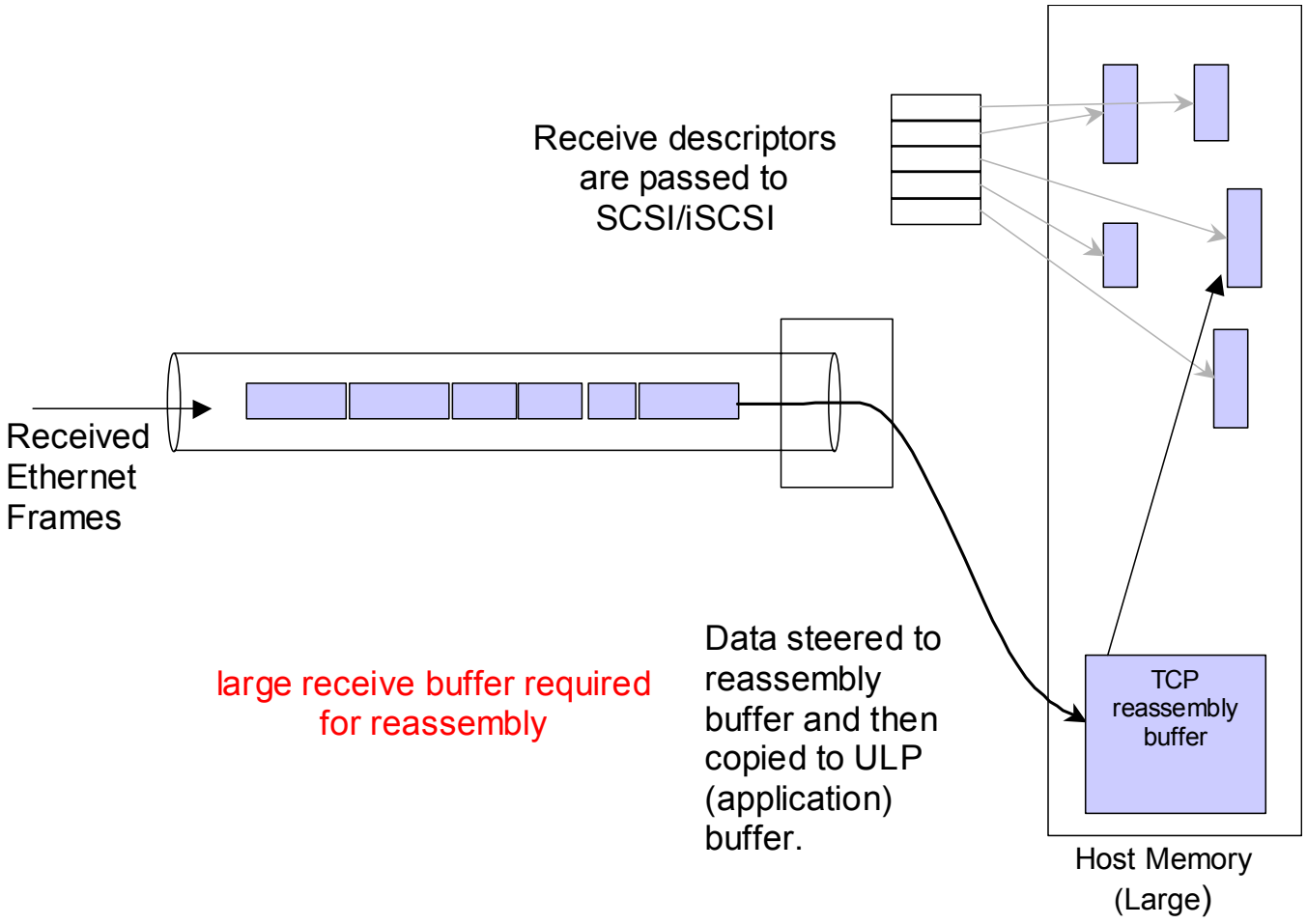
iSCSI Framing Presentation

Agilent Technologies

Storage Paradigm



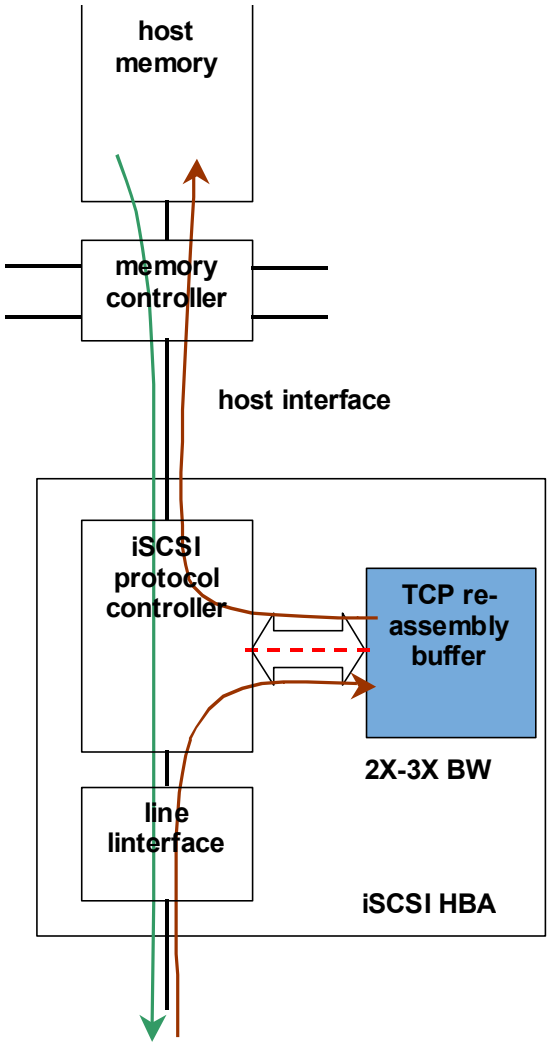
Network Paradigm



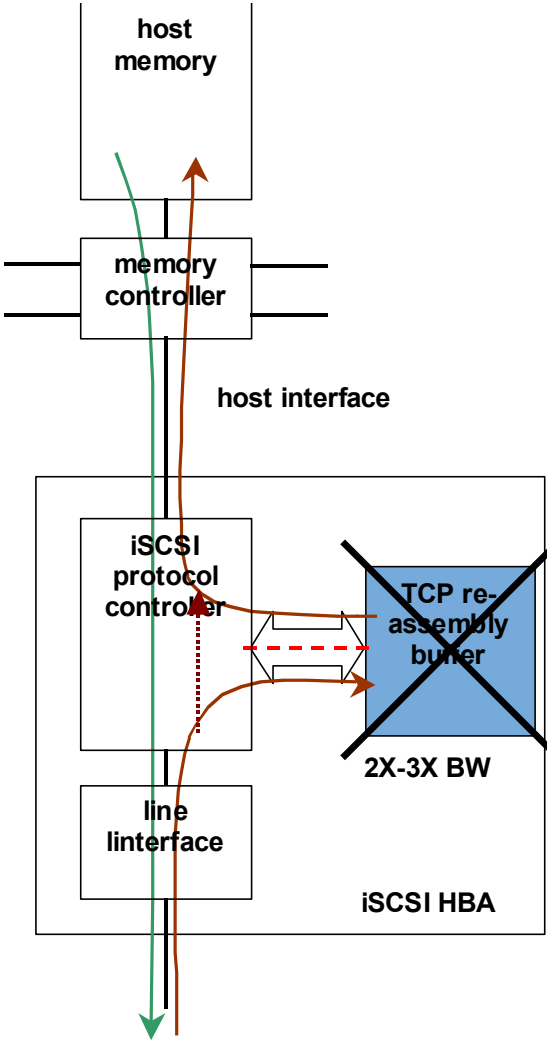
iSCSI Requirements

1. **iSCSI solutions must not utilize more of the system CPU and/or data busses than existing storage interconnect technologies.**
2. **iSCSI solutions must not cost more than existing storage interconnect technologies.**
3. **iSCSI solutions must have competitive latency and throughput to existing storage interconnect technologies.**

