



Flash Memory Summit

NVMe performance optimization and stress testing

Isaac Livny
Teledyne Corporation

Santa Clara, CA
August 2017



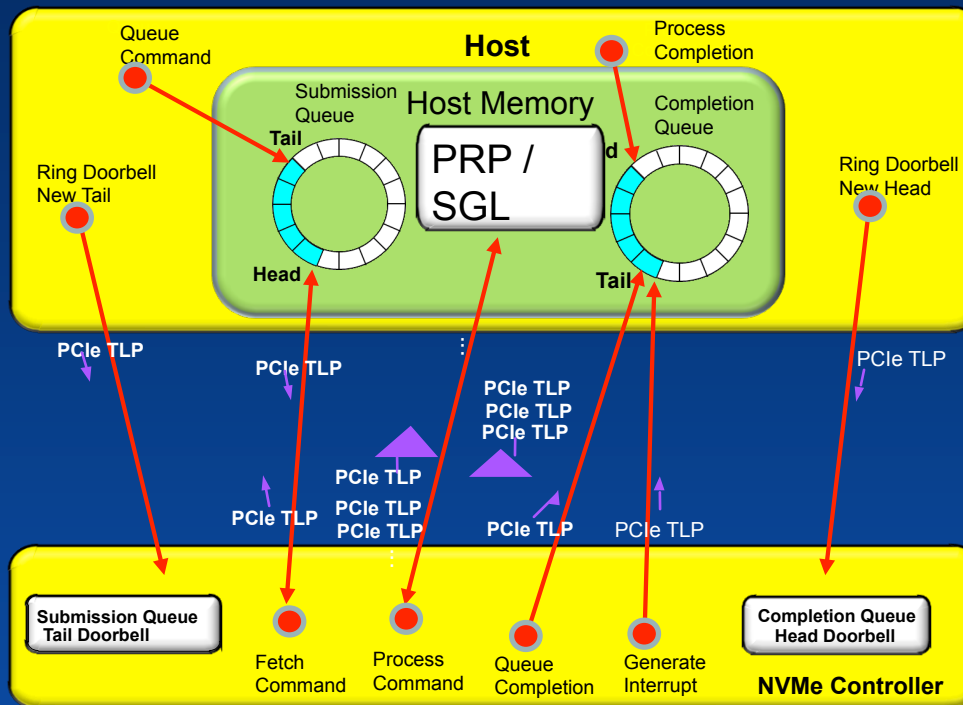
Flash Memory Summit

Agenda

- NVMe / NVMoF transfer overview
- PCIe performance analysis
 - NVMoF over CNA example
- NVMe performance analysis
 - LBA distribution analysis
 - Conditional performance analysis using scripting
- Stress testing using traffic generation
 - Script examples



NVMe complete transfer





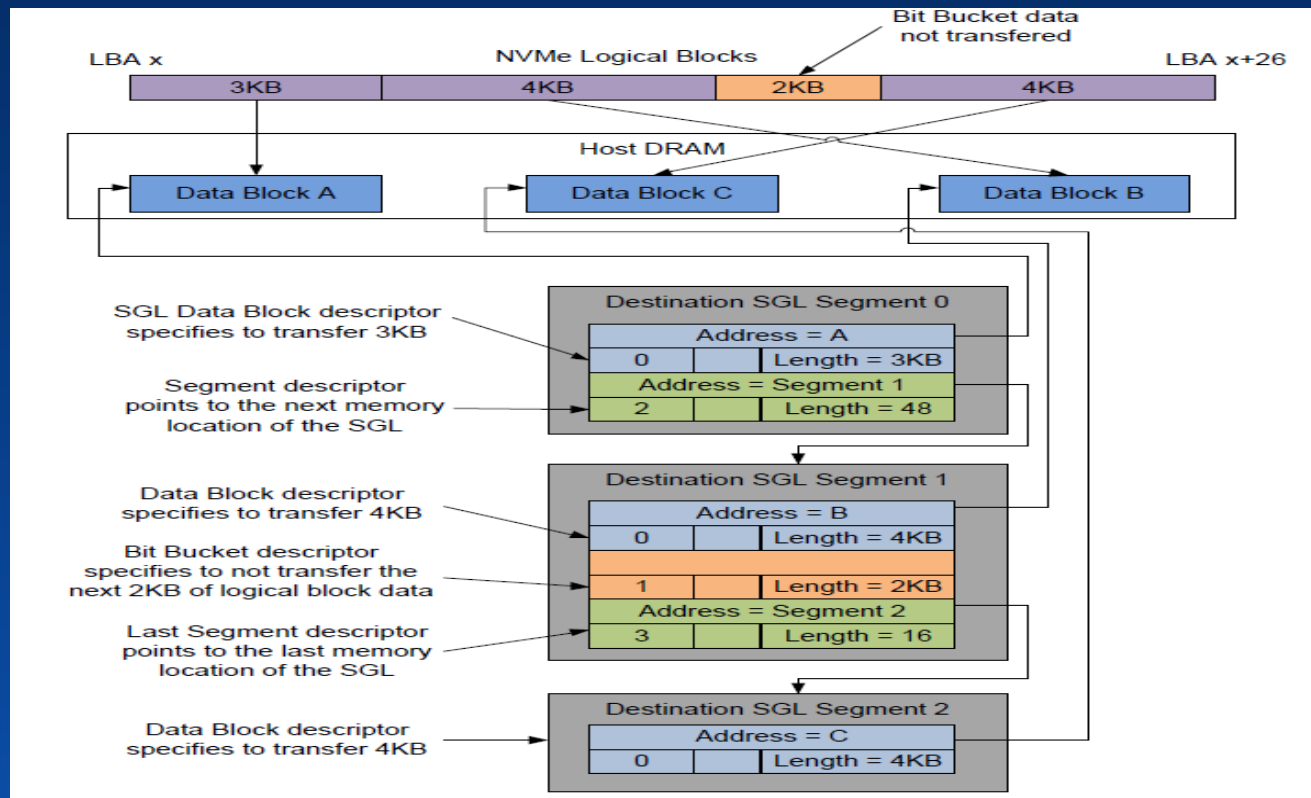
Flash Memory Summit

Each PRP data line in NVMe transaction view corresponds to pointer in a PRP list

NVMe Cmd	OPC	SQID	CQID	CID	Data	MPTR	PRP1	PRP2	SLBA	NLB	PRINFO	FUA	LR	DSM	ACCF	AC		
56	D	Read	0x0001	0x0001	0x0033	32768 dwords	00000000:00000000	00000000:82403818	00000000:87A1E000	00000000:000E3980	0x00FF	0x0	0	1		No frequency information provided		
NVMe 414	H	Device ID	QID	CID	IO SQT	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp								
414	H	001:00:0	0x0001	0x0033	SOYDBL	0x0034	SAMSUNG MZVPV256HDGL-00000	1	443.750 ns	0028 . 187 249 604 s								
NVMe 416	H	Device ID	QID	CID	Address	IO SQT	OPC	FUSE	PSDT	CID	NSID	MPTR	Address	PRP1	Address	PRP2	Address	SLBA
416	H	001:00:0	0x0001	0x0033	00000000:87A20CC0	0x0034	Read	Normal operation	PRP	0x0033	0x00000001	00000000:00000000	00000000:82403818	00000000:87A1E000	00000000:000E3980			
NVMe 417	H	Device ID	QID	CID	Address	PRP List	Data Len	Address List	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp					
417	H	001:00:0	0x0001	0x0033	00000000:87A1E000	0x00000040	32 addresses	SAMSUNG MZVPV256HDGL-00000		5	678.500 ns	0028 . 187 250 868 s						
Split Tra	R+	8.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC	VD	Address	Status	Data					
1451	R+	x4	Mem	MRd(32)	000:00:0	000:00:0	3	0	0	0	87A1E000	SC	0: 82404000 00000000 82405000 00000000 82406000 00000000 82407000 00000000					
Split Tra	R+	8.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC	VD	Address	Status	Data	Metrics	# LinkTrans	Time Delta	Time Stamp	
1452	R+	x4	Mem	MRd(32)	000:00:0	000:00:0	4	0	0	0	87A1E040	SC	16 dwords	2	688.500 ns	0028 . 187 251 546 s		
Split Tra	R+	8.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC	VD	Address	Status	Data	Metrics	# LinkTrans	Time Delta	Time Stamp	
1453	R+	x4	Mem	MRd(32)	000:00:0	000:00:0	5	0	0	0	87A1E080	SC	16 dwords	2	686.500 ns	0028 . 187 252 234 s		
Split Tra	R+	8.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC	VD	Address	Status	Data	Metrics	# LinkTrans	Time Delta	Time Stamp	
1454	R+	x4	Mem	MRd(32)	000:00:0	000:00:0	6	0	0	0	87A1E0C0	SC	12 dwords	2	110.309 us	0028 . 187 252 920 s		
Split Tra	R+	8.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC	VD	Address	Status	Data	Metrics	# LinkTrans	Time Delta	Time Stamp	
1455	R+	x4	Mem	MRd(32)	000:00:0	000:00:0	7	0	0	0	87A1E100	SC	4 dwords	2	-27.309 us	0028 . 187 363 230 s		
NVMe 418	D	Device ID	QID	CID	Address	PRP Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp					
418	D	001:00:0	0x0001	0x0033	00000000:82403818	0x00000150	506 dwords	SAMSUNG MZVPV256HDGL-00000		19	782.000 ns	0028 . 187 335 920 s						
NVMe 419	D	Device ID	QID	CID	Address	PRP Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp					
419	D	001:00:0	0x0001	0x0033	00000000:82404000	0x00000400	1024 dwords	SAMSUNG MZVPV256HDGL-00000		40	1.649 us	0028 . 187 336 702 s						
NVMe 420	D	Device ID	QID	CID	Address	PRP Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp					
420	D	001:00:0	0x0001	0x0033	00000000:82405000	0x00000400	1024 dwords	SAMSUNG MZVPV256HDGL-00000		40	1.647 us	0028 . 187 338 352 s						
NVMe 421	D	Device ID	QID	CID	Address	PRP Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp					
421	D	001:00:0	0x0001	0x0033	00000000:82406000	0x00000400	1024 dwords	SAMSUNG MZVPV256HDGL-00000		40	1.653 us	0028 . 187 340 000 s						
NVMe 422	D	Device ID	QID	CID	Address	PRP Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp					
422	D	001:00:0	0x0001	0x0033	00000000:82407000	0x00000400	1024 dwords	SAMSUNG MZVPV256HDGL-00000		40	1.659 us	0028 . 187 341 652 s						
NVMe 423	D	Device ID	QID	CID	Address	PRP Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp					
423	D	001:00:0	0x0001	0x0033	00000000:82408000	0x00000400	1024 dwords	SAMSUNG MZVPV256HDGL-00000		40	1.649 us	0028 . 187 343 312 s						



SGL descriptor types





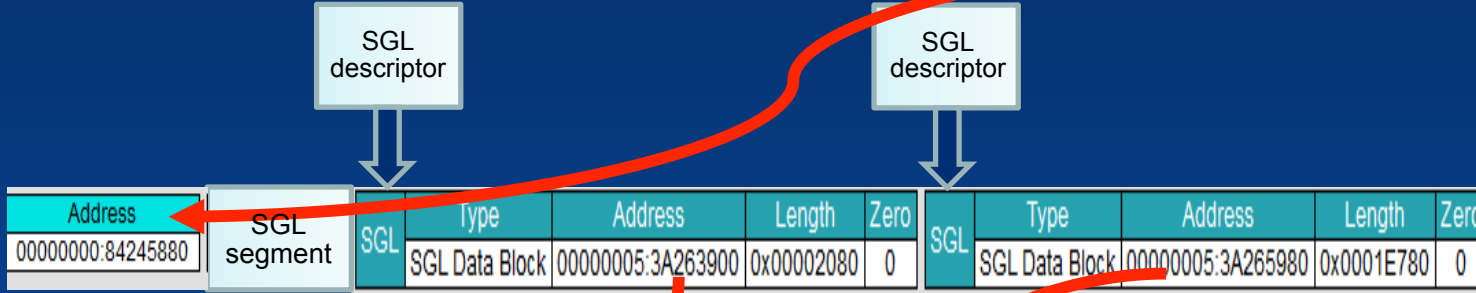
Flash Memory Summit

SGL decode transaction layer view

(1) command line, indicating the first SGL segment for the command and decoding its fields

IOSQ	OPC	FUSE	PSDT	CID	NSID	MPTR	Address	SGL	Type	Address	Length	Zero
	Read	Normal operation	SGL	0x0006	0x00000001		00000000:00000000	SGL	SGL Last Segment	00000000:84245880	0x00000020	0

(2) SGL segment line decoding its fields



(3) SGL data block per each SGL data block descriptor

Address	SGL Data	Data Len	Data	MN
00000005:3A263900		0x00000820	2080 dwords	SAMSUNG MZWLL1T6HEHP-00003
00000005:3A265980		0x000079E0	31200 dwords	SAMSUNG MZWLL1T6HEHP-00003



SGL decoding challenges

- Only the pointed to lines should show with the complete range
- Missing ranges should show as errors with tooltip pointing to missing ranges
- Duplicates should optionally show as errors with pointers in tooltips
- Account for bit bucket descriptors

