Supermicro's Storage AOC

Introduction

Supermicro has been and continues to employ and deliver storage products with the latest storage technologies. In this white paper I will only cover the AOC (Add On Card) portion. Supermicro's AOCproduct line offerings are in three categories:

- HBA – Host Bus Adapter where each physical device (i.e. PD) is presented to
  OS separately. If two PDs are connected to an HBA then OS will see two distinct devices.
- SW RAID – These AOCs are very similar to the architecture which is used in the HBA except they support minimal amount of RAID features (i.e. 0, 1 and 10). Both AOC as well as mother board's (i.e. MB) resources are used to process data. More than one PD can be configured into a virtual device (i.e. VD). The operating system (i.e. OS) will see VDs as distinct devices.
- RAID On Chip (i.e. ROC) based RAID – These cards are rich in RAID features. They support RAID level 0,1,10,5,6,50 and 60. This AOC contains cache memory and cache memory protection (i.e. Battery Back Unite (i.e. BBU) or Cachevault (Supecap+TFM)).

Now I will describe each of the above categories more in detail by covering their:

1. Features
2. Performance & Capacity
3. Cable Types
4. SKUs and model names

Features

- HBA (IT mode) –

This AOC provides one-to-one mapping of PDs (i.e. Physical Devices) and exports them back to the OS. It does not provide any RAID feature. It supports 8 devices if it is directly connected to a DA (i.e. Direct attached) backplane that would also support 8 devices. It however can support 122 or more PDs if it is connected to a backplane that uses an expander. The higher device support is possible by using the logic in the backplane expander. This AOC provides to the OS, all distinct devices that it can find. The OS can either use them as individual devices or use SW RAID to create OS level VDs that utilizes a desired RAID levels. Higher number of devices can be configured by using just a bunch of devices (i.e. JBOD – SMCI defines this as a chassis that is externally available for storage scalability). These JOBDs can be chained for scalability and grow a capacity in a enterprise software that can handle on-the-fly scalability such as hadoop (http://en.wikipedia.org/wiki/Hadoop) , Swift (*http://en.wikipedia.org/wiki/OpenStack)* and many other OpenSource Scalable storage products. This card does not need cache therefore no cache protection (i.e. BBU) is required.

- SW RAID (IR mode) –

This AOC provides minimal RAID features (i.e. RAID 0,1 and 10) and exports them in the form of VDs (i.e. Virtual Devices) back to the OS. Each VD is made of several PDs. For example a VD could be 3 PDs that are configured as RAID 0. It supports 8 devices if it is directly connected to a DA backplane that would also support 8 devices. You are allowed to have upto 8 PDs in each VD and upto 2 VDs. It however can support 63 or more PDs if it is connected to a backplane that uses an expander. In this case you are allowed to have 14 PDs in each VD and upto 2 VDs . Higher number of devices are possible by using the logic in both the backplane expander FW (i.e. Firmware – software residing on the hardware) and AOC's. Higher number of VDs could be configured by using drive groups (i.e. DG) and expander based JBOD chassis. SMCI defines JBOD as a expander based chassis that is externally available for storage scalability. This card does not need cache therefore no cache protection (i.e. BBU) is required. A user's guide is also available at SMCI's ftp site *(ftp://ftp.supermicro.com/driver/SAS/LSI/LSI_SAS_EmbMRAID_SWUG.pdf).*

- ROC based RAID –

This AOC uses RAID On Chip (i.e. ROC chip) to provide full RAID features (i.e. RAID 0,1, 10, 6,5,50,60) in the form of VDs (i.e. Virtual Devices) and exports them back to the OS. It supports 8 devices if it is directly connected to a DA backplane that would also support 8 devices. You are allowed to have upto 8 PDs in each VD and upto 8 VDs. It however can support 240 or more PDs if it is connected to a backplane that uses an expander. In this case you are allowed to have 240 PDs in each VD and upto 64 VDs. Higher number of VDs could be configured by using an expander based JBOD chassis and drive group (i.e. DG). This card also supports cache memory for higher performance. The cache memory is protected in the event of a power outage. Technologies that are used to do so are the old fashion BBU or the newer Cachevault (Supecap+TFM). The card also supports many other features that are normally associated w/ a RAID card, see the user's guide at *http://centraldb/cds/sites/default/files/download/user_guide/aoc/54385-00_RevD_12Gbs_MegaRAID_SAS_SW_UserGd_v1-0.pdf*. All features can be set or unset at BIOS and OS (command line or GUI app). SMCI provides latest version of these tools at its ftp site (*ftp://ftp.supermicro.com/driver/SAS/LSI/*).

