ARM Server Motherboard Design for Open Vault Chassis Hardware v0.3 MB-draco-hesperides-0.3

Author: Tom Schnell, Calxeda
tom@calxeda.com

1 Scope

This document defines the technical specifications for the Calxeda-based ARM motherboard used in the Open Vault storage enclosure for Open Compute Project servers.

2 Contents

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 an ARM-based server board that converts the Open Vault JBOD into a storage server. This specification assumes a Calxeda-based ARM solution that replaces the SAS expander board within the Open Vault chassis (red board shown below).

Figure 1 Open Vault

3.1 License

4 Server Board Features

4.1 Block Diagram

Figure 2 illustrates the functional block diagram of the ARM Server Board, using a Calxeda ECX-1000 SOC. Key points:

1) Supports 15 drives through SATA multipliers. This design is intended for cold storage where SATA bandwidth requirements are low.
2) Supports 4GB system memory. This is the maximum for a 32-bit CPU. Other CPUs can support higher memory capacities at a higher cost of power. A 32-bit CPU is deemed sufficient for cold storage.
3) The network interface can be differentiated based on bandwidth requirements and rack-level cabling preferences.

A "node" consists of the ARM SOC processor core, system memory, and local peripherals needed to boot a Linux operating system and run an application. Nodes are NOT coherent, rather, they are independent servers connected primarily through the network fabric.

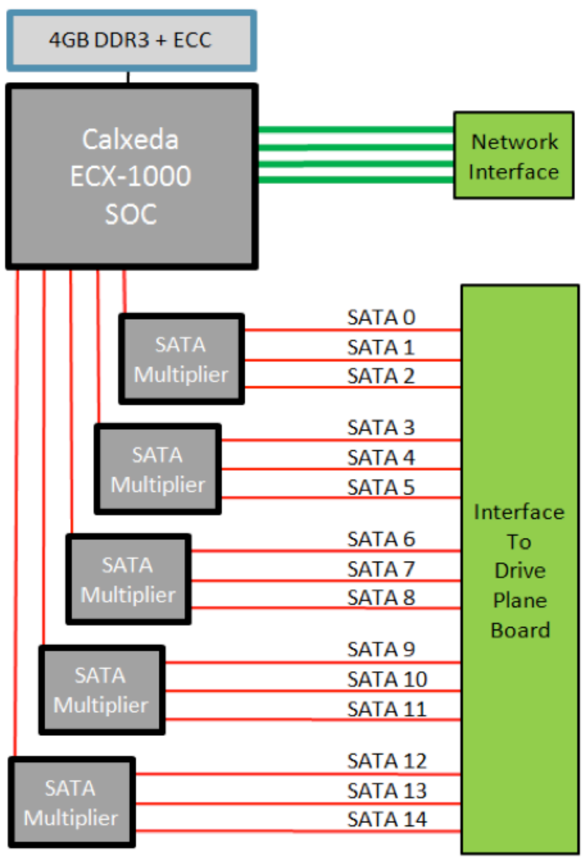

Figure 2 ARM Server Board Block Diagram

4.2 Network Options

4.2.1 Dual 1GbE Ports

The dual 1GbE port option provides the lowest cost network interface. The ECX1000 will provide SGMII interfaces to two 1Gbps PHY chips that drive the external cable interface through RJ45 connectors.

