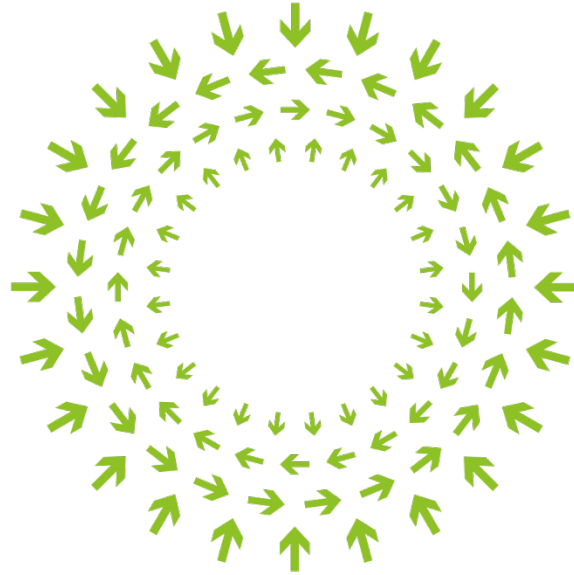




OPEN

Compute Project

Cold Storage Hardware v0.7 ST-draco-abraxas-0.7

Mike Yan, Storage Hardware Design Engineer, Facebook

Song Liu, Storage System Software Engineer, Facebook

1 Contents

1	Contents	2
2	Scope	4
3	Overview.....	4
3.1	License	4
4	Cold Storage Overview	4
4.1	Reference Documents	5
5	Data Center Requirements	5
5.1	Data Center Floor Plan	5
5.2	Networking Topology	6
6	Open Rack Requirements	7
6.1	Rack Power Budgets	7
6.2	Rack Power Zone	7
6.3	Rack Configuration	7
6.4	Mini-SAS Cabling	9
6.5	Rack Mechanical Requirements	10
7	Compute Node Requirements.....	12
8	Storage Unit Design Changes	14
8.1	Hard Disk Drive Selection	14
8.2	Bandwidth Calculation and Discussion.....	14
8.3	Open Vault Mechanical Change	15
8.4	Open Vault Thermal Changes.....	15
8.5	Open Vault FCB Hardware Changes	16
8.6	Open Vault DPB Hardware Changes.....	16
8.7	Open Vault SEB Hardware Changes	17
8.8	Open Vault Firmware Changes.....	17
9	System Software Requirements	19
9.1	HDD Access Mode	19
9.2	HDD Spin Controller	19
9.3	System Monitoring	19

10	Revision History	19
----	------------------------	----

2 Scope

This document discusses the requirements and specifications for implementing a Cold Storage hardware system in a specifically designed cold storage data center, based on the Open Vault storage system design for the Open Compute Project.

3 Overview

When data center design and hardware design move in concert, they can improve efficiency and reduce power consumption. To this end, the Open Compute Project is a set of technologies that reduces energy consumption and cost, increases reliability and choice in the marketplace, and simplifies operations and maintenance. One key objective is openness -- the project is starting with the opening of the specifications and mechanical designs for the major components of a data center, and the efficiency results achieved at facilities using Open Compute technologies.

One component of this project is a Cold Storage server, a high capacity, high storage density, low power consumption, and low cost hardware system for storing data that is accessed very infrequently.

3.1 License

As of April 7, 2011, the following persons or entities have made this Specification available under the Open Web Foundation Final Specification Agreement (OWFa 1.0), which is available at

<http://www.openwebfoundation.org/legal/the-owf-1-0-agreements/owfa-1-0>

Facebook, Inc.

You can review the signed copies of the Open Web Foundation Agreement Version 1.0 for this Specification at <http://opencompute.org/licensing/>, which may also include additional parties to those listed above.

Your use of this Specification may be subject to other third party rights. THIS SPECIFICATION IS PROVIDED "AS IS." The contributors expressly disclaim any warranties (express, implied, or otherwise), including implied warranties of merchantability, non-infringement, fitness for a particular purpose, or title, related to the Specification. The entire risk as to implementing or otherwise using the Specification is assumed by the Specification implementer and user. IN NO EVENT WILL ANY PARTY BE LIABLE TO ANY OTHER PARTY FOR LOST PROFITS OR ANY FORM OF INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER FROM ANY CAUSES OF ACTION OF ANY KIND WITH RESPECT TO THIS SPECIFICATION OR ITS GOVERNING AGREEMENT, WHETHER BASED ON BREACH OF CONTRACT, TORT (INCLUDING NEGLIGENCE), OR OTHERWISE, AND WHETHER OR NOT THE OTHER PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

4 Cold Storage Overview

As the requirements for cold data -- the data that's stored on disk but almost never read again, like legal data or backups of third copies of data -- keep on increasing dramatically,

there are huge demands for developing some form of Cold Storage system with the highest capacity and the lowest cost.