Performance & Capacity

Let's pick an earlier AOC which uses SAS-2 on PCIe Gen2 w/ either LSI's 2008 or 2108 ASICs. Let's use it as an example to calculate its maximum theoretical performance. It's PCIe Gen 2 side has 8 lanes at 500MB/s that adds up to 4GB/s (i.e. 8X500MB/s). The SAS-2 side uses 8 lanes at 6Gb/s, which adds up to 4.8 GB/s. In this case maximum IO performance is dictated by the slower interface that is the card's performance bottleneck which in this case would be 4GB/s. Let's review the bandwidth speed for different interfaces. Eight lanes of SAS-2 is 4.8GB/s, PCI-e Gen-2 is 4GB/s, SAS-3 is 9.6GB/s and PCI-e Gen-3 is 6.4GB/s. Using the above calculation, the following waves of AOCs, represent maximum theoretical performance:

- First wave - SAS-2/Gen-2 (LSI 2008/LSI 2108)= 4GB/s (bottleneck PCIe Gen-2)
- Second wave - SAS-2/Gen-3 (LSI 2308/LSI 2208)= 4.8GB/s (bottleneck SAS-2)
- Third wave - SAS-3/Gen-3 (LSI 3008/LSI 3108) = 6.4GB/s (bottleneck PCI-e Gen-3)
- Fourth wave – Will use PCI-e Gen-4 due in late 2016

Please note the actual performance is always approximately around 30% less than the theoretical performance. This is due to deficiencies in logic that is used in FW, driver and OS.

Let's use the following two realistic examples to achieve a desired capacity and performance.

Example #1 (Low end configuration):

Let's say you do not care about capacity and do not need any RAID features but need to achieve total read performance of 5GB/s. You need to first calculate total number of HDDs/SSDs that can yield 5GB/s. Read/write spec on performance of Toshiba's SAS-3/12gb PX02SMF020's is defined at *http://www.cdw.com/shop/products/Toshiba-PX02SMF020-solid-state-drive-200-GB-SAS-3/3203414.aspx#TS* to be 900MB/s for sequential read and 400MB/s for sequential write.

Now to find out total number of SSDs you must divide 5GB/s or 5000MB/s (desired) by 900 (read) which is 5.5 SSDs . This means you will need 6 SSDs. SMCI's HBA AOC (IT mode) supports up to 8 direct attached devices. You will not need an expander base chssis. Remember expanders are used when you need more than 8 devices. You need a DA based chassis. The HBA can realistically perform around 5GB/s. So now your HW configuration to achieve 5GB/s read, will be as below:

- One SAS-3/Gen-3 HBA AOC (SKU is AOC-S3008L-L8e)
- 6X Toshiba SAS-3/12Gb PX02SMF020 SSDs
- Any SAS-3 supermico DA backplane that supports at least 6 devices (i.e. SKU is BPN-SAS3-216A or BPN-SAS3-116A).
- Cables – See "Cable Types" below

Please see Supermiro lab's result for the above configurations using iometer in windows:

Sequential Read (MB/s) / Block Size	HDD-0	HDD-1	HDD-2	HDD-3	HDD-4	HDD-5	HDD-6	HDD-7
512B	66.63	66.72	66.62	66.80	66.33	65.43	66.57	66.63
1KB	133.68	133.23	133.30	133.59	133.66	133.57	133.29	133.33
2KB	267.59	267.21	267.32	266.80	267.96	267.84	267.48	267.50
4KB	532.39	531.87	532.23	531.88	532.40	531.52	533.83	533.01
8KB	806.74	800.67	804.81	807.22	806.31	806.23	804.67	803.91
16KB	880.90	882.52	881.16	881.69	880.21	881.77	881.17	880.91
32KB	907.55	910.67	910.04	910.27	907.66	907.68	909.01	909.54
64KB	955.49	956.13	952.68	958.27	954.38	955.41	950.75	954.03
128KB	982.81	980.48	981.47	976.50	977.92	979.49	976.78	977.06
256KB	989.26	992.12	990.73	990.89	988.86	988.06	990.75	987.63
512KB	999.67	994.53	992.68	998.16	993.19	995.44	989.93	990.88
1MB	987.97	988.57	985.66	983.87	984.63	986.01	983.14	986.39
2MB	971.47	973.44	971.47	971.19	966.36	968.49	969.37	970.50
REMARKS								
REF	1A1A1AL							
PBID								

Sequential Write (MB/s) / Block Size	HDD-0	HDD-1	HDD-2	HDD-3	HDD-4	HDD-5	HDD-6	HDD-7
512B	25.83	25.83	25.80	25.77	25.88	25.79	25.83	25.81
1KB	52.66	52.78	52.70	52.71	52.77	52.69	52.79	52.68
2KB	102.85	103.17	102.84	103.17	103.03	102.65	102.93	102.90
4KB	195.32	195.48	195.16	195.32	195.83	195.22	195.68	195.59
8KB	345.95	345.00	346.13	345.41	345.85	346.08	346.48	346.09
16KB	429.53	431.77	429.23	426.62	425.54	428.28	427.83	427.72
32KB	431.07	432.02	429.18	427.36	426.44	431.32	427.91	428.77
64KB	430.59	432.64	429.53	428.26	426.77	430.45	428.72	428.47
128KB	429.94	432.25	430.87	427.48	426.70	429.11	429.51	428.49
256KB	430.78	430.50	428.89	426.54	425.31	429.12	427.79	427.00
512KB	429.75	431.38	428.51	427.22	426.32	429.40	428.52	427.58
1MB	430.05	431.12	429.77	426.15	425.09	428.52	427.57	427.44
2MB	429.65	430.39	428.31	426.34	424.75	429.00	426.72	426.69

Example #2 (High end configuration):

Let's say you need full RAID features, 1 PB of storage and maximum sequential read performance of 15GB/s. You need to first calculate total number of HDDs/SSDs that can yield the capacity of 1PB. Hitachi's HUS726060ALS640 size is 6TB. Dividing 1PB (i.e. 1000TB) by 6TB, results in total number of HDDs which is 167 HDDs. Now to achieve 15GB/s, Read/write performance of Hitachi for SAS-2/6Gb HUS726060ALS640 is defined at *http://www.serversupply.com/HARD%20DRIVES/SAS-6GBITS/6TB-7200RPM/HITACHI%20/HUS726060ALS640.htm* to be 180 MB/s for read and 180 MB/s for write.

To find out total number of drives needed to achieve 15GB/s we must divide 15GB/s (i.e. 15000MB/s) by 180 MB/s. This results to 84 HDDs. So we need 167 HDDs to achieve 1PB of capacity and 84 of these drives will provide the required performance of 15GB/s. Since we need more than 8 devices to achieve our desired requirements, we must use an expander based backplane. SMCI's ROC based AOC supports at least 240 devices, when used with an expander. Again remember expanders are used when you need more than 8 devices. This AOC can realistically perform at least around 5GB/s. So now your HW configuration to achieve 15GB/s read performance and 1PB capacity will be as below:

- 4XSAS-3/Gen-3 ROC AOC (Total number of AOCs are derived from total number of chassis needed, see
- below) (SKU is AOC-S3108L-H8iR)
- 167 Hitachi HUS726060ALS640 SAS-2/6Gb PX02SMF020 SSDs
- 4X847 supermico JBOD chassis that use expander backplanes, each supports 45 devices. To get total of
- JBODs needed for 167 HDDs, divide 167 HDDs by 45 HDDs which is 4 JBODs. (SKU is 847E1CR1K28JBOD)
- Cables – See "Cable Types" below