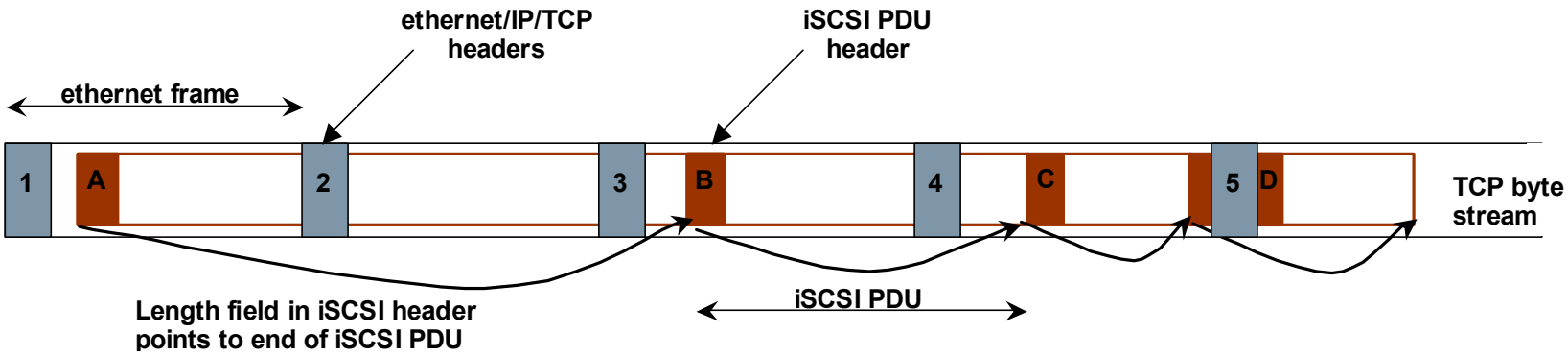
Solution to Requirement #1



What about Requirements #2 and #3?



The Framing Problem



How much memory is required?

Calculation of TCP packet buffer memory requirements for LFNs									
References:									
	"The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm",								
	Matthew Mathis, Jeffrey Semke and Jamshid Mahdavi (equation)								
Assumptions:									
	Assumptions made by reference #1 cited above								
	BER of LFN on order of 1e-11								
Using the following equation:									
	BW < (MSS/RTT)*1/sqrt(p)								
where:									
	BW is link bandwidth								
	MSS is maximum segment size in bits								
	RTT is Round Trip Time in secs								
	p is random packet loss at constant probability								
BER for a modern data center serial link is				1.00E-12					
BER for a modern LFN is realistically				1.00E-11					
1Gbs =	1.00E+09								
10Gbs =	1.00E+10								
Speed of light in fibre =			5.00E-06	sec/km					
A BER of	1.00E-11	will be used for this analysis…							
MSS =	1518	bytes *	8	bits/byte =	12144	bits			
p = BER * MSS =		1.21E-07							

[Scenario #1 - Network diameter at 10Gb/s]

BW =	1.00E+10	bits/s (10Gb/s)							
RTT = MSS/((10Gb/s)*sqrt(p)) =	3.48E-03								
network diameter = RTT / (speed of light in fibre) =				697	km or		348	km (one way)	
Summary - a 10Gb/s connection cannot go farther than				348	km assuming a realistic BER of				1.00E-11
The window size is RTT * 10Gbs / (8 bits/byte) =				4.36E+06	bytes =		4.2	MB	
A packet buffering solution would need approximat				**4.2**	**MB of buffering in this scenario.**				

[Scenario #2 - Network diameter at 1Gb/s]

BW =	1.00E+09	bits/sec (1Gbs)							
RTT = MSS/((1Gb/s)*sqrt(p)) =	3.48E-02								
network diameter = RTT / (speed of light in fibre) =				6970	km =		3485	km (one way)	
Summary - a 1Gb/s connection cannot go farther than				3485	km assuming a realistic BER of				1.00E-11
The window size is RTT * 1Gbs / (8 bits/byte) =				4.36E+06	bytes =		4.2	MB	
Since we have a 10Gb/s NIC and pipe, we can support 10 1Gb/s									
connections and therefore 10 windows of		4.2	MB for a total of				42	MB	
A packet buffering solution would need approximat				**42**	**MB of buffering in this scenario**				