NVMe Cmd	OPC	1 error(s)	SQID	CQID	CID	Data	MPTR	SGL 1	SLBA	NLB	PRINFO	FUA	LR	DSM	ACCF	ACCL	SEQR	INCOM		
0	D	Read	0x0001	0x0001	0x000C	33270 dwords	00000000.00000000	80 CE 2C 84 00 00 00 00 ...	00000000.0AC00200	0x00FF	0x4	0	0		No frequency information provided	None	0	0		
NVMe	Cmd	Device ID	QID	CID	Address	IO SQ	IO QDBL	MN	IO SQ	IO QDBL	MN	IO SQ	IO QDBL	MN	IO SQ	IO QDBL	MN	IO SQ	IO QDBL	MN
0	H	158:00:0	0x0001	0x000C	00000000.842CF300	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
1	H	158:00:0	0x0001	0x000C	00000000.842CF300	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
2	H	158:00:0	0x0001	0x000C	00000000.842CCE80	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
3	D	158:00:0	0x0001	0x000C	00000005.3A841800	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
4	D	158:00:0	0x0001	0x000C	00000005.3A849000	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
5	D	158:00:0	0x0001	0x000C	00000005.3AC08500	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
6	D	158:00:0	0x0001	0x000C	00000005.32F25200	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
7	D	158:00:0	0x0001	0x000C	00000005.32F31F28	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
8	D	158:00:0	0x0001	0x000C	00000005.32F36028	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
9	D	158:00:0	0x0001	0x000C	00000000.842C00C0	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1
10	H	158:00:0	0x0001	0x000C	00000000.842C00C0	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1	0x000D	0x000D	HUSMR7619BDP3Y1



Flash Memory Summit

PCIe Performance analysis

- PCIe is a split protocol
 - Allows new requests to cross old completions
 - Performance analysis to account for overlap
- PCIe FC credit
 - Accumulative credit accounting
 - Manages Bottlenecks



Flash Memory Summit

PCIe Performance Measurement Techniques

- performance criteria
 - Instantaneous performance metrics
 - Overall statistical analysis
 - Traffic summaries
 - Bus utilization charts
 - Conditional performance analysis using automated analysis scripting techniques



Performance metrics

- Response time – complete transfer time
 - First to last packet of split transaction
- Latency time – Time to data
 - End of request to start of completion
- Throughput – payload over response time
 - Total payload coincident with split transaction divided by response time

Split Tra	R←	2.5	Mem	MRd(64)	RequesterID	CompleterID	Tag	TC	VC ID	Address	Status	Data
17		x8		001:00000	001:00:0	000:00:0	69	0	0	00000008:C0000000	SC	512 dwords

Metrics	# LinkTrans	Resp. time	Latency	Pld. Bytes	Thrpt MB/s	Time Delta	Time Stamp
	5	3.484 us	1.260 us	2048	4226.342	32.000 ns	0000 . 050 010 250 s



Overall statistics, traffic summaries

PCIe timings			
Queue Utilization		Total Input / Output	
Bus Utilization			
	Upstream		Downstream
Link Utilization	0.156 %		0.246 %
Time Coverage	0.156 %		0.246 %
Bandwidth	2.984 MB/s		4.694 MB/s
Data Throughput	0.000 MB/s		3.754 MB/s
Packets/second	387337.474		111656.363
Split Transaction Performance			
	Minimum	Average	Maximum
Response Time	2.180 us	104.356 us	184.000 us
Latency	1.132 us	91.318 us	183.032 us
Throughput	0.026 MB/s	1793.359 MB/s	4741.034 MB/s

Resp. time (Min)	Resp. time (Avg)	Resp. time (Max)	Latency (Min)	Latency (Avg)	Latency (Max)
91.188 us	104.078 us	184.000 us	90.684 us	103.534 us	183.032 us
2.180 us	104.356 us	109.708 us	1.132 us	91.315 us	96.740 us

Requester -> Completer, Reads Δ	Total	Thrpt MB/s (Min)	Thrpt MB/s (Avg)	Thrpt MB/s (Max)
000:00:0 -> 001:00:0, Cfg TCO	18	0.026	199.078	1791.577
001:00:0 -> 000:00:0, Mem TCO	81601	0.437	1793.711	4741.034
	81619			



Flash Memory Summit

Use case: NVMoF over CNA

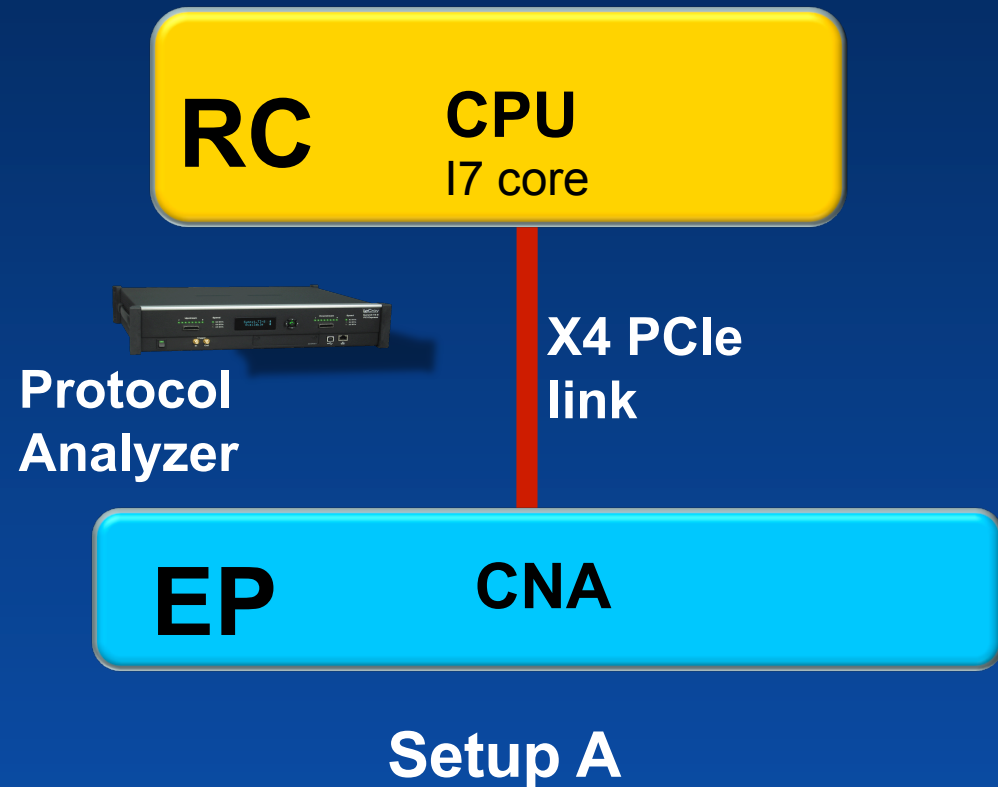
- PCIe performance analysis
 - Throuput, latency leading to credit analysis
- NVMof using CNA
 - Converged network adaptor
- CNA above switch used in NT mode
 - Non transparent bridging



Flash Memory Summit

- Setup A – Direct connection
 - No switch between root complex and endpoint.
- This setup shows 9.3 Gb/sec
 - No performance degradation
 - Baseline to establish the troubling component

Setup A – Host to Drive





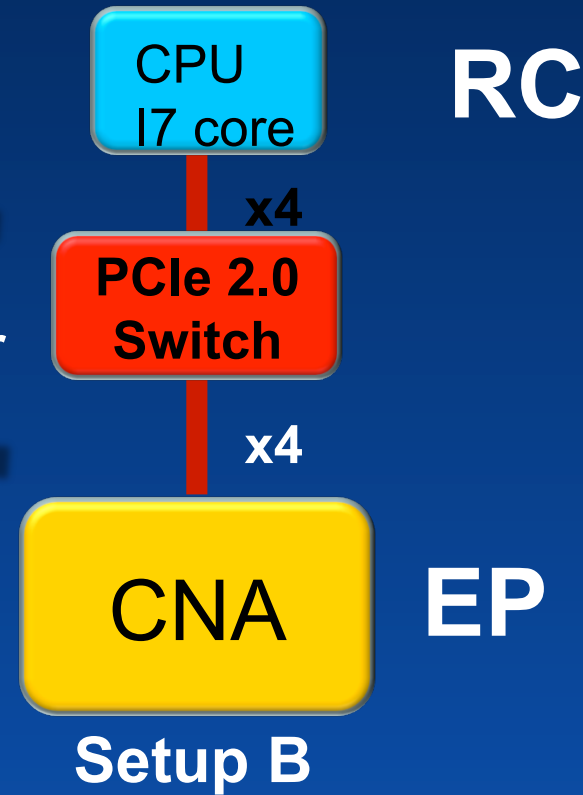
Setup B – connection through PCIe switch

- System uses a CNA Running NVMeF connected to a PCIe Switch.
- Switch connected to RC on i7 core.
- Need to determine root cause for a data throughput drop of 9.3Gig/sec to 5.5Gig/sec on an Optical 10GigE network

Cross Sync



Protocol Analyzer





Flash Memory Summit

Determine the root cause for performance degradation

- Identify performance degradation source
 - Host processor?
 - RC port?
 - Switch Primary port?
 - Switch secondary port?
 - EP port?
 - CNA?
- Once identified – can we tell what causes this source to limit performance?



The Analysis Process

- Is performance degradation reflected in PCIe link utilization?
- Is performance degradation observed on both primary and secondary links?
- Determine for each link if waiting for requests or stalling traffic
- If neither link is limiting performance the link is waiting for the requester, network or host
- If one of the links is limiting performance – determine which port
- What stalls port's performance



Flash Memory Summit

Setup A – instantaneous vs Overall performance

Teledyne LeCroy PETracer(TM) - PCI Express Protocol Analyzer - [E:\backup_for_Dec_2012\customers\lsi\ACP\AM2\active_port0_max_9.5Gig.pex]

File Setup Record Generate Report Search View Tools Window Help

Split Tra 3 R 5.0 Mem MRd(32) RequesterID 000:00000 CompleterID 000:00:0 Tag 3 TC 0 VC ID 0 Address 2F282000 Status SC Data 32 dwords Metrics # LinkTras 2 Resp. time 302.000 ns Latency 218.000 ns Thrpt MB/s 404.206 Pld. Bytes 128 Time Delta 12.000 ns

Link Tra 10 R 5.0 TLP Mem MWr(32) Length 32 RequesterID 010:00000 Tag 14 Address 2F35A700 1st BE 1111 Last BE 1111 Data 32 dwords VC ID 0 ExplicitACK Packet #37 Metrics # Packets 2 Resp. time 126.000 ns Pld. Bytes 128 Thrpt MB/s 968.812 Time Delta 80.000 ns

Timing Calculator - [active_port0_max_9.5Gig.pex]

From beginning of: Packet 0 To beginning of: Packet 217897
 Marker: Packet # 0 (start) Marker: Packet # 217897 (end)
 Time: 0.0000010200 secs Time: 0.0032283880 secs

Total Time: 3.227 milliseconds

Bus Utilization		
	Upstream	Downstream
Link Utilization	81.789 %	83.778 %
Time Coverage	81.688 %	83.668 %
Bandwidth	16357.90 Mb/s	16755.62 Mb/s
Data Throughput	1146.07 MB/s	1156.98 MB/s
Packets/second	24860505.53	42654881.62

Split Transaction Performance			
	Minimum	Average	Maximum
Response Time	176.000 ns	384.050 ns	784.000 ns
Latency	122.000 ns	227.470 ns	626.000 ns
Throughput (MB/s)	27.845	325.614	540.134

Memory Writes Performance			
	Minimum	Average	Maximum
Response Time	14.000 ns	98.630 ns	228.000 ns
Throughput (MB/s)	70.643	1314.133	1649.599

Status	Data	VC ID	ExplicitACK	# LinkTras	Resp. time	Latency	Thrpt MB/s	Pld. Bytes	Time Delta
SC	32 dwords	0	Packet #37	2	304.000 ns	220.000 ns	401.547	128	12.000 ns
ast BE 1111	32 dwords	0	Packet #44	2	124.000 ns	128	984.438	128	76.000 ns
SC	32 dwords	0	Packet #51	2	448.000 ns	218.000 ns	272.478	128	12.000 ns
ast BE 1111	32 dwords	0	Packet #59	2	126.000 ns	128	968.812	128	76.000 ns
SC	32 dwords	0	Packet #66	2	308.000 ns	224.000 ns	396.332	128	12.000 ns
ast BE 1111	32 dwords	0	Packet #73	2	126.000 ns	128	953.674	128	80.000 ns
SC	32 dwords	0	Packet #81	2	488.000 ns	280.000 ns	250.144	128	12.000 ns
ast BE 1111	32 dwords	0	Packet #88	2	128.000 ns	128	953.674	128	76.000 ns
SC	32 dwords	0	Packet #95	2	368.000 ns	282.000 ns	333.525	128	12.000 ns
ast BE 1111	32 dwords	0	Packet #102	2	186.000 ns	128	656.292	128	80.000 ns
SC	32 dwords	0	Packet #109	2	360.000 ns	276.000 ns	339.084	128	12.000 ns
ast BE 1111	32 dwords	0	Packet #116	2	184.000 ns	128	663.426	128	116.000 ns



Setup A viewed with Throughput chart and Packet Metrics

The screenshot displays a software interface with a menu bar (File, Setup, Record, Generate, Report, Search, View, Tools, Window, Help) and a toolbar. Below the toolbar are several data tables and charts. A red box highlights a 'Metrics' table with the following data:

	Resp. time	Latency	Thrpt MB/s	Pld. Bytes	Time
	356.000 ns	142.000 ns	171.447	64	12.00
Metrics	# Packets	Resp. time	Pld. Bytes	Thrpt MB/s	
	1	74.000 ns	128	1649.599	
Metrics	# Packets	Resp. time	Pld. Bytes	Thrpt MB/s	
	2	106.000 ns	128	1151.607	

Below the tables are two charts: 'SPLIT: Throughput Per Transaction' and 'Memory Writes: Throughput'. Both charts show throughput over time (176 to 226 μs). The 'SPLIT' chart shows a relatively stable throughput around 1.0 MB/s. The 'Memory Writes' chart shows a high, fluctuating throughput, reaching up to 1.0 MB/s.

