NVMe scaling with IO concurrency

Tobias Oberstein, Crossbar.io GmbH, 22.1.2017

Random 8KB IOPS, measured with FIO, directly over 16 raw NVMe block devices

IO Concurrency:			1	2	4	8	16	32	64	128	256	384	512	768	1024	1536	2048
READ		measured	**8.411**	**15.924**	**35.120**	**68.215**	**137.012**	**258.000**	**525.756**	**943.872**	**1.762.100**	**2.381.700**	**2.776.800**	**3.426.700**	**3.889.900**	**4.326.500**	**4.597.700**
		theor. linear scaling		16.822	33.644	67.288	134.576	269.152	538.304	1.076.608	2.153.216	3.229.824	4.306.432	6.459.648	8.612.864	12.919.296	17.225.728
		theor. linear scaling %		94,7%	104,4%	101,4%	101,8%	95,9%	97,7%	87,7%	81,8%	73,7%	64,5%	53,0%	45,2%	33,5%	26,7%
WRITE		measured	**61.482**	**124.920**	**241.626**	**452.979**	**847.455**	**1.300.700**	**1.467.700**	**1.656.400**	**1.953.500**	**2.033.700**	**2.020.100**	**2.113.700**	**2.208.800**	**2.265.200**	**2.249.600**
		theor. linear scaling		122.964	245.928	491.856	983.712	1.967.424	3.934.848	7.869.696	15.739.392	23.609.088	31.478.784	47.218.176	62.957.568	94.436.352	125.915.136
		theor. linear scaling %		101,6%	98,3%	92,1%	86,1%	66,1%	37,3%	21,0%	12,4%	8,6%	6,4%	4,5%	3,5%	2,4%	1,8%
R/W-70/30	**R**	measured	**6.912**	**14.004**	**28.225**	**55.033**	**104.963**	**191.081**	**335.743**	**571.823**	**864.544**	**1.082.700**	**1.311.300**	**1.518.600**	**1.739.600**	**1.832.500**	**1.871.100**
		theor. linear scaling		13.824	27.648	55.296	110.592	221.184	442.368	884.736	1.769.472	2.654.208	3.538.944	5.308.416	7.077.888	10.616.832	14.155.776
		theor. linear scaling %		101,3%	102,1%	99,5%	94,9%	86,4%	75,9%	64,6%	48,9%	40,8%	37,1%	28,6%	24,6%	17,3%	13,2%
	W	measured	**2.975**	**6.007**	**12.102**	**23.690**	**44.986**	**81.870**	**143.749**	**245.137**	**370.542**	**463.747**	**562.054**	**651.064**	**746.121**	**785.532**	**802.120**
		theor. linear scaling		5.950	11.900	23.800	47.600	95.200	190.400	380.800	761.600	1.142.400	1.523.200	2.284.800	3.046.400	4.569.600	6.092.800
		theor. linear scaling %		101,0%	101,7%	99,5%	94,5%	86,0%	75,5%	64,4%	48,7%	40,6%	36,9%	28,5%	24,5%	17,2%	13,2%

Random 8KB IOPS, measured with FIO, over Linux MD RAID-0 over 16 NVMe block devices