[Scenario #3 - Around the world]

The circumference of the globe at the equator is approximately					40000	km			
(we need to consider round-trip delay to keep the pipe full)									
RTT = distance * speed of light in fibre =		0.2000	sec						
BW < (MSS/RTT)*1/sqrt(p) =	1.74E+08	bits/sec =		174	Mb/sec				
The window size is RTT * BW / (8 bits/byte) =				4.36E+06	bytes =		4.2	MB	
Since we have a 10Gb/s NIC and pipe, we can support 10Gbs / BW =						57			
174	Mb/sec connections and therefore			57	windows of			4.2	MB
for a total of	238	MB							
A packet buffering solution would need approximat				**238**	**MB of buffering in this scenario.**				

Packet buffering at 10Gb is expensive, both from a cost and board space perspective

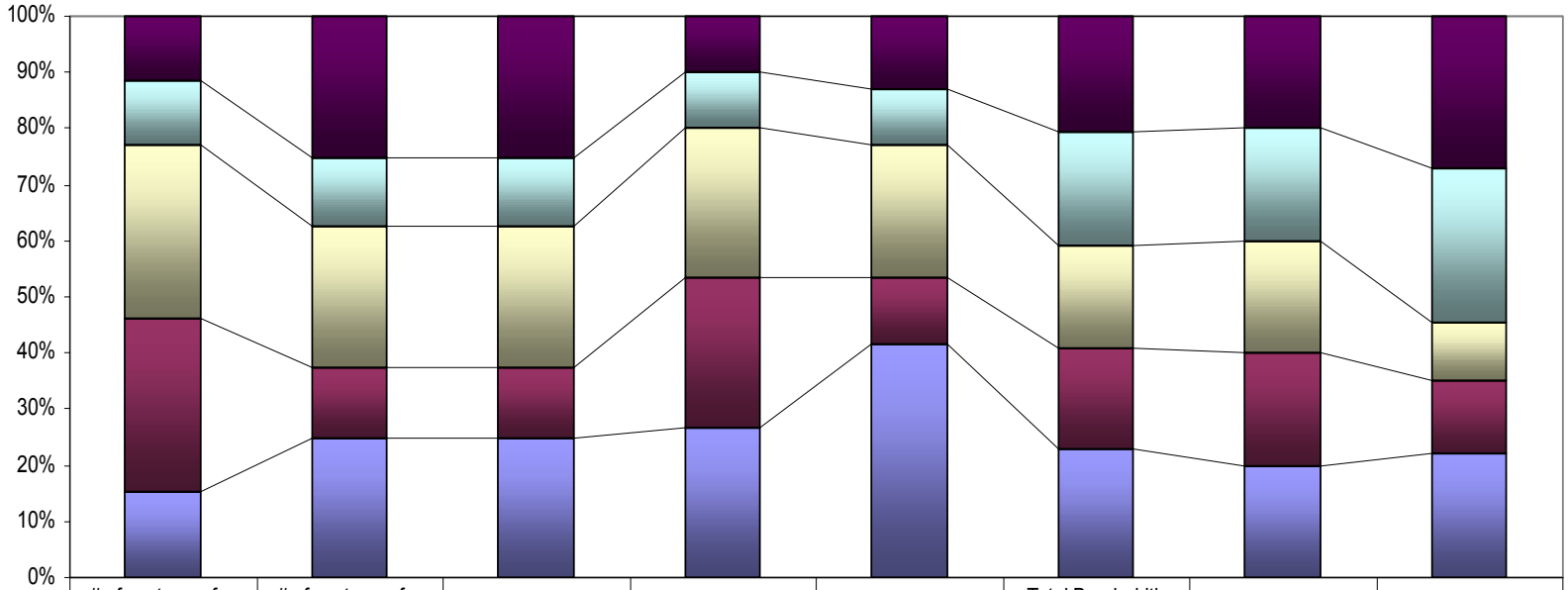
- **Packet buffering results in:**
  - **High pin count on protocol chips due to packet buffer interface with large bus width requirements**
  - **Higher parts count than competing technologies (FC) [memory]**
  - **Higher solution footprint than competing technologies (fc) [memory]**
  - **Higher power requirements (to drive all the extra parts)**

How much does Packet Buffering cost?