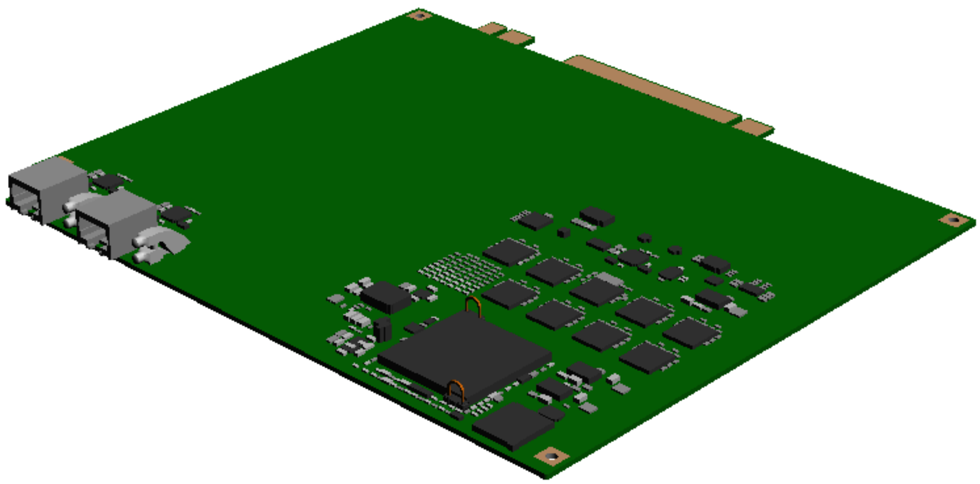

Figure 3 ARM Server Board with 2x 1Gb Ethernet Ports

Dual ports allow daisy-chained rack cabling or redundant TOR switches.

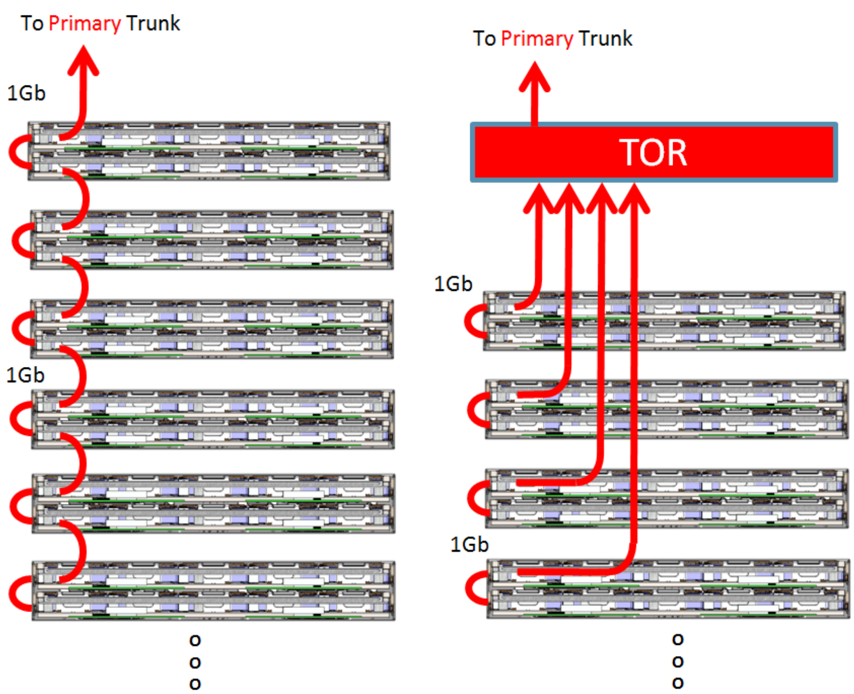

Figure 4 Networking Layout with 2x 1Gb Ethernet Ports

4.2.2 Quad CX4 Ports

Quad CX4 ports provide the lowest cost 10GbE network interface. The ECX1000 will provide XAUI interfaces to four CX4 connectors that drive the external cable interface.