Supermicro lab's report shows that using a RAID level 0, you can get a realistic performance of slightly above 4GB/s for both sequential read and write per chassis. See below:

RAID PERFORMANCE SAS3-846EL1					SAS3-847EL1			
	RAID 0	RAID 5	RAID 6	RAID 10	RAID 0	RAID 5	RAID 6	RAID 10
HDD QTY / RAID SIZE	24	24	24	24	20	20	20	20

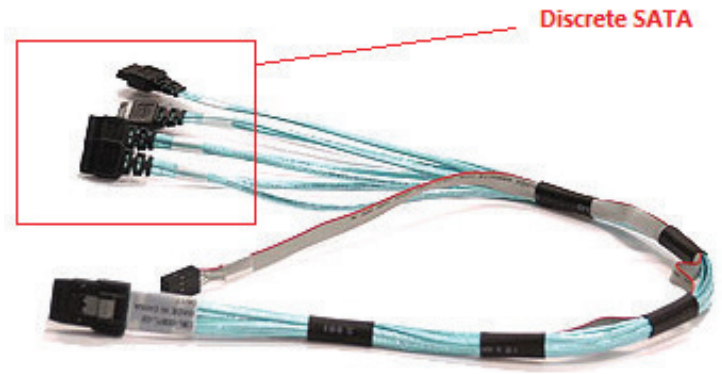
RAID PERFORMANCE SAS3-846EL1					SAS3-847EL1			
Sequential Read (MB/s) / Block Size	RAID-0	RAID-5	RAID-6	RAID-10	RAID-0	RAID-5	RAID-6	RAID-10
512B	99.99	55.42	100.12	107.41	119.54	118.95	119.53	99.17
1KB	161.41	95.99	163.98	197.55	208.11	210.46	210.64	272.81
2KB	224.74	188.97	223.46	290.28	355.58	354.17	351.66	458.76
4KB	365.25	350.92	391.52	473.26	626.65	625.23	630.34	450.33
8KB	650.05	709.11	709.02	1,036.12	1,039.38	1,047.55	1,049.69	996.68
16KB	1,592.06	1,253.01	1,669.70	1,751.76	1,766.13	1,777.92	1,789.92	1,676.30
32KB	3,295.17	1,710.53	3,123.28	2,103.34	3,421.16	3,249.22	3,066.36	1,714.13
64KB	3,945.18	2,473.01	3,293.17	2,123.31	3,421.56	3,235.79	3,079.97	1,725.85
128KB	4,037.69	3,313.57	3,348.59	2,173.96	3,396.03	3,247.05	3,074.51	1,730.28
256KB	4,089.13	3,674.99	3,393.50	2,296.16	3,423.83	3,160.46	3,027.66	1,734.65
512KB	4,102.69	3,646.10	3,396.23	2,873.39	3,423.87	3,200.48	3,054.52	2,097.27
1MB	4,117.45	3,713.64	3,398.66	3,152.20	3,423.97	3,207.74	3,048.26	2,402.54
2MB	4,124.46	3,702.42	3,389.18	3,248.54	3,422.51	3,194.67	3,049.03	2,409.83
REMARKS								