Cold Storage is designed as a bulk load fast archive. The typical use case is a series of sequential writes, but random reads. A Shingled Magnetic Recording (SMR) HDD with spin-down capability could be the most suitable and cost-effective technology for Cold Storage. But any HDD that comes at lowest purchase cost, with hardware system to power off the unused drives, to save power consumption thus greatly reducing operating costs, will also meet the target of lowest TCO for Cold Storage. To accommodate this special use case, a separate infrastructure dedicated for Cold Storage needs to be designed and deployed.

A Cold Storage system design comprises, but is not limited to, the following aspects:

- Ability to adopt current and future HDD technologies with the lowest cost
- Capability to power off HDDs that are not in use
- Modification of storage unit (based on Open Vault)
- Configuration of an OCP compute node
- Mini-SAS fan-out cable between the Open Vault and the OCP compute node
- Custom Open Rack for the configuration of the Cold Storage system
- Redefined topology for networking switch deployment
- New power consumption provisioning, and new data center floor plan, and so forth.

4.1 Reference Documents

The Cold Storage system is a revised version of an OCP data center and hardware, but also utilizes the following specifications, which can be found on the Open Compute Project website:

- Open Rack Hardware v1.0, Sept 18, 2012
(<http://opencompute.org/projects/open-rack/>)
- Open Rack Design Guide v1.0, Jan 19, 2013
(<http://opencompute.org/projects/open-rack/>)
- Intel Motherboard Hardware v2.0, Apr 11, 2012
(<http://opencompute.org/projects/motherboard-design/>)
- Intel Server for Open Rack v0.3, Jan 16, 2013
(<http://opencompute.org/projects/motherboard-design/>)
- Open Vault Storage Hardware v0.7, Jan 16, 2013
(<http://opencompute.org/projects/storage/>)

5 Data Center Requirements

5.1 Data Center Floor Plan

The floor layout and power requirements vary according to the total quantity of racks in the data center.

5.1.1 Floor Layout for Cold Storage Racks

The following floor plan is designed for one Cold Storage data center suite, including:

- 744 racks
- 31 rows for Cold Storage systems
 - o 30 rows of Cold Storage racks
 - o 1 row of protected racks for accessing servers that serve cold data

- 24 racks per row
- 1 network switch per 3 racks
- 1 power zone / power shelf per rack
- 2 compute nodes per rack
- 16 modified Open Vault storage units per rack

Figure 1 shows the layout:

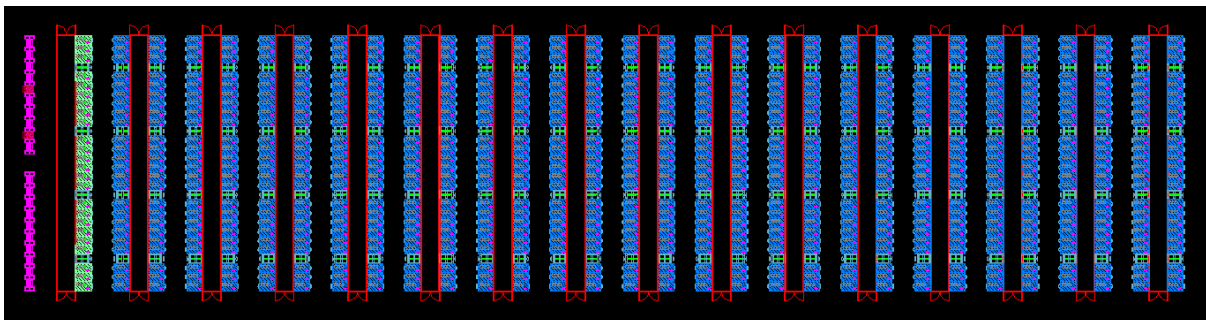


Figure 1 Floor Layout for Cold Storage Data Center Suite

5.1.2 Rack Power Requirements

The total power budget of a data center suite is about 1.3MW to 1.4MW.