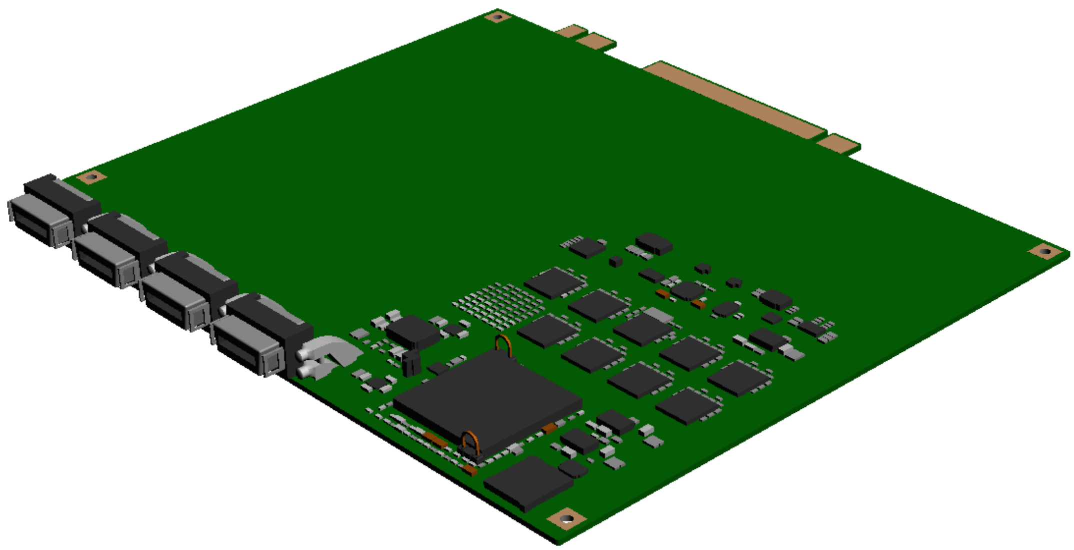

Figure 5 ARM Server Board with 4x CX4 Ports

The benefit of four 10Gbps ports are chassis-to-chassis connectivity within a rack and greater configurability of TOR bandwidth. Examples of rack cabling show these options.

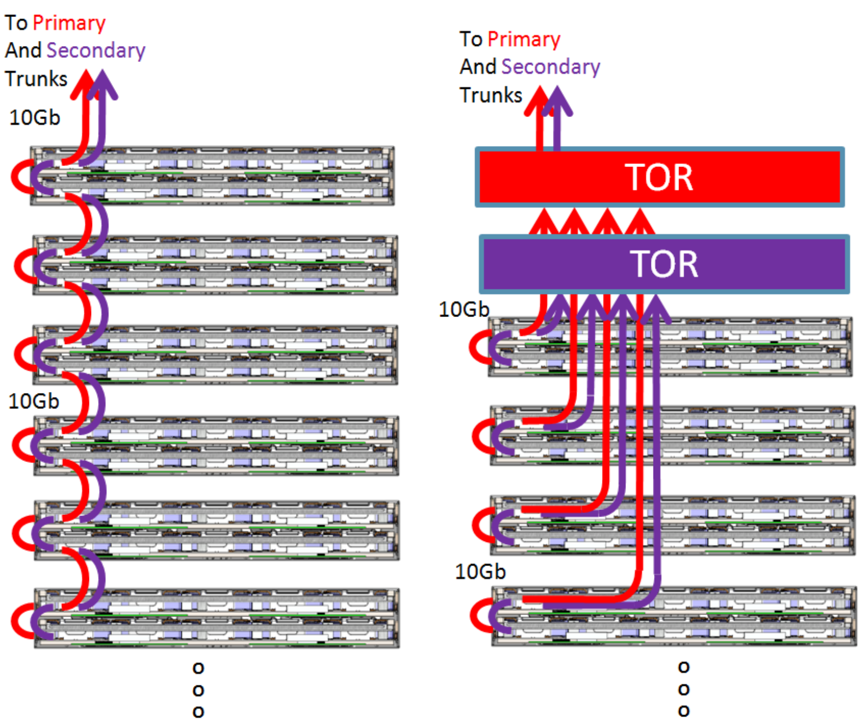

Figure 6 Networking Layout with 4x CX4 Ports

4.2.3 Quad QSFP Ports

Quad QSFP ports provide similar benefits as CX4, adding cost to gain more flexibility in physical media type and cabling. Note that each QSFP provides one 10GbE link (as opposed to 40GbE provided by QSFP+ which is not supported in this connector.)