RAID PERFORMANCE SAS3-846EL1					SAS3-847EL1			
Sequential Write (MB/s) / Block Size	RAID-0	RAID-5	RAID-6	RAID-10	RAID-0	RAID-5	RAID-6	RAID-10
512B	56.80	5.64	6.15	65.05	91.23	4.75	4.66	96.87
1KB	99.40	6.56	7.43	94.16	135.97	5.83	5.47	112.72
2KB	162.75	7.76	8.11	155.59	225.75	5.93	5.90	194.39
4KB	246.76	7.89	8.21	280.08	385.79	10.45	9.82	359.69
8KB	347.79	14.40	15.41	502.94	676.47	15.11	15.98	609.86
16KB	504.07	27.10	26.00	1,001.86	1,136.24	25.01	27.22	1,033.73
32KB	1,815.50	55.66	60.19	1,918.73	2,189.80	56.54	55.22	1,658.03
64KB	3,371.48	122.42	105.89	1,982.09	3,347.99	106.10	90.07	1,694.05
128KB	3,988.31	337.20	305.65	1,996.28	3,376.74	250.50	232.47	1,693.09
256KB	4,092.33	1,383.29	983.55	2,047.98	3,393.11	2,809.26	2,732.72	1,694.79
512KB	4,095.03	2,330.01	2,915.76	2,054.38	3,296.92	3,107.17	2,959.48	1,671.57
1MB	4,108.22	2,748.92	2,951.50	2,060.05	3,317.53	3,138.46	2,857.16	1,658.74
2MB	4,121.88	2,792.49	2,955.05	2,068.80	3,341.53	3,151.95	2,778.56	1,665.90
REMARKS								

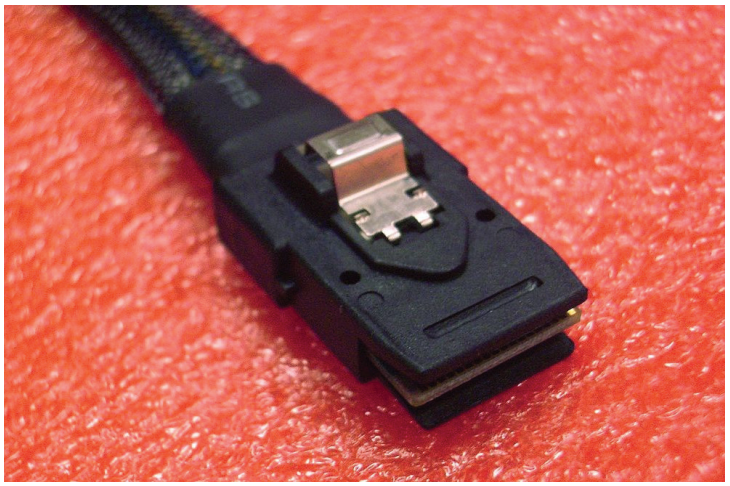
Cable types

AOCs and backplanes connect to each other in a variety of configurations. To understand how to select the correct cable types, you must know what connector-type is required on each side (i.e. AOC being on one side and backplane being on the other side). First let's cover the backplane side. There are three types of backplane nomenclatures that use the letters "TQ" and "A" for DA backplane and "E" for expander backplane. Below I will describe backplane's cable types:

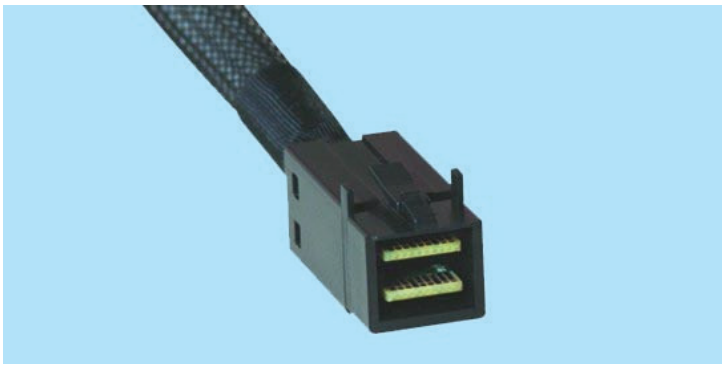
1. TQ backplanes (DA) - uses discrete SATA connectors



2. SAS-2 "A" backplanes (DA) – uses mini SAS ipass connectors (i.e. SFF-8087)



3. SAS-3 "A" backplanes (DA) – uses mini HD connectors (i.e. SFF-8643)



4. SAS-2 Expander backplanes - uses mini SAS ipass connectors (i.e. SFF-8087)

5. SAS-3 Expander backplanes - uses mini SAS HD connectors (i.e. SFF SFF-8643)

Below I will describe AOC's cable types:

1. SAS-2 storage AOCs – uses mini SAS ipass connectors (i.e. SFF-8087)
2. SAS-3 storage AOCs – uses mini SAS HD connectors (i.e. SFF-8643)

Please note that there are external versions of the above cable connectors not mentioned in this white paper.