- Consistent read and write throughput with full link utilization in both upstream and downstream directions.



Setup A viewed with Link Tracker

Flash Memory Summit

Split Tra	R	5.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC ID	Address	Status	Data	Metrics	# LinkTras	Resp. time	Latency	Thrpt MB/s	Pld. Bytes	Time Delta	Time Stamp
76	x4			00:00000	002:00:0	000:00:0	6	0	0	2F51E080	SC	32 dwords		2	258.000 ns	174.000 ns	473.141	128	16.000 ns	0000.000 007 748 s
Link Tra	R	5.0	TLP	Mem	MWr(32)	Length	RequesterID	Tag	Address	1st BE	Last BE	Data	VC ID	ImplicitACK	Metrics	# Packets	Time Delta	Time Stamp		
215	x4		3377		10:00000	32	002:00:0	8	2F2A9F80	1111	1111	32 dwords	0	Packet#502		1	76.000 ns	0000.000 007 764 s		

Link Tracker - Packet # 481

Time	Packet #	Upstream										Downstream									
00.000 007 616		FD	C1	8D	22	1F	F5	F3	A0												
00.000 007 618		FA	48	CC	F7	02	9A	E9	24												
00.000 007 620		CF	49	C0	F7	2A	3E	EF	B2												
00.000 007 622		F7	58	20	52	3E	CB	36	70												
00.000 007 624		DB	37	9E	A4	79	D2	FA	F9												
00.000 007 626		30	F4	33	D1	73	23	C3	EF												
00.000 007 628		01	76	1D	70	E5	2E	51	10												
00.000 007 630		D5	4F	24	4F	84	7A	A3	A3												
00.000 007 632		82	76	7F	3F	5E	C9	00	B6												
00.000 007 634		1B	59	E2	A6	C6	38	E2	26												
00.000 007 636		15	B2	30	8E	62	66	FC	AB												
00.000 007 638		BC	16	C0	75	5B	75	9F	28												
00.000 007 640		8B	88	98	58	D5	73	DB	05												
00.000 007 642		0F	D8	59	65	BD	DB	DF	C8												
00.000 007 644		FE	83	D8	95	9E	93	6E	B7												
00.000 007 646		6B	6E	85	04	88	05	EE	04												
00.000 007 648		60	57	09	A6	74	59	15	9C												
00.000 007 650		7C	6A	DB	EC	7A	C1	91	EC												
00.000 007 652		33	1C	93	DA	34	8B	18	F2												
00.000 007 654		DC	10	EC	AA	A2	7F	0C	34												
00.000 007 656		9A	A7	4F	FD	67	55	4D	FD												
00.000 007 658																					
00.000 007 660	481 (Upstream)	FB	17	34	1A																
00.000 007 662		90	80	92	B7																
00.000 007 664		98	9C	67	B7																
00.000 007 666	183 (Downstream)	9E	2F	CF	51	5C	2D	44	20												
00.000 007 668		A8	47	9B	FD	11	87	44	FD												
00.000 007 670	184 (Downstream)					FB	B4	4C	F9												
00.000 007 672		FB	D7	F5	9A	F0	F0	D0	F0												
00.000 007 674		A1	A1	81	A3	2B	2B	AB	29												
00.000 007 676		F2	F5	0D	D0	6E	6E	6E	AE												
00.000 007 678		60	D5	4A	E4	5B	4C	F4	BD												
00.000 007 680		5C	86	8A	7D	FF	9D	7E	A4												

- Link Tracker display shows Upstream and Downstream data transfer with full link utilization across all lanes in both directions.



Flash Memory Summit

Setup A vs Setup B Timing Calculator Comparisons

From beginning of: Packet 2 To beginning of: Packet 17452

Marker Marker

Time 0.0000010280 secs Time 0.0002592020 secs

Total Time: 258.174 microseconds

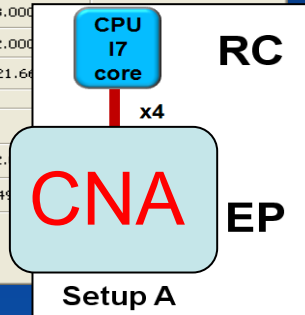
	Upstream	Downstream
Link Utilization	81.732 %	83.928 %
Time Coverage	81.625 %	83.820 %
Bandwidth	16346.34 Mb/s	16785.58 Mb/s
Data Throughput	1144.64 MB/s	1159.60 MB/s
Packets/second	24944417.33	42645657.58

	Minimum	Average	Maximum
Response Time	190.000 ns	383.960 ns	728.000 ns
Latency	138.000 ns	226.610 ns	502.000 ns
Throughput (MB/s)	38.532	325.765	521.644

	Minimum	Average	Maximum
Response Time	14.000 ns	98.140 ns	202.000 ns
Throughput (MB/s)	84.771	1321.288	1644.000

Note: 1 Mb= 1000 * 1000 bits and 1 MB = 1024 * 1024 bytes.

Calculate



Setup A

From beginning of: Packet 2424 To beginning of: Packet 30167

Marker Packet # 2424 (Marker #1) Marker Packet # 30167 (Marker #2)

Time 0.0000502540 secs Time 0.0006338780 secs

Total Time: 583.624 microseconds

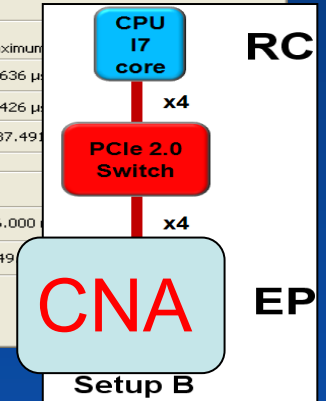
	Upstream	Downstream
Link Utilization	48.709 %	47.337 %
Time Coverage	48.650 %	47.282 %
Bandwidth	9741.89 Mb/s	9467.40 Mb/s
Data Throughput	643.02 MB/s	635.09 MB/s
Packets/second	20927857.66	26607884.53

	Minimum	Average	Maximum
Response Time	502.000 ns	1.129 μs	5.636 μs
Latency	430.000 ns	940.310 ns	5.426 μs
Throughput (MB/s)	5.537	154.453	237.491

	Minimum	Average	Maximum
Response Time	42.000 ns	157.180 ns	276.000 ns
Throughput (MB/s)	45.960	792.824	1649.000

Note: 1 Mb= 1000 * 1000 bits and 1 MB = 1024 * 1024 bytes.

Calculate



Setup B



Setup B: Cross sync between primary and secondary links

Split Tra	R	5.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	Address	Data	Metrics	# LinkTras	Time Delta	Time Stamp
4	R	x4	Mem	00:00000	000:00:0	001:00:0	5	A1462600	32 dwords	Metrics	2	76.000 ns	0000 . 000 001 550 s
5	R	x4	Mem	00:00000	000:00:0	001:00:0	2	A1462680	32 dwords	Metrics	2	164.000 ns	0000 . 000 001 626 s
6	R	x4	Mem	00:00000	000:00:0	001:00:0	6	A1462700	32 dwords	Metrics	2	88.000 ns	0000 . 000 001 790 s
7	R	x4	Mem	00:00000	000:00:0	001:00:0	0	A1462780	32 dwords	Metrics	2	136.000 ns	0000 . 000 001 878 s
8	R	x4	Mem	00:00000	000:00:0	001:00:0	1	A1462800	32 dwords	Metrics	2	84.000 ns	0000 . 000 002 014 s
9	R	x4	Mem	00:00000	000:00:0	001:00:0	4	A1462880	32 dwords	Metrics	2	240.000 ns	0000 . 000 002 098 s
10	R	x4	Mem	00:00000	000:00:0	001:00:0	5	A1462900	32 dwords	Metrics	2	112.000 ns	0000 . 000 002 338 s

C:\isaac\customers\lsi\ACP\cross_sync_tests\speed_6gig_pex.pex

Split Tra	R	5.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	Address	Data	Metrics	# LinkTras	Time Delta	Time Stamp
6	R	x4	Mem	00:00000	003:00:0	000:00:0	5	2F462600	32 dwords	Metrics	2	80.000 ns	0000 . 000 001 750 s
7	R	x4	Mem	00:00000	003:00:0	000:00:0	2	2F462680	32 dwords	Metrics	2	184.000 ns	0000 . 000 001 830 s
8	R	x4	Mem	00:00000	003:00:0	000:00:0	6	2F462700	32 dwords	Metrics	2	80.000 ns	0000 . 000 002 014 s
9	R	x4	Mem	00:00000	003:00:0	000:00:0	0	2F462780	32 dwords	Metrics	2	80.000 ns	0000 . 000 002 094 s
10	R	x4	Mem	00:00000	003:00:0	000:00:0	1	2F462800	32 dwords	Metrics	2	96.000 ns	0000 . 000 002 174 s
11	R	x4	Mem	00:00000	003:00:0	000:00:0	4	2F462880	32 dwords	Metrics	2	240.000 ns	0000 . 000 002 270 s
12	R	x4	Mem	00:00000	003:00:0	000:00:0	5	2F462900	32 dwords	Metrics	2	112.000 ns	0000 . 000 002 510 s

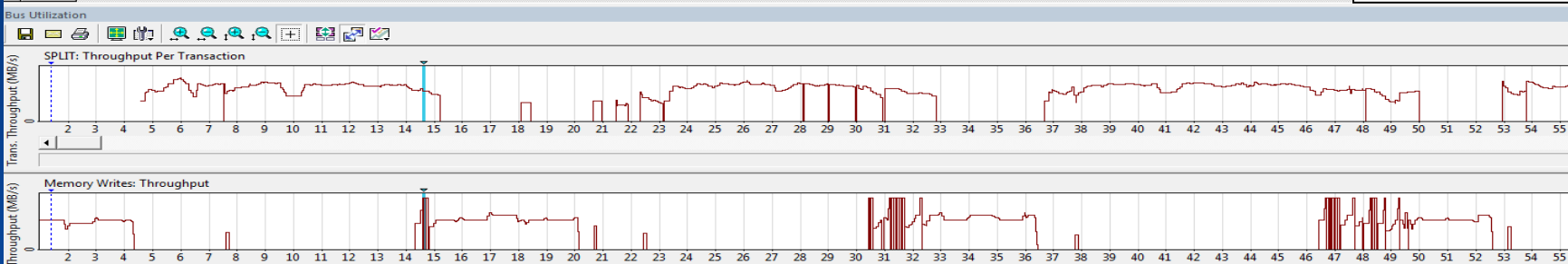
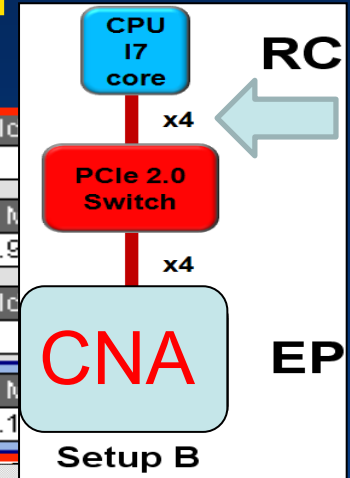


Flash Memory Summit

Setup B: Capture between RC and switch

Link Tra	R	5.0	TLP	Mem	MW(32)	Length	RequesterID	Tag	Address	
215	x4	872			10.00000	32	003:00:0	18	2F2F5B00	
Split Tra	86	x4	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC ID	A
				00:00000	003:00:0	000:00:0	0	0	0	2F
Link Tra	218	x4	874		10.00000	32	003:00:0	19	2F2F5B80	
Split Tra	87	x4	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC ID	A
				00:00000	003:00:0	000:00:0	2	0	0	2F
Link Tra	221	x4	876		10.00000	32	003:00:0	20	2F2F5C00	
Link Tra	223	x4	877		10.00000	32	003:00:0	21	2F2F5C80	
Link Tra	225	x4	878		10.00000	32	003:00:0	22	2F2F5D00	
Link Tra		x4	TLP	Mem	MW(32)	Length	RequesterID	Tag	Address	

Implicit ACK	Packet #	Metrics	# Packets	Resp. time	Plc
Packet #662			1	74.000 ns	
Metrics	# LinkTras	Resp. time	Latency	Thrpt M	
	2	444.000 ns	196.000 ns	274.9	
Implicit ACK	Packet #	Metrics	# Packets	Resp. time	Plc
Packet #670			1	74.000 ns	
Metrics	# LinkTras	Resp. time	Latency	Thrpt M	
	2	422.000 ns	176.000 ns	253.1	



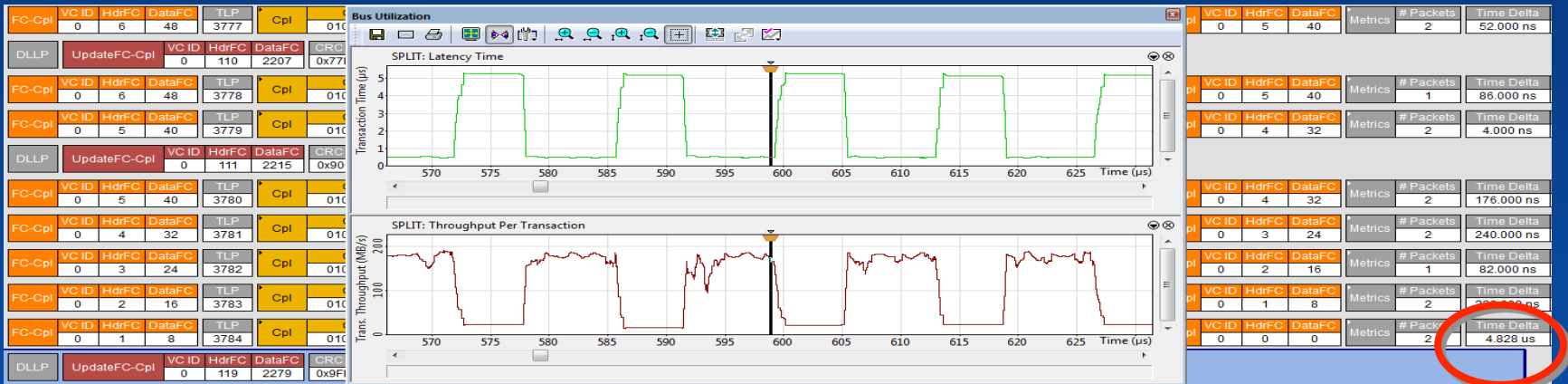
Read and write transfers are not overlapped. High Read throughput coincident with 0 write throughput and vice versa. Split view shows however low latencies for read completions



Flash Memory Summit

Setup B: Capture between switch and CNA

- What causes the long completion latencies?
- CNA in root mode
- Finite initial completion credits
- NT mode implemented in switch
- Completion credit is exhausted





Why need a tool to debug the serial links within the fabric?