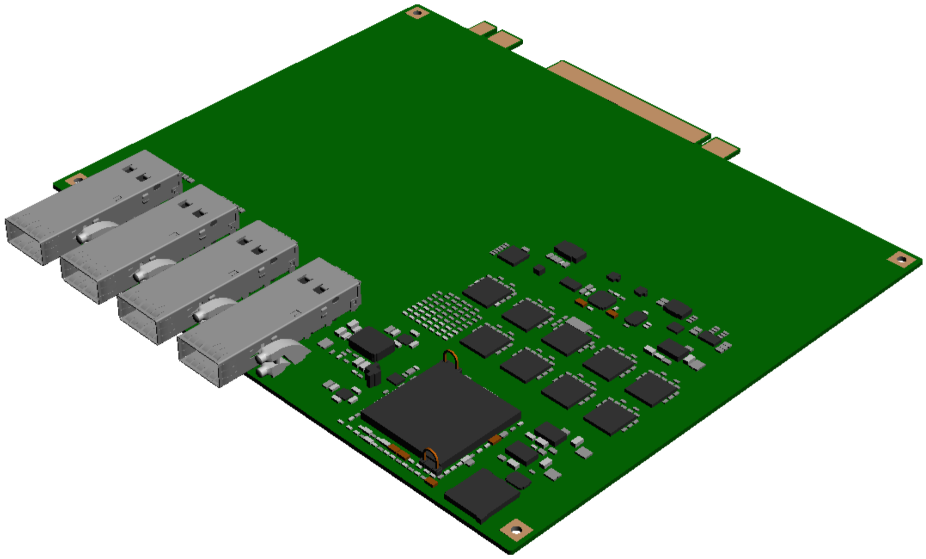

Figure 7 ARM Server Board with 4x QSFP Ports

4.2.4 Quad SFP+ Ports

Quad SFP+ ports provide similar benefits as CX4, adding cost to gain more flexibility in physical media type and cabling. A 10GbE PHY chip is required to convert XAUI to XFI to be driven through the installed SFP+ module.

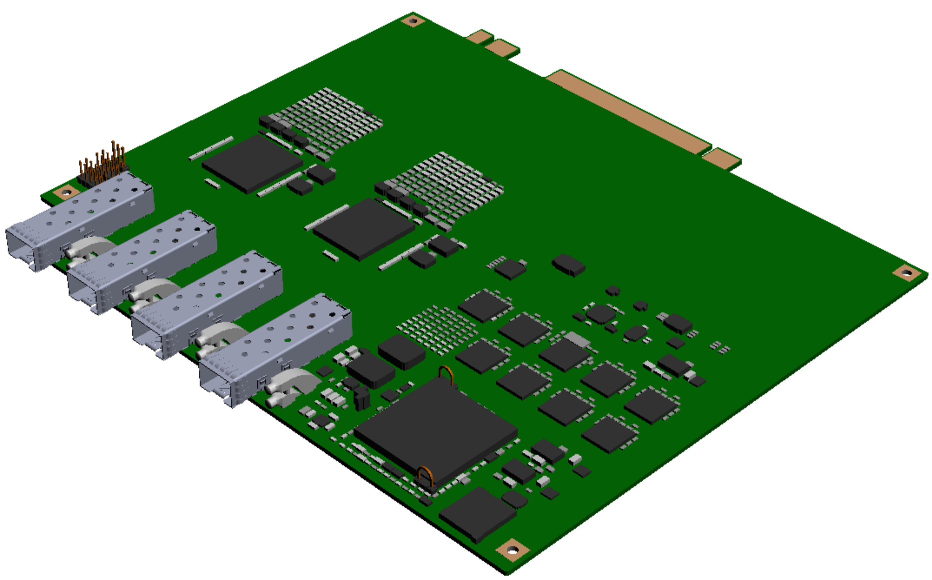

Figure 8 ARM Server Board with 4x SFP+ Ports

5 Node Details

From the block diagram in Figure 1, the SOC can be viewed generically as a processor complex with various quantities of SATA and Ethernet interfaces. This processor complex boots and runs an operating system, like Linux, which allows it to become a network-accessible device. On top of Linux, various software applications can execute, according to the target application. Therefore, the details of the SOC block can be vendor-specific.