To determine cable types, we need to know which AOC is connected to which backplane. You can then use the above to determine the correct cable type. Two examples are described below to explain how to do this:

Example #1:

Desired configuration is to connect a SAS-3 AOC to a SAS-2 "A" type DA backplane. That means you need a cable that has SFF-8643 on the AOC side and SFF-8087 on the backplane side. SMCI offers variety of cable length for this. Below are a few:

CBL-SAST-0507-01	Internal Mini-SAS to Mini-SAS HD 80cm w/ SB, 28AWG
CBL-SAST-0508-01	Internal Mini-SAS to Mini-SAS HD 50cm w/ SB, 30AWG

Example #2:

Desired configuration is to connect a SAS-3 AOC to a SAS-3 expander backplane. That means you need a cable that has SFF-8643 on both AOC and backplane side. Again SMCI offers variety of cable length for this. Below are a few:

CBL-SAST-0531	Internal Mini-SAS HD to Mini-SAS HD 80cm,30AWG,12Gb/s
CBL-SAST-0532	Internal Mini-SAS HD to Mini-SAS HD 50cm,30AWG,12Gb/s
CBL-SAST-0550	Internal Mini-SAS HD to Mini-SAS HD 25cm,30AWG,HF,RoHS/REACH

SKUs and Model Names

Item#	Model	Port	AOC Type
1	AOC-S3008L-L8e	8 internal ports, low-profile, 12Gb/s per port-Gen-3, 240HDD	HBA (IT mode)
2	AOC-S3008L-L8i	8 internal ports, low-profile, 12Gb/s per port-Gen-3, 63HDD	RAID 0,1,1E (IR mode)
3	AOC-S3108L-H8iR	8 internal ports, low-profile, 12Gb/s per port- Gen-3, 240HDD / ROC	RAID 0, 1, 5, 6, 10, 50, 60
4	AOC-S2308L-L8e	8 internal ports, low-profile, 6Gb/s per port-Gen-3, 122HDD	HBA (IT mode)
5	AOC-S2308L-L8i	8 internal ports, low-profile, 6Gb/s per port-Gen-3, 63HDD	RAID 0,1,1E (IR mode)
6	AOC-S2208L-H8iR	8 internal ports, low-profile, 6Gb/s per port-Gen-3, 240HDD / ROC	RAID 0, 1, 5, 6, 10, 50, 60
7	AOC-SAS2LP-MV8	8 internal ports, low-profile, 6Gb/s per port-Gen-2	HBA (IT mode)
8	AOC-SAS2LP-H4iR	4 internal & 4 external ports, low-profile, 6Gb/s per port-Gen-2, 240HDD / ROC	RAID 0, 1, 5, 6, 10, 50, 60
9	AOC-SAS2LP-H8iR	8 internal ports, low-profile, 6Gb/s per port-Gen-2, 240HDD / ROC	RAID 0, 1, 5, 6, 10, 50, 60
10	AOC-USAS2LP-H8iR	8 internal ports, low-profile, 6Gb/s per port-Gen-2, 240HDD / ROC	RAID 0, 1, 5, 6, 10, 50, 60
11	AOC-USAS2-L8X	8 internal ports, low-profile, 6Gb/s per port-Gen-2	X=i (RAID 0,1,1E) - IR X=e (HBA) - IT X=iR (RAID 0,1,10,5) – SR

Conclusion

Traditionally Supermicro prides itself in high quality, early development and TTM for its products. Supermicro follows a series of strict ISO approved processes where product development cycles must follow before they are manufactured. A strict approval method is required prior to product release. The processes begin from pre-concept and end at post-mortem. Peer review, strong test procedures, and early product inventory are few steps that are taken before a detail spec is presented on SMCI's website to all potential customers. To assure even higher quality, a final pilot run process is also put in place in Supermicro's production prior to product shipment. We pride ourselves in highest level of product quality and timely delivery by employing the latest technology and complying to strict quality guidelines in allof products.