- Need to see that the link width and speed come up correctly. This directly affects throughput.
- How to communicate to switch vendor the nature of the issue? An analyzer trace can prove the problem is with the switch. This is the correct way to show the root cause and communicate it to the vendor.
- The vendor may have never seen the issue since they do not use large networked storage fabrics.
- systems assembly/test engineers and system integrators need to be able to detect an interoperability problem and show evidence of the root cause of the problem and report to the silicon manufacturer



Flash Memory Summit

NVMe performance analysis

- As NVMe technology matures leading to a need to maximize performance
- What's special about NVMe performance vs general PCIe performance
- Differences in performance analysis techniques between SSD drives and traditional magnetic drives.



Flash Memory Summit

NVMe has different latency sources

- Doorbell to command submission
- Command submission to Data transfer
- Command submission to command completion
- Command completion to interrupt



NVMe performance criteria

- Response time
 - Transmission of the complete transfer from the beginning of the PCIe packet to the end of the last PCIe packet of this NVMe command
- Latency time
 - Time from the last PCIe packet of the NVMe command submission to the first PCIe packet of the NVMe command completion
- IOPS
 - # of overlapping NVMe commands from submission doorbell to completion doorbell



instantaneous performance metrics

Flash Memory Summit

NVMe Cmd	D	OPC	SQID	CQID	CID	Data	MPTR	PRP1	PRP2	SLBA	NLB	PRINFO	FUA	LR	DSM	ACCF
21		Read	0x0001	0x0001	0x0003	128 dwords	00000000:00000000	00000002:2DD9B000	00000000:00000000	00000000:00000000	0x0000	0x0	0	0	DSM	No frequency information provided
ACCL	SEQR	INCOM	EILBRT	ELBAT	ELBATM	ST	SCT	SC	Device ID	MN	Explicit SQyTDBL	Explicit IOSQ	Explicit IOCQ	Explicit CQyHDBL	NSID	Metrics
None	0	0	0x00000000	0x0000	0x0000	ST	Generic Command Status	Successful Completion	006:00:0	NVMeLeCroy000000	NVMe #120	NVMe #121	NVMe #123	NVMe #125	0x00000001	
# NVMe Trans	Resp. time	Latency	Pld. Bytes	Thrpt MB/s	IOPS	SDbl - CDbI	SDbl - CCmd	SCmd - CCmd	Time Delta	Time Stamp						
6	683.652 us	464.192 us	512	0.714	1462.733	682.240 us	631.476 us	466.192 us	692.512 us	0039.509104202 s						
22		Read	0x0001	0x0001	0x0004	128 dwords	00000000:00000000	00000002:2DFC3CC0	00000000:00000000	00000000:00000000	0x0000	0x0	0	0	DSM	No frequency information provided
ACCL	SEQR	INCOM	EILBRT	ELBAT	ELBATM	ST	SCT	SC	Device ID	MN	Explicit SQyTDBL	Explicit IOSQ	Explicit IOCQ	Explicit CQyHDBL	NSID	Metrics
None	0	0	0x00000000	0x0000	0x0000	ST	Generic Command Status	Successful Completion	006:00:0	NVMeLeCroy000000	NVMe #126	NVMe #127	NVMe #129	NVMe #131	0x00000001	
# NVMe Trans	Resp. time	Latency	Pld. Bytes	Thrpt MB/s	IOPS	SDbl - CDbI	SDbl - CCmd	SCmd - CCmd	Time Delta	Time Stamp						
6	812.300 us	509.776 us	512	0.601	1231.072	810.888 us	760.884 us	511.840 us	13.514 ms	0039.509796714 s						
23		Read	0x0001	0x0001	0x0005	128 dwords	00000000:00000000	00000002:2DFCD980	00000000:00000000	00000000:00000000	0x0000	0x0	0	0	DSM	No frequency information provided
ACCL	SEQR	INCOM	EILBRT	ELBAT	ELBATM	ST	SCT	SC	Device ID	MN	Explicit SQyTDBL	Explicit IOSQ	Explicit IOCQ	Explicit CQyHDBL	NSID	Metrics
None	0	0	0x00000000	0x0000	0x0000	ST	Generic Command Status	Successful Completion	006:00:0	NVMeLeCroy000000	NVMe #132	NVMe #133	NVMe #135	NVMe #137	0x00000001	
# NVMe Trans	Resp. time	Latency	Pld. Bytes	Thrpt MB/s	IOPS	SDbl - CDbI	SDbl - CCmd	SCmd - CCmd	Time Delta	Time Stamp						
6	681.364 us	470.128 us	512	0.717	1467.644	679.984 us	626.980 us	472.144 us	696.584 us	0039.523310538 s						
24		Read	0x0001	0x0001	0x0006	128 dwords	00000000:00000000	00000002:2DD9D000	00000000:00000000	00000000:00000000	0x0000	0x0	0	0	DSM	No frequency information provided
ACCL	SEQR	INCOM	EILBRT	ELBAT	ELBATM	ST	SCT	SC	Device ID	MN	Explicit SQyTDBL	Explicit IOSQ	Explicit IOCQ	Explicit CQyHDBL	NSID	Metrics
None	0	0	0x00000000	0x0000	0x0000	ST	Generic Command Status	Successful Completion	006:00:0	NVMeLeCroy000000	NVMe #138	NVMe #139	NVMe #141	NVMe #143	0x00000001	
# NVMe Trans	Resp. time	Latency	Pld. Bytes	Thrpt MB/s	IOPS	SDbl - CDbI	SDbl - CCmd	SCmd - CCmd	Time Delta	Time Stamp						
6	780.748 us	475.760 us	512	0.625	1280.823	779.112 us	726.268 us	477.808 us	804.880 us	0039.524007122 s						



Response time, Latency time

NVMe Cmd	D	OPC	SQID	CQID	CID	Data	MPTR	PRP1	PRP2	SLBA	NLB	PRINFO	FUA	LR	DSM	ACCF	ACCL	SEQR	INCOM	EILBRT	ELBAT	ELBATM	ST	
21		Read	0x0001	0x0001	0x0003	128 dwords	00000000:00000000	00000002:2DD9B000	00000000:00000000	00000000:00000000	0x0000	0x0	0	0		No frequency information provided	None	0	0	0x00000000	0x0000	0x0000		
Generic Command Status		Successful Completion	Device ID	MN	Explicit SQyDBL	Explicit IOSQ	Explicit IOCC	Explicit CQyHDBL	NSID	Metrics	# NVMe Tr	Resp. time	Latency	Cmd. Bytes	Thrpt M	IOPS	SDbl - CDBl	DBl - CCmd	SCmd - CCmd					
Time Delta		Time Stamp	006:00:0	NVMeLeCroy000000	NVMe #120	NVMe #121	NVMe #123	NVMe #125	0x00000001	6		683.652 us	464.192 us	512	0.71	1462.733	682.240 us	31.476 us	466.192 us					
Time Delta		Time Stamp	1.428 us	0039_509_104_202 s																				
NVMe	H	Device ID	QID	CID	SQyDBL	IO SQT	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp													
120		006:00:0	0x0001	0x0004	0x0004	0x0004	NVMeLeCroy000000	1	1	1.428 us	0039_509_104_202 s													
Link Tra	R	2.5	TLP	Mem	MW(r)32	Length	RequesterID	Tag	Address	1st BE	Last BE	Data	VC ID	Explicit ACK	Metrics	# Packets	Time D	Time Stamp						
1927		x1	929	010:00000	1	000:00:0	0	FE201008	1111	0000	040000000	0	Packet #377117	2	1.428	0039_509_104_202 s								
Packet	R	2.5	TLP	Mem	MW(r)32	Length	RequesterID	Tag	Address	1st BE	Last BE	Data	LCRC	Time Delta	Time Stamp									
377116		x1	929	010:00000	1	000:00:0	0	FE201008	1111	0000	1 dwozd	0x551D6F4D	1.428 us	0039_509_104_202 s										
Packet	R	2.5	DLLP	ACK	AckNak_Seq_Num	CRC 16	Time Delta	Time Stamp																
377117		x1	929	0xD93D	163.856 us	0039_509_105_630 s																		
NVMe	H	Device ID	QID	CID	Address	IOSQ	OPC	FUSE	PSDT	CID	NSID	MPTR	Address	PRP1	Address	PRP2	Address	SLBA	NLB	PRINFO	PRCHK	PRACT	FUA	LR
121		006:00:0	0x0001	0x0003	00000002:2E0830C0	Read	Normal operation	PRP	0x0003	0x00000001	00000000:00000000	00000002:2DD9B000	00000000:00000000	00000000:00000000	0x0000	000	0	0	0	0	0	0	0	
AF		AL	SR	I	EILBRT	ELBAT	ELBATM	MN	Metrics	# Link & Split Trans	Time De	Time Stamp												
No frequency info		None	0	0	0x00000000	0x0000	0x0000	NVMeLeCroy000000	1	358.208	0039_509_269_486 s													
NVMe	D	Device ID	QID	CID	Address	PRP Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp											
122		006:00:0	0x0001	0x0003	00000002:2DD9B000	128 dwords	0x00000080	128 dwords	NVMeLeCroy000000	4	1	107.984 us	0039_509_627_694 s											
NVMe	D	Device ID	QID	CID	Address	IOCC	SQHD	SQID	CID	P	DW0	Reserved	ST	SCT	SC	MNR	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp			
123		006:00:0	0x0001	0x0003	00000002:2E093030	0x0004	0x0001	0x0003	1	0	0x00000000	0	0	Generic Command Status	Successful Completion	0	NVMeLeCroy000000	1	21.648 us	0039_509_735_678 s				
NVMe	D	Device ID	QID	CID	Address	Interrupt	Type	Vector	Message	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp										
124		006:00:0	0x0001	0x0003	00000000:FEE3F00C	MSI	1	0x000049A9	NVMeLeCroy000000	1	1	29.116 us	0039_509_732_635 s											
NVMe	H	Device ID	QID	CID	Address	IO COH	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp													
125		006:00:0	0x0001	0x0004	0x0004	NVMeLeCroy000000	1	1.380 us	0039_509_786_442 s															
Link Tra	R	2.5	TLP	Mem	MW(r)32	Length	RequesterID	Tag	Address	1st BE	Last BE	Data	VC ID	Explicit ACK	Metrics	# Packets	Time D	Time Stamp						
1936		x1	931	010:00000	1	000:00:0	0	FE20100C	1111	0000	040000000	0	Packet #377135	2	1.380	0039_509_786_442 s								



SAS vs NVMe IOPS definition

- SAS Definition
 - $IOPS = 1 / \text{Latency}$
- NVMe definition
 - $IOPS = \# \text{ commands} / \text{Sdbl-CDbl}$
 - Example $1 / 631240 \text{ usec} = 1462.733$