- To hold 744 racks, the maximum power budget for each rack is about 1.75kW to 1.88kW on average

5.2 Networking Topology

The data center is connected via a 10G network, but a TOR switch is not deployed in every rack. This is because each rack has only two 10G ports. One area switch is deployed for every three racks, instead. As you can see in Figure 2, only the middle rack has a switch. So the total number of ports in use for downstream networking in every area switch is six.

The switch uses SFP+ passive copper cables for the 10G network.

6 Open Rack Requirements

Estimated power consumption for a Cold Storage rack is:

- ## 6.2 Rack Power Zone

- One power shelf, 3 OpenU height
- Power shelf is located in the middle of the rack
- 5 PSU modules needed to supply 2.8kW maximum with 4+1 capability
- Center bus bar only

Each Cold Storage rack contains 2 compute nodes and 16 Open Vault systems:

-
- 7

- Ratio of compute node to HDD is 1:240

The rack configuration is shown in Figure 3. The 10G switch will be populated according to network topology.

6.3.1 Location of SEB

In order to support SATA HDD in Cold Storage, the SEB in the Knox for Cold Storage system has to be populated on the left side, or A-side, as referred to in the Open Vault storage documents.

Cold Storage Rack, 1:240			
1U	10G Switch		
1U	Empty		
2U	Empty	Winterfell	Empty
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
3U	Power Shelf		
2U	Empty	Winterfell	Empty
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	
2U	SEB 1A SEB 2A	Knox for Cold	

Figure 3 Cold Storage Rack Configuration

6.4 Mini-SAS Cabling

For better fault tolerance and simpler maintenance and service for technician, fan-out cable is preferred other than multiple levels of cascading. Figure 4 shows the Mini-SAS cabling for Cold Storage rack. Figure 5 gives the port mapping of the SAS HBA card and the Mini-SAS cable.

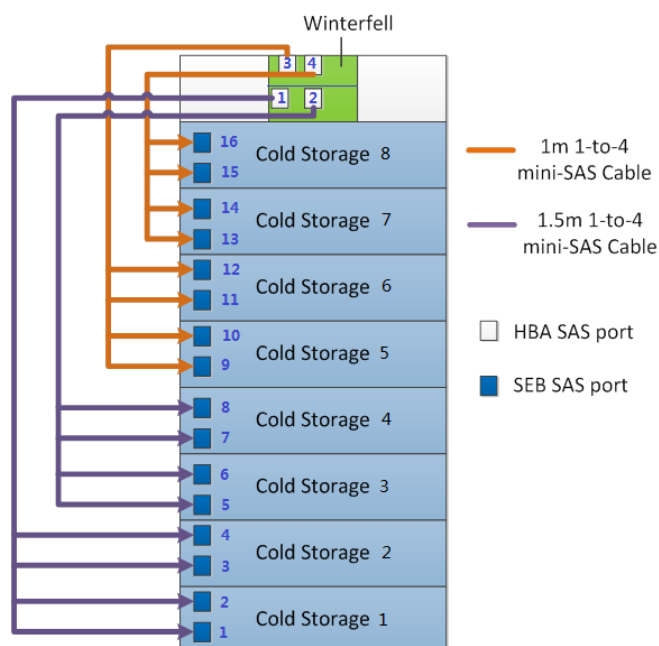


Figure 4 Mini-SAS Cabling for Cold Storage Rack

HBA 2	Port 3	(Cable 3-1)	Port 4	(Cable 4-1)
HBA 1	Port 1	(Cable 1-1)	Port 2	(Cable 2-1)
Cold Storage 8	Tray 16	Cable 4-5		
	Tray 15	Cable 4-4		
Cold Storage 7	Tray 14	Cable 4-3		
	Tray 13	Cable 4-2		
Cold Storage 6	Tray 12	Cable 3-5		
	Tray 11	Cable 3-4		
Cold Storage 5	Tray 10	Cable 3-3		
	Tray 9	Cable 3-2		
Cold Storage 4	Tray 8	Cable 2-5		
	Tray 7	Cable 2-4		
Cold Storage 3	Tray 6	Cable 2-3		
	Tray 5	Cable 2-2		
Cold Storage 2	Tray 4	Cable 1-5		
	Tray 3	Cable 1-4		
Cold Storage 1	Tray 2	Cable 1-3		
	Tray 1	Cable 1-2		

Figure 5 Mini-SAS Port Mapping for Cold Storage Server

6.5 Rack Mechanical Requirements

6.5.1 Rack Height

The Cold Storage rack is the same height as a standard Open Rack.