For a Calxeda-based solution, Calxeda provides the underlying firmware that provides power management, fabric management, and IPMI support. The ODM is responsible for programming the physical cards in production.

5.1 Linux

The Calxeda EnergyCore ECX-1000 SOC requires running Linux kernel v3.2+ and distributions such as Ubuntu Server 12.04 LTS based on that version. The ODM is responsible for production-level installs and OS support.

5.2 PXE Boot

The Uboot firmware supports PXE boot. When PXE booting, the system first attempts to boot from the first Ethernet interface (eth0). If a PXE boot on the first Ethernet interface fails, the firmware attempts to PXE boot from the second Ethernet interface (eth1).

5.3 Other Boot Options

The firmware also supports booting from SATA. The Uboot firmware provides the capability to select boot options.

5.4 Remote Firmware Update

The firmware can be updated remotely under these scenarios:

- Scenario 1: Sample/Audit firmware settings
    - Return current firmware settings, or
    - Save/export firmware settings in a human-readable form that can be restored/imported (as in scenario 2)
- Scenario 2: Update firmware with pre-configured set of firmware settings
    - Update/change multiple firmware settings
    - Reboot
- Scenario 3: firmware/firmware update with a new revision
    - Load new firmware/firmware on machine and update, retaining current firmware settings
    - Reboot

Additionally, the update tools have the following capabilities:

- Update from the operating system over the LAN – the Linux OS standards are RedHat and Ubuntu Linux.
- Can be scripted for no user interaction (like prompts)
- Can be scripted and propagated to multiple machines

5.5 Event Log

The ODM must provide a system access interface and application software to retrieve and clear the event log from the firmware, including, at minimum, a Linux application for the Ubuntu operating system and driver as needed. The event log must be retrieved and stored as a readable text file that is easy to handle by a scripting language under Linux. Each event record includes enhanced information identifying the error source device's vendor ID and device ID. Calxeda supports IPMI commands that provide access to the Event Log.

6 Hardware Monitoring

The motherboard does not employ a traditional out of band monitoring solution. The ODM needs to provide a system access interface and application to retrieve hardware monitoring sensor readings. Lm_sensors is the preferred tool for hardware monitoring under Linux; the ODM ensures Lm_sensors works. The sensors to be read include voltage, temperature, and fan speed.

6.1 Thermal Sensors

The motherboard has several thermal sensors:

- Each CPU must have a temperature sensor
- Inlet temperature, and located in the front of the motherboard
- Outlet temperature, and located in the rear of the motherboard

The sensors should make sure that no CPU throttling is triggered due to thermal issues, under the following environmental conditions:

- Inlet temperature lower than 30C (including 30C), and 0 inch H2O pressure
- Inlet temperature higher than 30C but lower than 35C (including 35C), and 0.01 inch H2O pressure

The sensors should make sure that the total airflow rate for the chassis is lower than 89CFM, including PSU.

In the event that one fan fails, an inlet temperature of 30C with 0 inch H2O pressure environment is used to verify thermal sensors.

6.2 Fan Control Algorithm

The ODM must provide an optimized fan control algorithm based on the thermal solution of the system including fan, heat sink, and air duct. Fan speed control should set system fans running at lowest speed and provide enough damping to avoid speed vibration.

7 Power System

7.1 Input Voltage

7.1.1 Input Voltage Level

The nominal input voltage delivered by the power supply is 12.5VDC. The motherboard can accept and operate normally with an input voltage tolerance range between 10.8V and 13.2V. The motherboard's under-voltage protection level is 10V or less.

7.1.2 Capacitive Load

To ensure compatibility with the system power supply, the motherboard cannot have a capacitive load greater than 4000µF. The capacitive load of the motherboard cannot exceed the maximum value of 4000µF under any operating condition listed in Environmental Requirements.