Categorized Performance analysis

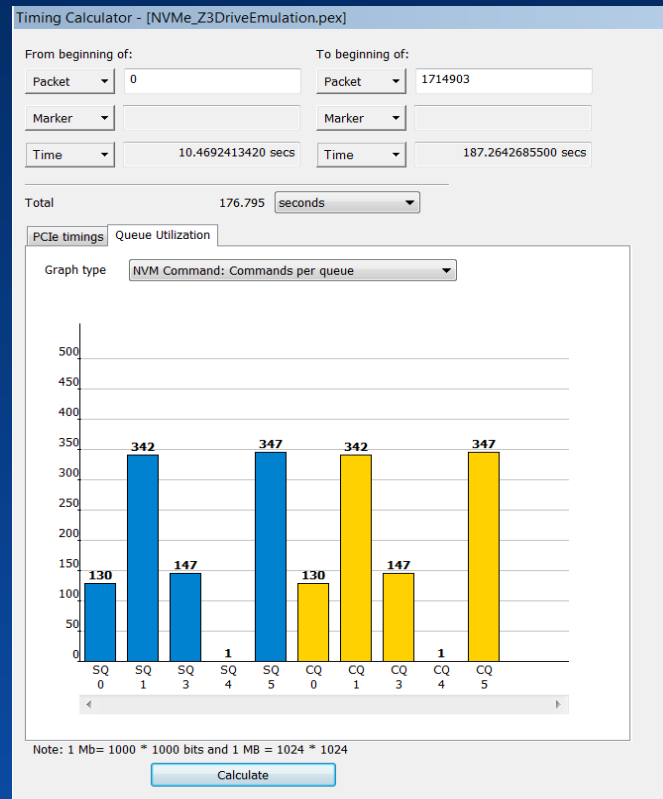
Read Requests Performance

Requester -> Completer	Total	Thrpt MB/s (Min)	Thrpt MB/s (Avg)	Thrpt MB/s (Max)	Resp. time (Min)	Resp. time (Avg)	Resp. time (Max)	Latency (Min)	Latency (Avg)	Latency (Max)
000:03:0 -> 000:00:0, Cfg TCO	416	0.995	21.175	1502.353	378.000 ns	1.172 us	3.834 us	204.000 ns	1.009 us	3.662 us
000:03:0 -> 002:00:0, Cfg TCO	5314	7.914	6189.918	11387.661	338.000 ns	376.760 ns	482.000 ns	174.000 ns	210.050 ns	304.000 ns
000:03:0 -> 003:04:0, Cfg TCO	5198	7.981	6833.574	11387.661	338.000 ns	382.470 ns	478.000 ns	174.000 ns	216.120 ns	304.000 ns
000:03:0 -> 004:00:0, Cfg TCO	1044	0.245	22.834	1526.181	2.522 us	3.152 us	15.562 us	2.360 us	2.986 us	15.396 us
000:03:0 -> 003:05:0, Cfg TCO	5194	8.048	6841.318	11387.661	338.000 ns	382.450 ns	474.000 ns	174.000 ns	215.810 ns	312.000 ns
000:03:0 -> 005:00:0, Cfg TCO	1024	0.093	23.177	1507.059	2.554 us	3.202 us	41.170 us	2.390 us	3.035 us	40.994 us
000:03:0 -> 003:06:0, Cfg TCO	5190	8.048	6844.803	11387.661	338.000 ns	382.440 ns	474.000 ns	174.000 ns	216.210 ns	312.000 ns
000:03:0 -> 006:00:0, Cfg TCO	1024	0.082	23.386	1523.765	2.526 us	3.198 us	46.490 us	2.370 us	3.032 us	46.322 us
000:03:0 -> 003:07:0, Cfg TCO	5190	7.981	6845.320	11387.661	338.000 ns	382.410 ns	478.000 ns	174.000 ns	216.160 ns	304.000 ns
000:03:0 -> 007:00:0, Cfg TCO	1024	0.081	23.116	1521.356	2.530 us	3.212 us	46.918 us	2.386 us	3.045 us	46.774 us
000:03:0 -> 004:00:0, Mem TCO	2089	0.005	0.889	16.124	3.954 us	6.335 us	807.882 us	3.796 us	6.168 us	807.708 us
004:00:0 -> 000:00:0, Mem TCO	277	155.702	186.383	468.472	274.000 ns	332.680 ns	456.000 ns	178.000 ns	230.490 ns	300.000 ns
000:03:0 -> 005:00:0, Mem TCO	2083	0.005	0.892	15.949	3.962 us	6.252 us	804.430 us	3.790 us	6.084 us	804.254 us
005:00:0 -> 000:00:0, Mem TCO	190	155.249	184.474	214.913	284.000 ns	329.720 ns	378.000 ns	168.000 ns	227.480 ns	274.000 ns
000:03:0 -> 006:00:0, Mem TCO	2078	0.005	0.873	16.188	3.962 us	6.371 us	805.082 us	3.786 us	6.203 us	804.902 us
006:00:0 -> 000:00:0, Mem TCO	541	61.527	1042.631	2765.319	288.000 ns	425.760 ns	526.000 ns	174.000 ns	248.020 ns	354.000 ns
000:03:0 -> 007:00:0, Mem TCO	2071	0.005	0.859	1.904	3.958 us	6.363 us	813.442 us	3.794 us	6.195 us	813.274 us
007:00:0 -> 000:00:0, Mem TCO	535	58.388	1051.490	2770.340	288.000 ns	428.030 ns	516.000 ns	178.000 ns	248.270 ns	338.000 ns
004:00:0 -> 000:31:6, Mem TCO	429	78.250	148.991	200.774	304.000 ns	416.230 ns	780.000 ns	192.000 ns	312.830 ns	676.000 ns
005:00:0 -> 000:31:6, Mem TCO	393	109.382	153.033	204.816	298.000 ns	402.730 ns	558.000 ns	194.000 ns	300.040 ns	454.000 ns
006:00:0 -> 000:31:6, Mem TCO	637	62.196	271.715	2601.922	302.000 ns	399.540 ns	614.000 ns	190.000 ns	283.240 ns	510.000 ns
007:00:0 -> 000:31:6, Mem TCO	872	44.703	285.667	2726.789	302.000 ns	395.230 ns	578.000 ns	198.000 ns	277.180 ns	474.000 ns
	42813									



Timing Calculator Queue View

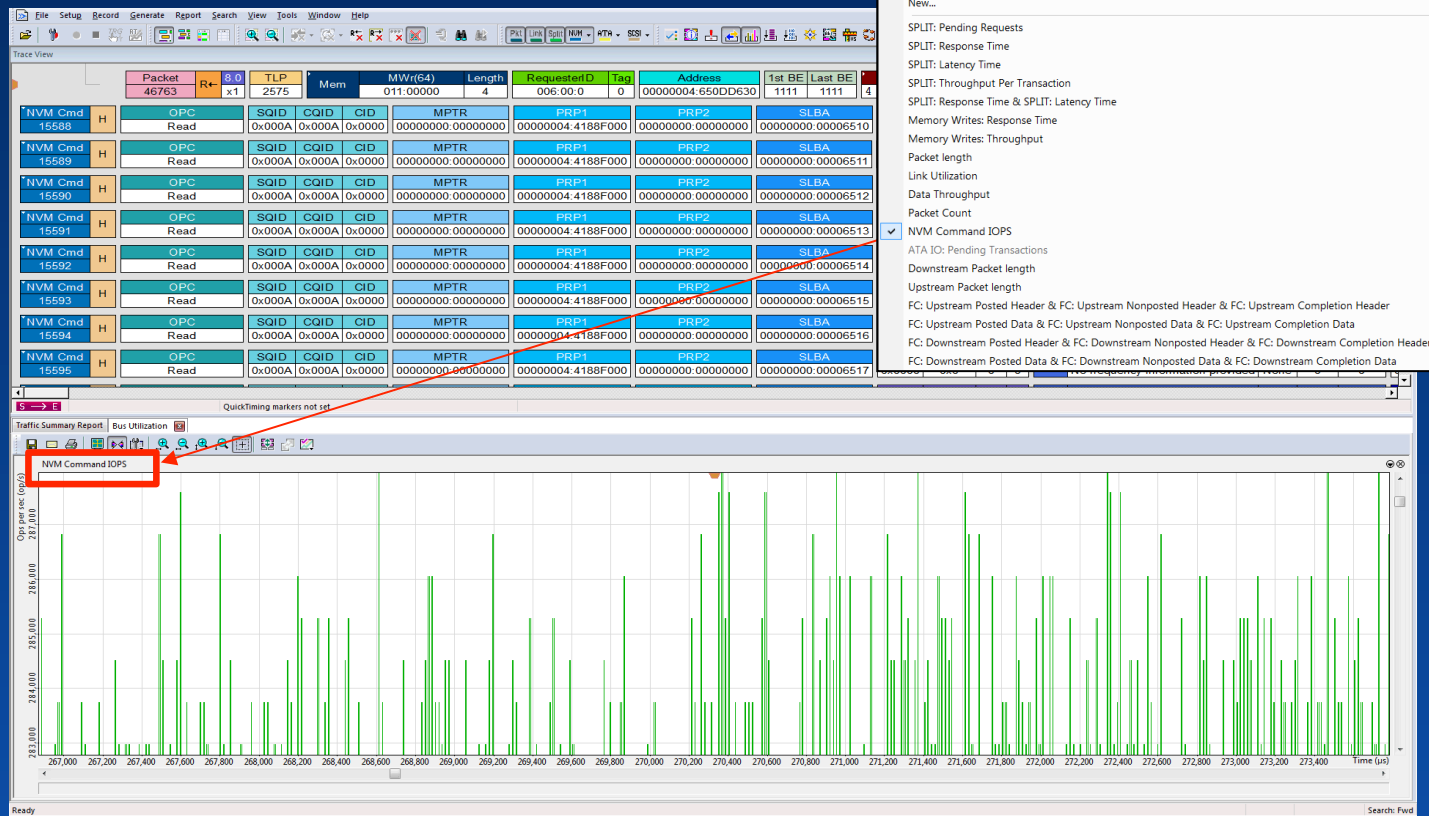
- View submission and completion Queues
- Compare Queues to see where overloading is occurring
- Verify if submission and completion queues are equal and that nothing was lost





Flash Memory Summit

NVMe Command IOPS Statistical chart





Flash Memory Summit

Long Recordings

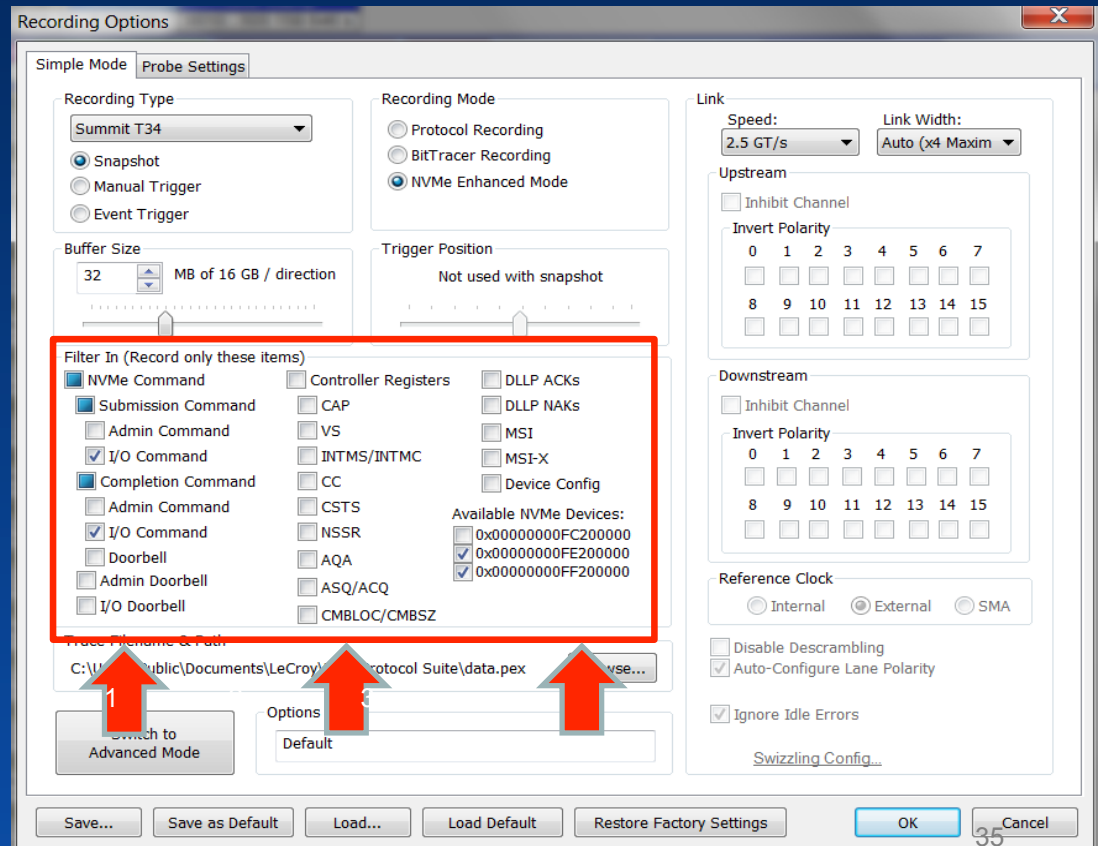
Memory Utilization Model Assumptions

- 16GB memory dedicated per direction
- Capture Duration Doubles when using expanded mode
- Recording stops as soon as either side fills up
- SSD rate for read is 2 GB / sec or 16 Gb/sec
 - This implies a Gen3 x4 link with 60% utilization
- Assume 16 pages / command
- Assume 2 doorbells
- Assume no interrupt aggregation
 - Dropped idles, SKPs, EDSs, DLLPs
- Each TLP occupies between 1.09-1.43 memory blocks on average



NVMe Enhanced mode – long recordings

- 3 Columns of filter in items
 - Entities that form NVMe Commands
 - NVMe Control Registers
 - PCIe entities related to NVMe traffic





Flash Memory Summit

Conditional performance analysis using Verification scripting

- Extract metrics within a defined range
- LBA drive access pattern
- Queue access distribution
- Low power states entry / exist
- Multiple NVMe commands per TLP referenced by time stamps



VSE script example

```
OnStartScript()
{
    ReportText("OnStartScript called...");
    ReportText("\n\nRunning...\n");
    EventCount = 0;
    SendAllChannels();
    SendAllTraceEvents();
    SendLevelOnly( _NVMC );
    #SendLevelOnly( _SPLIT );
    filePtr = OpenFile("C:\\Documents\\test.csv");
    WriteString(filePtr,"Start time, Response time,
        latencyTime, LBA, Length, QID, FUA"); }
```