The total height of the rack is 41 OpenU.

6.5.2 Weight of a Fully Populated Rack

The Cold Storage rack is heavier than a standard Open Rack. The weights of various rack components are as follows:

- Each Open Vault with 4TB HDD: 50 kg
- Each compute node: 23 kg
- Each power shelf with 5 PSU modules: 27 kg
- Cables and switches: 20 kg
- Rack itself: 230 kg
- Total weight: 1,123 kg

6.5.3 Modification on Rack

To support the increased weight of the Cold Storage rack, the supporting mechanisms on both sides of the rack are enhanced. This support mechanism uses the space of the two fans on both sides. It is aligned with the modified Open Vault chassis, which does not include Fan 1 and Fan 6. Additionally it is noted that these fans are not required due to low power consumption of the system. By keeping the compute node in the middle position, the rack needs only the center bus bar. This simplifies the design and reduces the cost.



Figure 6 Custom Open Rack for Cold Storage

7 Compute Node Requirements

The compute node is an Open Compute Project Intel motherboard, v2.0.

- Motherboard:
 - o CPU: x2, Sandy Bridge 2.2 GHz (95W)
 - o RAM: 144GB (2 x 16GB + 14 x 8GB)
- System boot drive:
 - o 2TB SATA
- NIC:

- o 10G network interface card, Mellanox CX3, single port
- HBA card:
 - o Two dual-port SAS HBA card for each server. From LSI: SAS 9207-8e (Silicon D1, Supporting PCI-E Gen3)
- Mini-SAS fan-out cable:
 - o Each server has 16 SAS lanes from the two dual-port HBA cards; using four mini-SAS fan-out cables enables each server to connect to 8 Open Vaults (16 trays)
 - o Each mini-SAS fan-out cable provides links from 1 port SFF-8088 with 4 lanes to 4 ports SFF-8088 with 1 lane in each
 - o Recommended cable lengths: 1.5 meters for cable assemblies 1 and 2 (Cold Storage system 1 to 4). 1.0 meter for cable assemblies 3 and 4 (Cold Storage system 5 to 8). Please refer to Figure 4.
 - o The following images show some details of the mini-SAS fan-out cable:



Figure 7 Mini-SAS Fan-out Cable

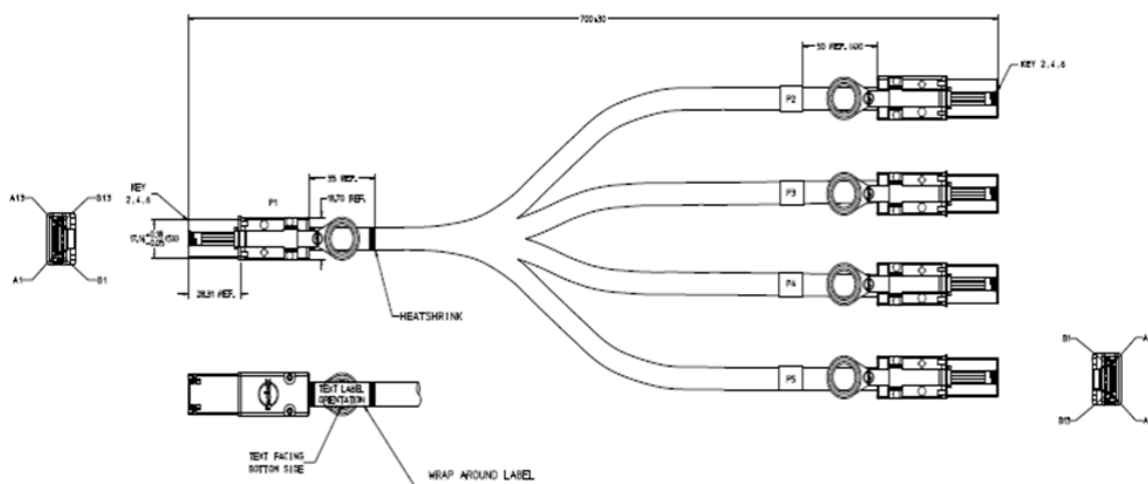


Figure 8 Drawing of Mini-SAS Fan-out Cable

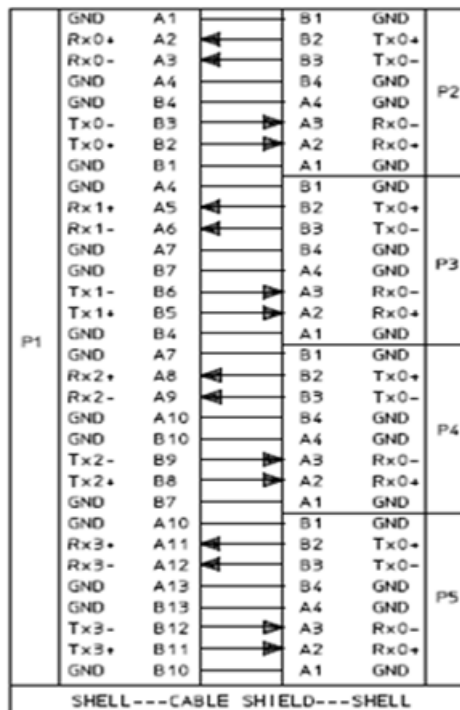


Figure 9 Pin-out of Mini-SAS Fan-out Cable

8 Storage Unit Design Changes

8.1 Hard Disk Drive Selection

To achieve low cost and high capacity, SMR hard disk drives will be supported in the Cold Storage system. Also desktop-level HDDs or archive HDDs are acceptable. To reduce power consumption, only one drive of the 15 on an Open Vault tray will be active at a given time. This will also reduce the vibration to adjacent HDDs and is deemed a good thing for reliability.

- Drive types: SMR, Desktop, Archive, or any LBA based drive that offers the lowest potential TCO
- Interface: SATA, 3G and 6G
- Capacity: 4TB and larger

8.2 Bandwidth Calculation and Discussion

Ideally on average, one HDD will provide 80 MB/s bandwidth. Therefore, if $80 \times 16 = 1,280$ MB/s, the aggregated bandwidth to a compute node is about 1.3 GB/s.

This bandwidth can serve the “backup” needs for cold storage in normal use cases. But for “restore” needs, higher bandwidth will be expected. In the case that a little bit more power consumption is allowed, up to three HDDs can be powered on in each tray. In theory this gives much higher bandwidth to a compute node so it won’t be bottlenecked by storage unit: $80 \times 3 \times 16 = 3,840$ MB/s, or about 3.9 GB/s.

Rack power consumption for this case will be:

- Storage unit (Open Vault with 6 HDD spinning): 120W

- Compute node: 300W
- Network switch: 200W
- Power budget without network switch: $120 \times 16 + 300 \times 2 = 2,520W$
- Power budget with network switch: $120 \times 16 + 300 \times 2 + 200 = 2,720W$
- Power budget for every three racks with one network switch: $2,520 \times 2 + 2,720 = 7,760W$

8.3 Open Vault Mechanical Change

The Open Vault chassis mechanical design must be modified to block the open areas of fan 1 and fan 6 to prevent air turbulence. The lower case needs a flexible tooling that can deal with either 6 fan slots for normal storage or 4 fan slots for Cold Storage.

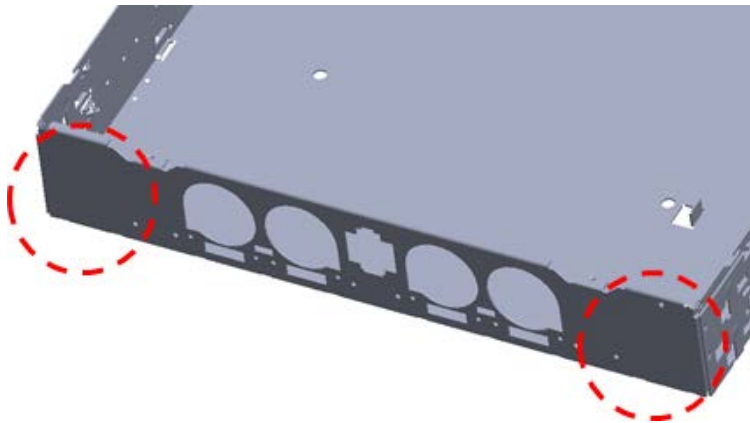


Figure 10 Blocking the Fan Opening

8.4 Open Vault Thermal Changes

One simple fan control strategy utilizes the same fan model and control logic as a normal storage Open Vault system. The fan control algorithm in the Open Vault firmware is used, and the only modification is to import a new fan curve.