7.2 Hard Drive Power

The motherboard does NOT supply power to the system's 15 hard drives. T

7.3 System VRM Efficiency

The ODM supplies high efficiency VRMs for all other voltage regulators over 20W not defined in this specification. All voltage regulation modules over 20W have 91% efficiency over the 30% to 90% load range.

7.4 Power On

The motherboard powers on upon application of power to the input connector. The use of a power button is not required. The motherboard always resumes operation upon restoration of power in a power failure event.

8 I/O System

8.1 Network

For the SFP+ network configuration, the motherboard has a Vitesse VSC8488 dual Ethernet transceiver that interfaces to the 10Gb SFP+ cages. It has an SMbus connection to the SOC.

The firmware supports PXE boot on all network ports on the motherboard.

8.2 Reboot on WOL in S0 State

Reboot on WOL (ROW) is a feature that repurposes the traditional Wake on LAN (WOL) signal to reboot the motherboard. While the system is in S0 state (running), when a WOL packet is received, the wakeup signal causes a hardware reboot of the motherboard. ROW does not require the power supply to cycle its output.

8.3 SATA

The motherboard supports 3Gbps SATA channels.

9 Mechanical

9.1 PCB Thickness

To ensure proper alignment of the edge connectors and mounting within the mechanical enclosure, the board must have 85mil (2.16mm) thickness.

9.2 Silkscreen

The silkscreen is white in color and includes labels for these components:

- CPU numbers
- Ethernet port numbers
- [LEDs]
- [Switches]

10 Environmental Requirements

The motherboard meets the following environmental requirements:

- Gaseous Contamination: Severity Level G1 per ANSI/ISA 71.04-1985
- Ambient operating temperature range: -5°C to +45°C
- Operating and storage relative humidity: 10% to 90% (non-condensing)
- Storage temperature range: -40°C to +70°C
- Transportation temperature range: -55°C to +85°C (short-term storage)

The full OpenVault system also meets these requirements. In addition, the full system has an operating altitude with no de-ratings of 1000m (3300 feet).

10.1 Vibration and Shock

The motherboard meets shock and vibration requirements according to the following IEC specifications: IEC78-2-(*) and IEC721-3-(*) Standard & Levels. The testing requirements are listed in.

	Operating	Non-Operating
Vibration	0.5g acceleration, 1.5mm amplitude, 5 to 500 Hz, 10 sweeps at 1 octave / minute for each of the three axes (one sweep is 5 to 500 to 5 Hz)	1g acceleration, 3mm amplitude, 5 to 500 Hz, 10 sweeps at 1 octave / minute for each of the three axes (one sweep is 5 to 500 to 5 Hz)
Shock	6g, half-sine 11mS, 5 shocks for each of the three axes	12g, half-sine 11mS, 10 shocks for each of the three axes

Figure 9 Vibration and Shock Requirements

11 Prescribed Materials

11.1 Disallowed Components

The following components are not used in the design of the motherboard:

- Components disallowed by the European Union's Restriction of Hazardous Substances Directive (RoHS 6)
- Trimmers and/or potentiometers
- Dip switches

11.2 Capacitors and Inductors

The following limitations apply to the use of capacitors:

- Only aluminum organic polymer capacitors made by high quality manufacturers are used; they must be rated 105C
- All capacitors have a predicted life of at least 50,000 hours at 45C inlet air temperature, under worst conditions
- Tantalum capacitors are forbidden
- SMT ceramic capacitors with case size > 1206 are forbidden (size 1206 are still allowed when installed far from the PCB edge and with a correct orientation that minimizes risks of cracks)
- Ceramic material for SMT capacitors must be X7R or better material (COG or NP0 type should be used in critical portions of the design)

Only SMT inductors may be used. The use of through hole inductors is disallowed.

11.3 Component De-rating

For inductors, capacitors, and FETs, de-rating analysis should be based on at least 20% de-rating.