IO Concurrency:			1	2	4	8	16	32	64	128	256	384	512	768	1024	1536	2048
READ		measured	**8.125**	**16.227**	**32.514**	**68.252**	**131.004**	**257.886**	**517.045**	**955.146**	**1.712.600**	**1.815.100**	**1.857.200**	**1.474.800**	**1.456.200**	**1.451.500**	**1.442.700**
		theor. linear scaling		16.250	32.500	65.000	130.000	260.000	520.000	1.040.000	2.080.000	3.120.000	4.160.000	6.240.000	8.320.000	12.480.000	16.640.000
		theor. linear scaling %		99,9%	100,0%	105,0%	100,8%	99,2%	99,4%	91,8%	82,3%	58,2%	44,6%	23,6%	17,5%	11,6%	8,7%
WRITE		measured	**62.076**	**122.440**	**235.765**	**447.516**	**827.292**	**1.236.800**	**1.235.100**	**1.235.300**	**1.264.600**	**1.263.700**	**1.242.300**	**1.265.600**	**1.232.900**	**1.237.500**	**1.239.700**
		theor. linear scaling		124.152	248.304	496.608	993.216	1.986.432	3.972.864	7.945.728	15.891.456	23.837.184	31.782.912	47.674.368	63.565.824	95.348.736	127.131.648
		theor. linear scaling %		98,6%	95,0%	90,1%	83,3%	62,3%	31,1%	15,5%	8,0%	5,3%	3,9%	2,7%	1,9%	1,3%	1,0%
R/W-70/30	**R**	measured	**6.814**	**14.117**	**28.302**	**57.097**	**106.936**	**193.242**	**339.172**	**574.925**	**913.081**	**951.073**	**949.782**	**944.325**	**952.329**	**957.530**	**957.075**
		theor. linear scaling		13.628	27.256	54.512	109.024	218.048	436.096	872.192	1.744.384	2.616.576	3.488.768	5.233.152	6.977.536	10.466.304	13.955.072
		theor. linear scaling %		103,6%	103,8%	104,7%	98,1%	88,6%	77,8%	65,9%	52,3%	36,3%	27,2%	18,0%	13,6%	9,1%	6,9%
	W	measured	**2.931**	**6.049**	**12.143**	**24.505**	**45.923**	**82.851**	**145.524**	**246.448**	**391.516**	**407.710**	**407.194**	**404.646**	**408.157**	**410.316**	**410.179**
		theor. linear scaling		5.862	11.724	23.448	46.896	93.792	187.584	375.168	750.336	1.125.504	1.500.672	2.251.008	3.001.344	4.502.016	6.002.688
		theor. linear scaling %		103,2%	103,6%	104,5%	97,9%	88,3%	77,6%	65,7%	52,2%	36,2%	27,1%	18,0%	13,6%	9,1%	6,8%

Random 8KB IOPS, measured with FIO, over XFS over Linux MD RAID-0 over 16 NVMe block devices

IO Concurrency:			1	2	4	8	16	32	64	128	256	384	512	768	1024	1536	2048
READ		measured	**8.393**	**16.529**	**33.910**	**69.439**	**129.680**	**260.787**	**517.743**	**969.833**	**1.743.200**	**2.377.400**	**2.382.400**	**2.608.900**	**2.759.400**	**2.797.300**	**2.805.800**
		theor. linear scaling		16.786	33.572	67.144	134.288	268.576	537.152	1.074.304	2.148.608	3.222.912	4.297.216	6.445.824	8.594.432	12.891.648	17.188.864
		theor. linear scaling %		98,5%	101,0%	103,4%	96,6%	97,1%	96,4%	90,3%	81,1%	73,8%	55,4%	40,5%	32,1%	21,7%	16,3%
WRITE		measured	**76.467**	**144.884**	**271.683**	**502.248**	**896.512**	**1.363.100**	**1.151.900**	**1.148.800**	**168.484**	**32.908**	**33.842**	**27.680**	**27.985**	**29.054**	**28.090**
		theor. linear scaling		152.934	305.868	611.736	1.223.472	2.446.944	4.893.888	9.787.776	19.575.552	29.363.328	39.151.104	58.726.656	78.302.208	117.453.312	156.604.416
		theor. linear scaling %		94,7%	88,8%	82,1%	73,3%	55,7%	23,5%	11,7%	0,9%	0,1%	0,1%	0,0%	0,0%	0,0%	0,0%
R/W-70/30	**R**	measured	**7.616**	**15.110**	**29.746**	**59.655**	**113.547**	**203.396**	**343.793**	**575.670**	**868.207**	**85.320**	**87.375**	**125.948**	**153.599**	**161.851**	**160.108**
		theor. linear scaling		15.232	30.464	60.928	121.856	243.712	487.424	974.848	1.949.696	2.924.544	3.899.392	5.849.088	7.798.784	11.698.176	15.597.568
		theor. linear scaling %		99,2%	97,6%	97,9%	93,2%	83,5%	70,5%	59,1%	44,5%	2,9%	2,2%	2,2%	2,0%	1,4%	1,0%
	W	measured	**3.264**	**6.482**	**12.732**	**25.489**	**48.727**	**87.237**	**147.452**	**246.854**	**372.283**	**36.669**	**37.508**	**54.083**	**65.850**	**69.458**	**68.705**
		theor. linear scaling		6.528	13.056	26.112	52.224	104.448	208.896	417.792	835.584	1.253.376	1.671.168	2.506.752	3.342.336	5.013.504	6.684.672
		theor. linear scaling %		99,3%	97,5%	97,6%	93,3%	83,5%	70,6%	59,1%	44,6%	2,9%	2,2%	2,2%	2,0%	1,4%	1,0%

latency optimized
best balance (throughput without overload)
througput maximized