VSE script example

```
ProcessEvent()  
{  
    respTime= in.Metric_ResponseTime;  
    latencyTime=in.Metric_LatencyTime;  
    time=in.Time;  
    throughput=in.Metric_Throughput;  
    CMD = in.nvmcCommandOpCode;  
    NLB = in.Read_NLB;  
    SLBA0 = in.Read_SLBA_DW0;  
    SLBA1 = in.Read_SLBA_DW1;  
    SLBA = in.Read_SLBA;  
    SQID = in.nvmcSubmissionQueueID;  
    if( CMD!= 1 )    FUA = 0;  
    else    FUA = in.Write_FUA;  
    if (SQID!= 0) {ReportText(FormatEx("%s,%s,%d,%s,%d,%d,%d", CSV_Val_TimeStamp_Seconds(in.Time ),  
    CSV_Val_TimeStamp_Seconds(respTime), CMD, (SLBA), (NLB+1), SQID, FUA));}  
    if (SQID!= 0) {WriteString(filePtr,FormatEx("%s,%s,%s,%s,%d,%d,%d",  
    CSV_Val_TimeStamp_Seconds(in.Time ), CSV_Val_TimeStamp_Seconds(respTime),  
    CSV_Val_TimeStamp_Seconds(latencyTime), (SLBA), (NLB+1), SQID, FUA));}  
    if( EventCount == MAX_NUMBER_OF_EVENTS ) ScriptPassed();  
    EventCount++;  
    return Complete();}
```

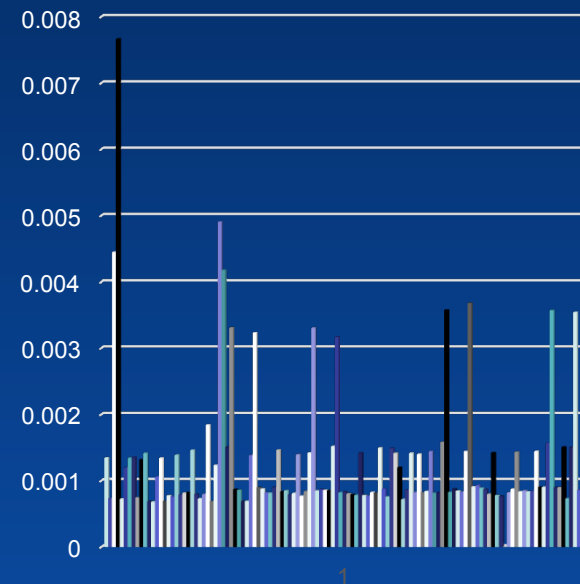


Flash Memory Summit

SSD traffic statistics

Start time	Response time	Command	LBA	Length	QID	FUA
97.44338	0.000819348	20x8000000000000000	20x8000000000000000	1	3	0
97.44427	0.00087518	20x8000000000000000	20x8000000000000000	1	3	0
97.44522	0.001432268	20xC000000000000000	20xC000000000000000	4	3	0
97.4476	0.0008339	20x8000000000000000	20x8000000000000000	1	5	0
97.44904	0.00085046	20x8000000000000000	20x8000000000000000	1	1	0
97.45029	0.000832564	20x8000000000000000	20x8000000000000000	1	3	0
97.45125	0.000837508	20x8000000000000000	20x8000000000000000	1	3	0
97.4522	0.00144462	20xC000000000000000	20xC000000000000000	4	3	0
97.45464	0.000901324	20x8000000000000000	20x8000000000000000	1	5	0
97.4576	0.000903012	20x8000000000000000	20x8000000000000000	1	3	0
97.4586	0.001566428	20xC000000000000000	20xC000000000000000	4	3	0
97.46027	0.003569364	20x8000000000000000	20x8000000000000000	13	3	0
97.46389	0.00089634	20x7FE8070000000000	20x7FE8070000000000	1	3	0
97.46483	0.00089858	20x80F4030000000000	20x80F4030000000000	1	3	0
97.46662	0.001507236	20x8000000000000000	20x8000000000000000	1	5	0
97.47032	0.00072846	20x8000000000000000	20x8000000000000000	1	1	0
97.47108	0.001501092	20xC000000000000000	20xC000000000000000	4	3	0
97.47268	0.003540652	20x8000000000000000	20x8000000000000000	13	3	0
97.47632	0.000855788	20x7FE8070000000000	20x7FE8070000000000	1	3	0
97.47727	0.0008609	20x80F4030000000000	20x80F4030000000000	1	3	0
97.47907	0.000801716	20x8000000000000000	20x8000000000000000	1	3	0
97.48059	0.000653244	20x8000000000000000	20x8000000000000000	1	5	0
97.48137	0.000685132	20x8000000000000000	20x8000000000000000	1	5	0
97.48209	0.001389612	20xC000000000000000	20xC000000000000000	4	5	0
97.48421	0.00069338	20x8000000000000000	20x8000000000000000	1	5	0
97.48543	0.000734724	20x8000000000000000	20x8000000000000000	1	5	0

Response time





Flash Memory Summit

Traffic generation

- The use of traffic generators to stress test an NVMe device and characterize its performance independent of a specific platform.



Generation script for stress testing

- High doorbell entry
- Fast completions

Address Space	Location	Offset	Size	Data
Write	Mem64	0x00010000	0x00000040	64 bytes

Packet	2.5	TLP	Mem	MWr(32)	Length	RequesterID	Tag	Address	1st BE	Last BE	Data	LCRC
1	x1	0	Mem	MWr(32)	1	064:02:0	0	EC031008	1111	0000	55000000	0xA1C03DFF

Wait TLP Header	Timeout	Fmt	Type	Address [31:0]
	Infinite	3DW header, with data	MRd/MWr	FFFFFFFF:FEE20000

Packet	2.5	TLP	Mem	MWr(32)	Length	RequesterID	Tag	Address	1st BE	Last BE	Data	LCRC
3	x1	0	Mem	MWr(32)	1	064:02:0	1	EC03100C	1111	0000	01000000	0xAA9510A1

S → E QuickTiming markers not set

Generation Script Editor

```
15 Structure=NVMem
16 {
17     Location = Mem64
18     Offset = 0x10000
19     NVMeStructType=NVMCommand
20     OpcodeNvm = Read
21     CID = 0
22     NamespaceId = 1
23     PRP1_Low = 0x3F246000
24     PRP1_High = 0x8
25     NumLBlocks = 1 }
26 ; Write Queue 1 doorbell
27 packet="Temp_OneDwordWrite"
28 {
29     Tag = 0
30     Address = ( CONTROLLER_REGISTERS_BASE + 0x1008 )
31     Payload = ( 55000000 ) }
32 ; Wait for the Controller to process the command. The last thing would be the MSI-X interrupt at vector 1
33 wait=TLP{
34     TLPType = MWr32
35     Address = CQ1_INT_VECTOR_ADDRESS }
36 ; Write Queue 1 Completion Queue Head
37 packet="Temp_OneDwordWrite"{
38     Address = ( CONTROLLER_REGISTERS_BASE + 0x100C )
39     Payload = ( 55000000 ) }
```



NVMe write with queueing

Link Tra	R←	2.5	TLP	Mem	MRd(64)	Length	RequesterID	Tag	Address	1st BE	Last BE	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
34	R←	x8	2852	Mem	MRd(64)	512	001:00:0	69	00000008:C0000000	1111	1111	0	Packet #22966	Metrics	2	32.000 ns
35	R←	x8	2853	Mem	MRd(64)	512	001:00:0	70	00000008:C0000800	1111	1111	0	Packet #22976	Metrics	1	32.000 ns
36	R←	x8	2854	Mem	MRd(64)	512	001:00:0	71	00000008:C0001000	1111	1111	0	Packet #22976	Metrics	1	32.000 ns
37	R←	x8	2855	Mem	MRd(64)	512	001:00:0	72	00000008:C0001800	1111	1111	0	Packet #22976	Metrics	1	80.000 ns
38	R←	x8	2856	Mem	MRd(64)	512	001:00:0	73	00000008:C0002000	1111	1111	0	Packet #22976	Metrics	1	32.000 ns
39	R←	x8	2857	Mem	MRd(64)	512	001:00:0	74	00000008:C0002800	1111	1111	0	Packet #22976	Metrics	1	32.000 ns
40	R←	x8	2858	Mem	MRd(64)	512	001:00:0	75	00000008:C0003000	1111	1111	0	Packet #22976	Metrics	1	32.000 ns
41	R←	x8	2859	Mem	MRd(64)	512	001:00:0	76	00000008:C0003800	1111	1111	0	Packet #22976	Metrics	1	80.000 ns
42	R←	x8	2860	Mem	MRd(64)	512	001:00:0	77	00000008:C0004000	1111	1111	0	Packet #22976	Metrics	2	32.000 ns
43	R←	x8	2861	Mem	MRd(64)	512	001:00:0	78	00000008:C0004800	1111	1111	0	Packet #22986	Metrics	1	32.000 ns
44	R←	x8	2862	Mem	MRd(64)	512	001:00:0	79	00000008:C0005000	1111	1111	0	Packet #22986	Metrics	1	32.000 ns
45	R←	x8	2863	Mem	MRd(64)	512	001:00:0	80	00000008:C0005800	1111	1111	0	Packet #22986	Metrics	1	80.000 ns
46	R←	x8	2864	Mem	MRd(64)	512	001:00:0	81	00000008:C0006000	1111	1111	0	Packet #22986	Metrics	1	32.000 ns
47	R←	x8	2865	Mem	MRd(64)	512	001:00:0	82	00000008:C0006800	1111	1111	0	Packet #22986	Metrics	1	32.000 ns



Flash Memory Summit

NVMe write with no queuing

Link Tra	R→	2.5	TLP	Mem	MRd(64)	Length	RequesterID	Tag	Address	1st BE	Last BE	VC ID	Explicit ACK	Metrics	# Packets	Time Delta	Time Stamp	
2014	x8	3319	Mem	001:00000	512	001:00:0	24	00000008:C0009800	1111	1111	0	Packet #31575		2	96.000 ns	0000.050423234 s		
2015	x8	365	Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
				010:01010	128	001:00:0	196	000:00:0	SC	0	1536	0x00	128 dwords	0	Packet #31589		2	268.000 ns
2016	x8	366	Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
				010:01010	128	001:00:0	186	000:00:0	SC	0	1024	0x00	128 dwords	0	Packet #31595		2	308.000 ns
2017	x8	367	Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
				010:01010	128	001:00:0	186	000:00:0	SC	0	512	0x00	128 dwords	0	Packet #31600		2	272.000 ns
2018	x8	368	Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
				010:01010	128	001:00:0	197	000:00:0	SC	0	2048	0x00	128 dwords	0	Packet #31606		2	8.000 ns
2019	x8	3320	Mem	MRd(64)	Length	RequesterID	Tag	Address	1st BE	Last BE	VC ID	Explicit ACK	Metrics	# Packets	Time Delta	Time Stamp		
				001:00000	512	001:00:0	25	00000008:C000A000	1111	1111	0	Packet #31590		2	272.000 ns	0000.050424186 s		
2020	x8	369	Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
				010:01010	128	001:00:0	197	000:00:0	SC	0	1536	0x00	128 dwords	0	Packet #31612		2	272.000 ns
2021	x8	370	Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
				010:01010	128	001:00:0	187	000:00:0	SC	0	1024	0x00	128 dwords	0	Packet #31614		2	264.000 ns
2022	x8	371	Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
				010:01010	128	001:00:0	187	000:00:0	SC	0	512	0x00	128 dwords	0	Packet #31620		2	228.000 ns
2023	x8	372	Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta
				010:01010	128	001:00:0	198	000:00:0	SC	0	2048	0x00	128 dwords	0	Packet #31626		2	188.000 ns
2024	x8	3321	Mem	MRd(64)	Length	RequesterID	Tag	Address	1st BE	Last BE	VC ID	Explicit ACK	Metrics	# Packets	Time Delta	Time Stamp		
				001:00000	512	001:00:0	26	00000008:C000A800	1111	1111	0	Packet #31615		2	80.000 ns	0000.050425410 s		

Santa Clara, CA
August 2017



Flash Memory Summit

NVMe performance optimization and stress testing

Teledyne LeCroy (Protocol Solutions Group)

3385 Scott Boulevard
Santa Clara, CA 95054

Phone: 800-909-7211 or 408-727-6600

Fax: 408-727-0800

Email Sales: contact.corp@teledynelecroy.com

Email Support:

psgsupport@teledynelecroy.com (Protocol Analyzers)

Web Site: <http://teledynelecroy.com/>

Phone Support: 1-800-909-7112 or 408-653-1260



Flash Memory Summit

Backup

- SGL decoding challenges
- PCIe throughput analysis
- Long recordings analysis



Flash Memory Summit

SGL decoding challenges

If (1) was a segment descriptor (2) can contain SGL segment descriptors, data descriptors, bit bucket descriptors or last segment descriptor.

- For each segment descriptor or data block, keyed data block or last segment descriptor there will be only ONE corresponding NVMe transaction line matching the corresponding address in the descriptor. This is how we implement a PRP transaction level line, based on PRP list
- For a bit bucket descriptor there will be NO corresponding NVMe transaction line
- No case will produce an NVMe transaction line that will not correspond to an address specified in a descriptor from a preceding segment
- Duplications, missing data or incorrect addresses should not create new transaction layer lines.
- Addresses outside the descriptor address range definition (address and length) will be classified unassociated traffic
- Missing address should be marked as error with a tooltip listing the missing address range
- Duplications can be optionally shown as errors, pointing to the duplicated packets.
- See examples in the bottom for cases where multiple NVMe transaction lines are shown for same descriptor



SGL decoding challenges

Flash Memory Summit

- Only the pointed to lines should show as NVMe transaction lines containing the complete range.
- If there are duplications they should collapse into that line.
- If addresses are missing they should be indicated as errors with tooltips showing the missing addresses.