Packet Buffer Analysis

Miscellaneous Parameters							
Link Throughput (Gb/s):			1.250E+09				
Buffer Interface Bandwidth reqt			5.000E+09	// 4x link rate: http://www.eetimes.com/story/OEG20000619S0011			
Buffer Interface Storage reqt			2.684E+08	// 256 MB	// http://www.ezchip.com/html/linktotech.html		
Bits/byte	8	Bytes/K	1024	Bytes/M	1048576		
SGRAM bits/part:			32				
DDR SDRAM bits/part:			16				
RDRAM bits/device:			16				
DDR SDRAM BW Derating Factor:			80%				
RDRAM BW Derating Factor:			90%				
RAC Cost:			$5				
RAM Types							
		BW per pin (b/s)	BW/part (B/s)	Package Area (cm^2)	Storage (MB)	Cost/part	Add'l Cost
Graphic DDR SDRAM (x32)		5.000E+08	1.600E+09	3.9904	16	$11	
Commodity DDR SDRM (x16)		4.000E+08	6.400E+08	2.2586	16	$5	
Commodity DDR SDRM (x32)		4.000E+08	6.400E+08	2.2586	32	$13	
Direct RDRAM		1.066E+09	1.919E+09	1.2463	16	$13	$5.00
Direct RDRAM		1.066E+09	1.919E+09	1.9186	32	$26	$5.00
Packet Buffer		No Parity or ECC!					
		SGDRAM	DDR SDRAM (32M)	DDR SDRAM (16M)	RDRAM (32M)	RDRAM (16M)	
# of parts req for BW		4	8	8	3	3	
# of parts req for Storage		16	8	16	8	16	
Solution part count		16	8	16	8	16	
Bus width (bits)		128	128	128	48	48	
Total Area (cm^2)		63.8464	18.0688	36.1376	15.3488	19.9408	
Total Bandwidth (B/s)		6.400E+09	5.120E+09	5.120E+09	5.756E+09	5.756E+09	
Total Storage (MB)		256	256	256	256	256	
Granularity (MB)		64	256	128	96	48	
Total Cost		$172	$104	$80	$213	$213	

Packet Buffer Analysis



	# of parts req for BW	# of parts req for Storage	Solution part count	Bus width (bits)	Total Area (cm^2)	Total Bandwidth (B/s)	Total Storage (MB)	Total Cost
RDRAM (16M)	3	16	16	48	19.9408	5.756E+09	256	$213
RDRAM (32M)	3	8	8	48	15.3488	5.756E+09	256	$213
DDR SDRAM (16M)	8	16	16	128	36.1376	5.120E+09	256	$80
DDR SDRAM (32M)	8	8	8	128	18.0688	5.120E+09	256	$104
SGDRAM	4	16	16	128	63.8464	6.400E+09	256	$172

The Solution Has This Property...

Provide a mechanism for transferring data over TCP/IP in which packets are sufficiently self describing that suitably designed NICs can place data directly in the ultimate receive location so that data copies are not required even in the midst of packet loss.

Potential Solutions

- Inband (TCP unaware):
  - Periodic marker
  - Special characters (s/w intensive)
  - Fixed length PDUs

Out of band (TCP aware):

  - Urgent pointer (rejected by IETF)
  - TCP options to allow PDU alignment with TCP header (rejected by IETF)
  - ULP Framing (WARP) (cannot be mandated by iSCSI spec)

Potential Solutions cont'd

- **Use a different transport:**
  - **SCTP**
  - **UDP**

Recommended Solution: Mandate implementation of Markers

- Markers are the only viable solution because:
  - Markers are embedded within the TCP stream, and all TCP stacks can implement them
  - ULP Framing cannot be mandated by the iSCSI spec, because ULP Framing is a change to the stack and cannot be mandated

- Markers must be mandatory to implement, and mandatory to be provided if asked for in the negotiation process

If Markers are not Mandated: Recommend NO framing solution(s)

- **If no framing solution is mandated, solutions will have to implement TCP packet buffering anyway to accommodate those solutions that do not provide framing – thus, framing will not be used anyway.**