In the typical use case, only one HDD is powered on at a time in each HDD tray, so at most only two HDDs will be powered on in a 2xOpenU Open Vault system at any given time. All reads and writes in the tray go to those two drives. As with most regular PWM controlled fan models, the current Open Vault fan has the minimum fan speed and power consumption at 30% of the PWM level. It cannot go below this limit. If four fans are running together at this level, the system becomes overcooled and fan power consumption is higher than needed.

To achieve more efficient fan control, a low power fan model is used for Cold Storage. With circuit modifications by fan vendors, this low power fan can ramp down all the way to 10% of PWM.

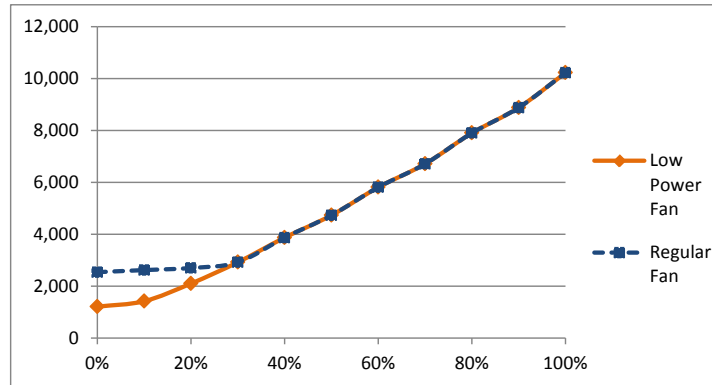


Figure 11 Example PWM and Fan Speed for a Low Power Fan

8.5 Open Vault FCB Hardware Changes

8.5.1 Hardware Versioning for FCB

In order to identify different hardware versions, there is a GPI pin for firmware to check during initialization: 0 for normal storage and 1 for Cold Storage, as Figure 12 shows. R357 or R360 will be populated accordingly.

Knox	FCB HW Rev	R357 (Pull-up)	R360 (Pull-down)
Normal Storage	0	No-POP	POP
Cold Storage	1	POP	No-POP

Figure 12 FCB Hardware Version Control

8.5.2 BOM Variance for Four Fans

To align with the Open Rack for Cold Storage, fan 1 and fan 6 are removed. Thus, the Open Vault for Cold Storage has only four fans (fan 2 to fan 5). As a result, the fan 1 and fan 6 connectors are depopulated from the FCB.



Figure 13 BOM Variance for Four Fans

8.6 Open Vault DPB Hardware Changes

8.6.1 Hardware Versioning for DPB

In order to identify different hardware versions, there is a GPI pin for the firmware to check during initialization: 0 for normal storage and 1 for Cold Storage, as Figure 14 shows. R406 or R407 will be populated accordingly.

Knox	DPB HW Rev	R406 (Pull-up)	R407 (Pull-down)
Normal Storage	0	No-POP	POP
Cold Storage	1	POP	No-POP

Figure 14 DPB Hardware Version Control

8.6.2 HDD Power Control Logics Change

To keep all HDDs at power-off status without inrush current, when the system is initially powered on, the Open Vault DPB has a logic change on the power control circuitry. Details are shown in Figure 15. Related AND gates are changed to NAND gates: U3, U6, U9, U12, U15, U18, U21, U24, U27, U30, U33, U36, U39, U42 and U45. Due to this change, the Open Vault DPB for normal storage is different from DPB for cold storage.

SEB A HDD Power Control	SEB B HDD Power Control	HDD Power Control on DPB for Normal Storage	HDD Power Control on DPB for Cold Storage
0	0	OFF	ON
0	1	OFF	ON
1	0	OFF	ON
1	1	ON	OFF

Figure 15 HDD Power Control Logics Change

8.7 Open Vault SEB Hardware Changes

In current cold storage architecture, fan-out is preferred than SAS expansion. So on the SEB, the two internal Mini-SAS connectors have been depopulated. Figure 16 shows the details.