Remove

NVMe	H	Device ID	QID	CID	Address	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length
529	H	157.00.0	0x0001	0x0008	00000020.7E0427A80	SGL	SGL Data Block	00000020.7E009800	0x00002080	0	SGL	SGL Data Block	00000020.7E00B880	0x0001E780
		Time Delta	Time Stamp											
		3.969 us	0047.532 993 218 s											
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
538	H	157.00.0	0x0001	0x0008	00000020.7E009800	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	383.750 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
539	H	157.00.0	0x0001	0x0008	00000020.7E009A00	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	412.500 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
540	H	157.00.0	0x0001	0x0008	00000020.7E009C00	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	406.250 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
541	H	157.00.0	0x0001	0x0008	00000020.7E009E00	SGL Data	0x00000110	272 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	4	804.500 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
543	H	157.00.0	0x0001	0x0008	00000020.7E00A220	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	402.000 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
544	H	157.00.0	0x0001	0x0008	00000020.7E00A420	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	416.500 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
546	H	157.00.0	0x0001	0x0008	00000020.7E00A620	SGL Data	0x00000110	272 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	4	796.250 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
548	H	157.00.0	0x0001	0x0008	00000020.7E00AA40	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	408.250 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
549	H	157.00.0	0x0001	0x0008	00000020.7E00AC40	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	414.500 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
550	H	157.00.0	0x0001	0x0008	00000020.7E00AE40	SGL Data	0x00000110	272 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	4	792.250 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
551	H	157.00.0	0x0001	0x0008	00000020.7E00B060	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	430.250 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
552	H	157.00.0	0x0001	0x0008	00000020.7E00B460	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	408.500 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
553	H	157.00.0	0x0001	0x0008	00000020.7E00B860	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	416.250 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
554	H	157.00.0	0x0001	0x0008	00000020.7E00B880	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	392.000 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
555	H	157.00.0	0x0001	0x0008	00000020.7E00BA80	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	406.250 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
556	H	157.00.0	0x0001	0x0008	00000020.7E00BC80	SGL Data	0x00000088	136 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	2	410.500 ns		
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta		
557	H	157.00.0	0x0001	0x0008	00000020.7E00BE80	SGL Data	0x00000110	272 dwords	SAMSUNG MZWLL1T6HEHP-00003	Metrics	4	798.250 ns		

Remove



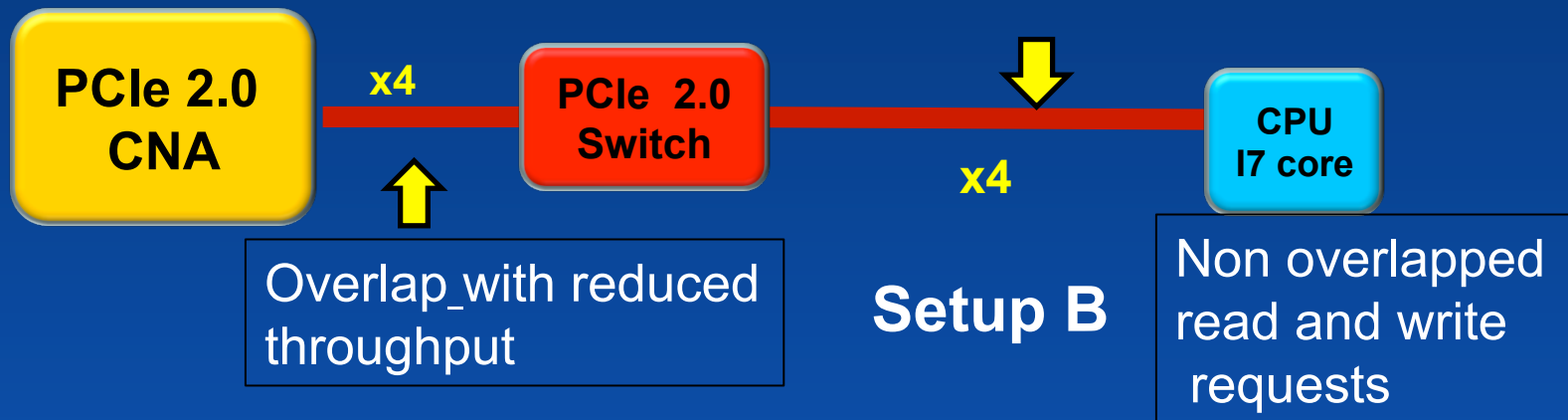
SGL – multiple data block descriptors

NVMe	H	Device ID	QID	SCyTDBL	IO SQT	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp												
0		158.00:0	0x0003		0x0037	HUSMR7638BDP3Y1		1	268.000 ns	0005.202.880.068 s												
NVMe	H	Device ID	QID	CID	Address	IOSQ	OPC	FUSE	PSDT	CID	NSID	MPTR	Address	SGL	Type	Address	Length	Zero	SLBA	NLB	PRINFO	PRCHK
1		158.00:0	0x0003	0x0006	00000000:843C4D80		Write	Normal operation	SGL	0x0006	0x00000001		00000000:00000000	SGL	SGL Segment	00000000:8404BDD8	0x00000100	0	00000000:00001000	0x00FF	PRINFO	100
PRACT	FUA	LR	AF	AL	SR	ILBRT	LBAT	LBATM	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp									
0	1	0	No frequency info	None	0	0	0x00000000	0x0000	0xFFFF	HUSMR7638BDP3Y1		1	6.380 us	0005.202.880.336 s								
NVMe	H	Device ID	QID	CID	Address	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL	
2		158.00:0	0x0003	0x0006	00000000:8404BDD8	SGL	SGL Data Block	00000005:39E0E500	0x00000208	0	SGL	SGL Data Block	00000005:39E0E800	0x00000208	0	SGL	SGL Data Block	00000005:39E0EB00	0x00000208	0	SGL	
Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL			
SGL Data Block	00000005:39E0EE00	0x00000208	0	SGL	SGL Data Block	00000005:39E0F100	0x00000208	0	SGL	SGL Data Block	00000005:39E0F400	0x00000208	0	SGL	SGL Data Block	00000005:39E0F700	0x00000208	0	SGL			
Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL			
SGL Data Block	00000005:39E0FA00	0x00000208	0	SGL	SGL Data Block	00000005:39E0FD00	0x00000208	0	SGL	SGL Data Block	00000005:39E10000	0x00000208	0	SGL	SGL Data Block	00000005:39E10300	0x00000208	0	SGL			
Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL	Type	Address	Length	Zero	SGL			
SGL Data Block	00000005:39E10600	0x00000208	0	SGL	SGL Data Block	00000005:39E10900	0x00000208	0	SGL	SGL Data Block	00000005:39E10C00	0x00000208	0	SGL	SGL Data Block	00000005:39E10F00	0x00000208	0	SGL			
Address	Length	Zero	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp															
00000000:8404BEF0	0x00000100	0	HUSMR7638BDP3Y1		2	5.612 us	0005.202.886.716 s															
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp									
3		158.00:0	0x0003	0x0006	00000005:39E0E500		0x00000082	130 dwords	HUSMR7638BDP3Y1		2	1.956 us	0005.202.892.328 s									
NVMe	H	Device ID	QID	CID	Address	SGL Data	Data Len	Data	MN	Metrics	# Link & Split Trans	Time Delta	Time Stamp									
4		158.00:0	0x0003	0x0006	00000005:39E0E800		0x00000082	130 dwords	HUSMR7638BDP3Y1		2	31.342 us	0005.202.894.284 s									



Setup B - The Analysis Process

- What is preventing the full duplex bus utilization of the read and write requests from maximizing data throughput in Setup B as opposed to Setup A configuration that shows maximum link utilization in both directions?





Setup B – instantaneous vs Overall performance

Link Tra	R	5.0	TLP	Mem	MWrr(32)	Length	RequesterID	Tag	Address	1st BE	Last BE	Data	VC ID	ExplicitACK	Packet #110	Metrics	# Packets	Resp. time	Pld. Bytes	Thrpt MB/s	Time Delta
34	x4	774	Mem	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
35	x4	775	Mem	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
36	x4	776	Mem	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
37	x4	777	Mem	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
38	x4	778	Mem	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
39	x4	779	Mem	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
40	x4	780	Mem	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
41	x4	781	Mem	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
0	x4	MRd	000:00	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
1	x4	MRd	000:00	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
2	x4	MRd	000:00	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
3	x4	MRd	000:00	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
4	x4	MRd	000:00	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
5	x4	MRd	000:00	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		
6	x4	MRd	000:00	010:00000	32	003:00:0	6	2F533980	1111	1111	32 dwords	0	Packet #110	Metrics	2	128.000 ns	128	953.674	80.000 ns		

Timing Calculator - [read_write_traffic_between_x86_and_plx.pex]

From beginning of: Packet 0 To beginning of: Packet 91227

Marker: Time 0.0000010020 secs Time 0.0021786420 secs

Total Time: 2.178 milliseconds

Bus Utilization		
	Upstream	Downstream
Link Utilization	41.876 %	40.971 %
Time Coverage	41.826 %	40.925 %
Bandwidth	8375.20 Mb/s	8194.17 Mb/s
Data Throughput	544.33 MB/s	550.34 MB/s
Packets/second	19043092.52	22849506.81

Split Transaction Performance			
	Minimum	Average	Maximum
Response Time	184.000 ns	363.540 ns	774.000 ns
Latency	118.000 ns	196.070 ns	598.000 ns
Throughput (MB/s)	40.799	328.028	554.865

Memory Writes Performance			
	Minimum	Average	Maximum
Response Time	36.000 ns	121.160 ns	230.000 ns
Throughput (MB/s)	76.294	1018.831	1649.599

inkTras	Resp. time	Latency	Thrpt MB/s	Pld. Bytes	Time Delta
2	304.000 ns	164.000 ns	200.774	64	160.000 ns
2	344.000 ns	166.000 ns	354.856	128	32.000 ns
2	406.000 ns	216.000 ns	300.666	128	48.000 ns
2	428.000 ns	242.000 ns	285.211	128	80.000 ns
2	432.000 ns	244.000 ns	282.570	128	80.000 ns
2	440.000 ns	246.000 ns	277.433	128	104.000 ns
2	408.000 ns	220.000 ns	299.192	128	80.000 ns



Setup B viewed in Link Tracker

Split Tra	5.0	Mem	MRd(32)	RequesterID	CompleterID	Tag	TC	VC ID	Address	Status	Data	Metrics	# LinkTras	Resp. time	Latency	Thrpt MB/s	Pld. Bytes	Time Delta	Time Stamp
17537	x4	00.00000	000.00.0	001.00.0	6	0	0	A12C9F80	SC	32 dwords		2	4.652 µs	4.450 µs	26.240	128	12.000 ns	0000_003 214 038 s	

Link Tra	5.0	TLP	Mem	MW(32)	Length	RequesterID	Tag	Address	1st BE	Last BE	Data	VC ID	Explicit ACK	Metrics	# Packets	Resp. time	Pld. Bytes	Thrpt MB/s	Time Delta	Time Stamp
52376	x4	1163	10.00000	32	000.00.0	22	A1512380	1111	1111	32 dwords	0	Packet#153706	2	168.000 ns	128	726.609	76.000 ns	0000_003 214 050 s		
52377	x4	1164	10.00000	32	000.00.0	23	A1512400	1111	1111	32 dwords	0	Packet#153709	2	168.000 ns	128	726.609	76.000 ns	0000_003 214 126 s		

Time	Packet #	Upstream	Downstream
00.003 213 952		A3	A3
00.003 213 954		32	27
00.003 213 956		19	6B
00.003 213 958		F8	F2
00.003 213 960		45	E3
00.003 213 962		BF	97
00.003 213 964		E7	CE
00.003 213 966		48	5C
00.003 213 968		40	81
00.003 213 970		A5	3E
00.003 213 972		9C	A2
00.003 213 974		65	B5
00.003 213 976		3B	7F
00.003 213 978		E3	A4
00.003 213 980		CF	AB
00.003 213 982		17	8D
00.003 213 984		76	0A
00.003 213 986		A6	83
00.003 213 988		85	BF
00.003 213 990		9A	46
00.003 213 992		F7	77
00.003 213 994		3C	A4
00.003 213 996		60	F5
00.003 214 000		D6	1F
00.003 214 002		84	EE
00.003 214 004		72	C2
00.003 214 006		72	B9
00.003 214 008		72	A4
00.003 214 010		57	AE
00.003 214 012		77	AE
00.003 214 014		CD	0C
00.003 214 016			FB

- One direction Memory writes
- No downstream completions



Setup B: between RC and switch FC Buffer Credit analysis

FC Credits updating properly early in trace

Packet 4283	R→	5.0 x4	FC-Cpl	VC ID 0	HdrFC 126	DataFC 2036	TLP 1824	Cpl	CplID 10:01010	Length 32	RequesterID 003:00:0	Tag 1	CompleterID 000:00:0	Status SC	BCM 0	Byte Cnt 128	Lwr Addr 0x00	Data 32 dwords	LCRC 0x4842C3D7	FC-Cpl	VC 0	
Packet 4284	R→	5.0 x4	FC-NP	VC ID 0	HdrFC 71	DataFC infinite	TLP 1687	Mem	MRd(32)	Length 32	RequesterID 003:00:0	Tag 3	Address 2F32A500	1st BE 1111	Last BE 1111	LCRC 0xBC41270C	FC-NP	VC ID 0	HdrFC 70	DataFC infinite	Time Delta 16.000 ns	Tin 0
Packet 4286	R→	5.0 x4	FC-P	VC ID 0	HdrFC 113	DataFC 300	TLP 1688	Mem	MWrr(32)	Length 32	RequesterID 003:00:0	Tag 15	Address 2F627200	1st BE 1111	Last BE 1111	Data 32 dwords	LCRC 0x43199B05	FC-P	VC ID 0	HdrFC 112	DataFC 292	Tin 6