Along with this change, the SAS expander chip has been changed from 28 ports to 20 ports for further cost savings. The part number of this new low-port-count expander IC is LSISAS2x20. It is from LSI, and is pin-to-pin compatible with the original part (LSISAS2x28).



Figure 16 SEB Hardware Change

8.8 Open Vault Firmware Changes

Changes required for the Open Vault firmware are listed below.

8.8.1 Adapt to 20 Ports SAS Expander

The SEB firmware has been modified so that it works with the new 20 ports SAS expander (LSISAS2x20).

8.8.2 Hardware Revision Detection

The SEB firmware, for both the DPB revision and the FCB revision, needs to detect Hardware Revision pins before entering Cold Storage mode.

8.8.3 Fan Module Status Reporting

Since Fan 1 and Fan 6 in are not included in the Open Vault Cold Storage system, the SEB firmware is modified accordingly to handle this case; it doesn't report an error about the absence of Fan 1 and Fan 6.

8.8.4 Fan Control Strategy

The firmware for the fan control in an Open Vault Cold Storage system is implemented as follows:

- The firmware signature in the FCB EEPROM indicates the Unified Fan Control circuits and algorithm.
- A new fan curve has been implemented based on thermal results tested with a low-power fan as described in Section 8.4.

8.8.5 HDD Spin-up and Power Control

When powering up an Open Vault Cold Storage system, the firmware needs to ensure that all 15 HDDs on one tray are not spun up at the same time. This is to avoid a big inrush current or unnecessary power consumption. This special spin-up control mechanism is achieved by keeping all HDD power rails set to "off" status when the system is initialized.

8.8.6 SATA Data Rate Selection and Discussion

Open Vault storage has been designed for 6G SAS. It meets all signal integrity specifications. For SATA 6G, there are two channels that cannot meet the specifications, HDD 9 and HDD 14. To keep all 15 channels consistent, current Cold Storage SEB firmware fixes the SATA data rate for all HDDs at 3G.

In case there's need for higher throughput from Cold Storage system, it will be allowed for the user to set auto negotiate, allow channels to support 6G.

8.8.7 Signal Integrity Related Parameter Setting

For all expander SAS channels, need to fine-tune MDIO setting based on signal integrity testing results on final SMR drive at 3G SATA data rate.

For expander channel 7 and 8 that go into SAS/SATA signal re-driver, also may fine-tune the parameter setting according to final SMR drive electrical characteristics.

8.8.8 Mini-SAS Port Link Status

Since there is only 1 SAS link from the head node to each Open Vault tray (SEB), the firmware on the Open Vault Cold Storage unit must be modified as follows for the mini-SAS port link status:

- With x1 SAS link in mini-SAS port, do not turn on the red LED for fault indication; keep the mini-SAS port link status LED blue for normal operation.
- No error code for "Mini-SAS Loss of Link".
- No event log for "Mini-SAS Link Error".

Mini-SAS Port Link Status	Blue LED	Red LED
SAS Links (x1 ~ x4) Health	ON	OFF
No SAS Links	OFF	OFF

Figure 17 Mini-SAS Port Status LED Definition for Cold Storage

9 System Software Requirements

9.1 HDD Access Mode

- Check with drive before powering off the drive, make sure it finishes internal housekeeping tasks.
- Multiple drives are powered up in parallel most of the time across systems.
- Most reads/writes are to large files, and mostly sequential.
- Application reorders outstanding requests to maintain queue depth of one request to each HDD, at any given time.

9.2 HDD Spin Controller

A HDD spin controller ensures that there is only one active disk in a tray.

- Power-on a specific HDD before accessing it. It may need to spin-up if has been recently spun-down.
- Power-off a specific HDD after it finishes access.
- Within an Open Vault system, power-on the two HDDs in the same slot on both trays.

9.3 System Monitoring

Because SATA was adopted for the HDD interface, the expander firmware in Open Vault cannot support S.M.A.R.T. information inquiry, therefore the upper level utilities or application software is required to support HDD S.M.A.R.T. information polling.

10 Revision History

Version	Changes
0.5	Initial public version.
0.6	• Expand HDD selection to desktop and archive drives. • Add Rack cabling and port mapping. • Add DPB and SEB changes. • Add discussion for power on up to three drives. • Add discussion for SATA 6G supporting.

0.7	Incorporated wider team review feedbacks. Edited for clarity and grammar.
-----	--