FC Credits updating properly later in trace

Packet 29939	R→	5.0 x4	DLLP	ACK	AckNak_Seq_Num 583	CRC 16 0xA117	Idle 0.000 ns	Time Stamp 0000.000720128 s														
Packet 29940	R→	5.0 x4	FC-NP	VC ID 0	HdrFC 70	DataFC infinite	TLP 3264	Mem	MRd(32)	Length 32	RequesterID 003:00:0	Tag 7	Address 2F411D00	1st BE 1111	Last BE 1111	LCRC 0xA11AD34B	FC-NP	VC ID 0	HdrFC 69	DataFC infinite	Time Delta 14.000 ns	Tin 00
Packet 29942	R→	5.0 x4	FC-P	VC ID 0	HdrFC 113	DataFC 300	TLP 3265	Mem	MWrr(32)	Length 32	RequesterID 003:00:0	Tag 17	Address 2F58DC00	1st BE 1111	Last BE 1111	Data 32 dwords	LCRC 0x59DE7FFB	FC-P	VC ID 0	HdrFC 112	DataFC 292	Tin 14

FC updating properly for posted and non posted transactions



Setup B: Capture between switch and EP

Flash Memory Summit

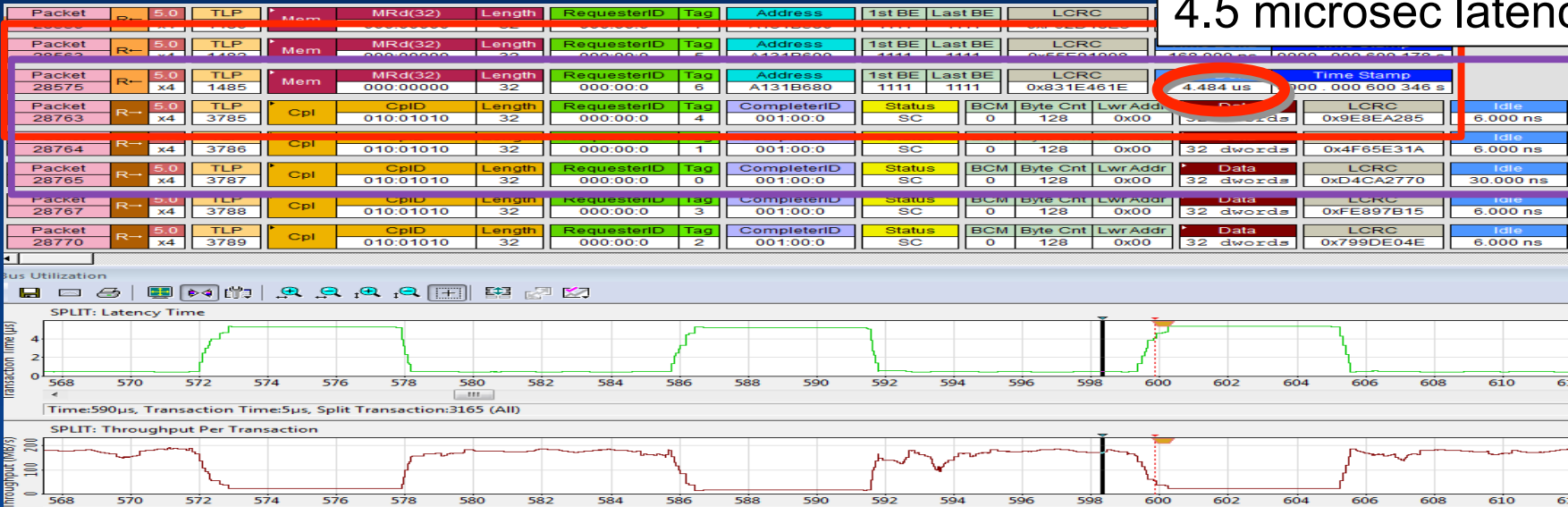
- Here we observe some overlap, but read throughput goes significantly down during write transfers.





Setup B: viewed in Packet Mode

4.5 microsec latency

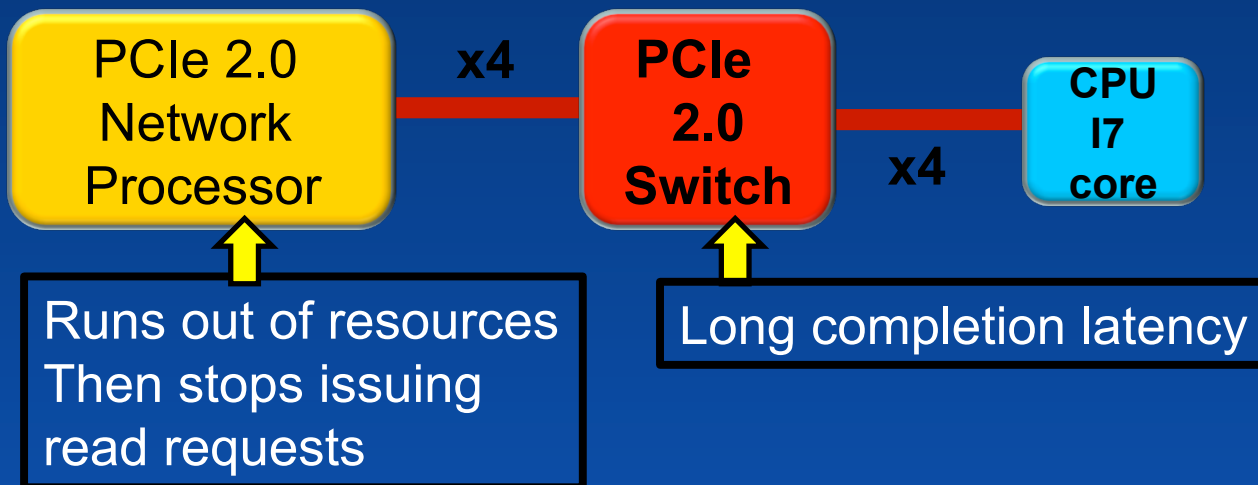


- We now display read transactions with their corresponding completions.
- Read requests tend to accumulate, varying their tags, with no completions until read requests stop.



The Analysis process

- The root cause for the stalled requests is the long completion latency from the switch side.
- We now question – why does the endpoint Network Processor stop issuing requests?
- Option 1: The endpoint Network Processor has to allocate resources for the queued up completions. This continues until the endpoint NP runs out of resources and then stops issuing read requests, note the 4.5 usec stall of read requests once all resources/tags have been exhausted.





Flash Memory Summit

Read Completion Credit analysis

Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
	010:01010	32	000:00:0	2	001:00:0	SC	0	128	0x00	32 dwords	0x6A40BEE2	52.000 ns
UpdateFC-Cpl	VC ID	HdrFC	DataFC	CRC 16	Time Delta	Time Stamp						
	0	109	2199	0x4B92	28.000 ns	0000 . 000 598 898 s						
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
	010:01010	32	000:00:0	5	001:00:0	SC	0	128	0x00	32 dwords	0xD5D48F20	52.000 ns
UpdateFC-Cpl	VC ID	HdrFC	DataFC	CRC 16	Time Delta	Time Stamp						
	0	110	2207	0x77FC	14.000 ns	0000 . 000 598 978 s						
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Idle
	010:01010	32	000:00:0	1	001:00:0	SC	0	128	0x00	32 dwords	0x4F451C69	12.000 ns
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
	010:01010	32	000:00:0	7	001:00:0	SC	0	128	0x00	32 dwords	0x9942B7F7	4.000 ns
UpdateFC-Cpl	VC ID	HdrFC	DataFC	CRC 16	Time Delta	Time Stamp						
	0	111	2215	0x9061	76.000 ns	0000 . 000 599 082 s						
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
	010:01010	32	000:00:0	4	001:00:0	SC	0	128	0x00	32 dwords	0xF8658EEB	176.000 ns
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
	010:01010	32	000:00:0	0	001:00:0	SC	0	128	0x00	32 dwords	0xDB75DD52	240.000 ns
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
	010:01010	32	000:00:0	2	001:00:0	SC	0	128	0x00	32 dwords	0xB09014F1	82.000 ns
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
	010:01010	32	000:00:0	5	001:00:0	SC	0	128	0x00	32 dwords	0x5CCAD1C4	222.000 ns
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Time Delta
	010:01010	32	000:00:0	7	001:00:0	SC	0	128	0x00	32 dwords	0x75CE9B7E	4.828 us
UpdateFC-Cpl	VC ID	HdrFC	DataFC	CRC 16	Time Delta	Time Stamp						
	0	119	2279	0x9FEF	124.000 ns	0000 . 000 604 706 s						
Cpl	CplID	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	LCRC	Idle
	010:01010	32	000:00:0	4	001:00:0	SC	0	128	0x00	32 dwords	0x9E8EA285	6.000 ns



Flash Memory Summit

Do we need to upgrade our hardware?

- Do we need more speed, higher link width, faster processor or more hardware acceleration?
 - Upgrading will create more heat issues so new mechanics and cooling mechanisms may be needed.
 - Upgrading means higher cost
 - Upgrading means more complexity
 - Upgrading means longer time to market and extended project completion
- With the correct tool one will be able to tell if his complex fabric has low performance because of low bandwidth and needs to be upgraded, or because of an unutilized link, pointing to the root cause of the low performance and exact component / issue to be addressed



Flash Memory Summit

NVMe Performance bottlenecks

- Performance bottlenecks in NVMe based platforms
 - PCI express flow control credit analysis.
 - Tradeoffs between memory resource allocation and data throughput
 - Virtual channels manage traffic in fabric
 - Allocate traffic classes to Virtual channels



Flash Memory Summit

Trace Expert- Performance Analysis

- Drill down to  get more reports

Performance Analysis

- [Link Transaction Performance](#)
- [Split Transactions Performance](#)
- [NVMe Performance](#)

Link Transaction Performance

Performance

Transaction Type	Total	# Packets (Min)	# Packets (Avg)	# Packets (Max)	Resp. time (Min)	Resp. time (Avg)	Resp. time (Max)	Pld. Bytes (Min)	Pld. Bytes (Avg)	Pld. Bytes (Max)
MsgD	1	2	2.00	2	1.412 us	1.412 us	1.412 us	4	4.00	4
CfgRd0	1072	2	2.00	2	1.236 us	1.375 us	1.620 us	0	0.00	0
CpID	565132	2	2.00	2	84.000 ns	1.636 us	2.396 us	1	63.87	64
CfgWr0	86	2	2.00	2	1.316 us	1.401 us	1.604 us	1	2.47	4
Cpl	163	2	2.00	2	68.000 ns	188.340 ns	308.000 ns	0	0.00	0
MWr(32)	2939	2	2.00	2	76.000 ns	1.011 us	1.692 us	4	4.00	4
MRd(32)	3	2	2.00	2	1.404 us	1.625 us	1.740 us	0	0.00	0
MRd(64)	71725	2	2.00	2	76.000 ns	202.820 ns	468.000 ns	0	0.00	0
MWr(64)	29700	2	2.00	2	148.000 ns	696.740 ns	988.000 ns	16	123.37	128
670821										

Memory Writes Performance


Requester, TC	Total	Resp. time (Min)	Resp. time (Avg)	Resp. time (Max)	Pld. Bytes (Min)	Pld. Bytes (Avg)	Pld. Bytes (Max)	Thrpt MB/s (Min)	Thrpt MB/s (Avg)	Thrpt MB/s (Max)
000:00:0, TCO	1973	1.268 us	1.407 us	1.692 us	4	4.00	4	2.255	2.715	3.008
006:00:0, TCO	30667	76.000 ns	681.170 ns	988.000 ns	4	119.60	128	10.039	162.247	207.603
32640										

Split Transactions Performance

Overall Performance

Requester -> Completer	Total	# Packets (Min)	# Packets (Avg)	# Packets (Max)	Resp. time (Min)	Resp. time (Avg)	Resp. time (Max)
000:00:0 -> 000:00:0	6	2	2	2	102.880 us	123.771 us	179.312 us
000:00:0 -> 006:00:0	1155	2	2	2	99.648 us	118.388 us	647.712 us
006:00:0 -> 000:00:0	71718	2	2	9	1.848 us	6.673 ms	9.859 ms
72879							

More Reports





Recording duration using 32K DW NVMe Transfers

Best Case
Gen 1 x1 Lane Width

Worst Case
Gen 3 x4 Lane Width

TLP size (dw)	Link Utilization (%)				
	100	90	80	70	60
32	139.8164	155.3516	174.7705	199.7378	233.0274
64	128.817	143.13	161.0213	184.0243	214.695
128	123.3173	137.0192	154.1466	176.1676	205.5289
256	120.5675	133.9638	150.7093	172.2392	200.9458
512	119.1925	132.4361	148.9907	170.275	198.6542
1024	118.5051	131.6723	148.1313	169.293	197.5084
2048	118.1613	131.2904	147.7017	168.8019	196.9356
4096	117.9895	131.0994	147.4868	168.5564	196.6491

TLP size (dw)	Link Utilization (%)				
	100	90	80	70	60
32	8.875066	9.861185	11.09383	12.67867	14.79178
64	8.176861	9.085402	10.22108	11.68123	13.6281
128	7.827759	8.69751	9.784699	11.18251	13.04627
256	7.653208	8.503564	9.56651	10.93315	12.75535
512	7.565932	8.406591	9.457415	10.80847	12.60989
1024	7.522294	8.358105	9.402868	10.74613	12.53716
2048	7.500475	8.333862	9.375594	10.71496	12.50079
4096	7.489566	8.32174	9.361958	10.69938	12.48261

****Note: Capture Duration Doubles when using T34 expanded mode**