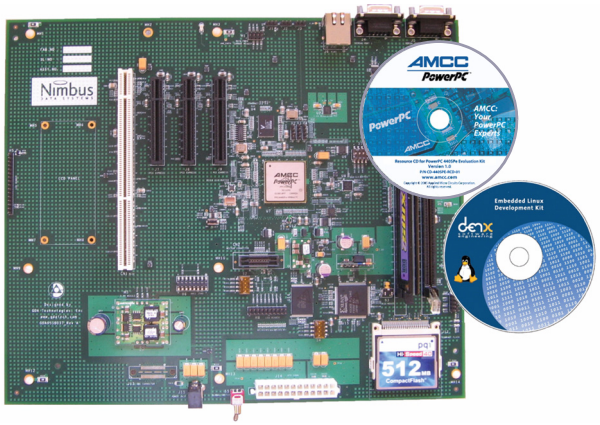
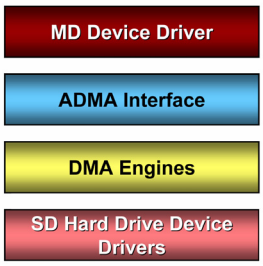
High Performance Linux RAID Stack for AMCC Storage Processors

Optimized RAID stack leverages the performance of on-chip DMA engines. Industry-standard interfaces support off-the-shelf Linux middleware and drivers for I/O controller cards, accelerating development time for storage systems.



Product Highlights

- Complete high-performance RAID stack using standard Linux based components
- Stack comprises LVM, MDADM, and hardware-accelerated MD driver
- Supports RAID 0, 1, 5, 6, 10, 50
- RAID 5 and 6 computations performed in PowerPC 440SP/SPe DMA engines
- MD driver uses standard ADMA interface to access on-chip DMA engines
- MD driver enhancements minimize impact of Linux overhead on system performance
- Supports rapid incorporation of new I/O controller cards provided with Linux device drivers
- Software bundle includes drivers for 1Gbps/10Gbps Ethernet and SAS/SATA I/O controllers
- Low cost 440SPe platform (Katmai) facilitates evaluation of complete software bundle
- Katmai board provides 3 PCIe, one PCI-X, 1 Gbps Ethernet, and 2 serial ports
- Cross development tools available for Linux and Windows hosts
- System level demonstrations provided by AMCC on request



BUILT ON Power™

Product Overview

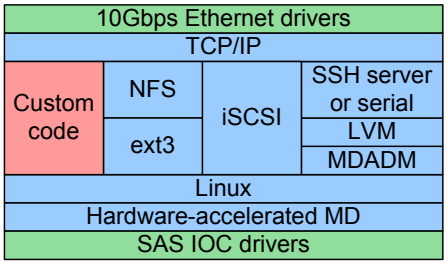
AMCC and DENX Software Engineering have architected a complete high-performance RAID Stack based on standard Linux components. The stack runs on the AMCC PowerPC 440SP and 440SPe storage processors. The stack comprises LVM (Logical Volume Manager), MDADM (MD driver management utility), and a hardware accelerated MD driver. The stack supports RAID levels 0, 1, 5, 6, 10, and 50. RAID 5 and 6 computations are performed using on-chip DMA engines. The MD driver accesses the DMA based hardware accelerators via the standard ADMA interface. In addition to hardware acceleration, the MD driver is enhanced to improve system throughput. The MD driver enhancements minimize the impact of Linux overhead on system performance.

The solution minimizes time to market for OEMs and ODMs. For example, introducing a new I/O controller only requires inserting its Linux driver. The software bundle includes drivers for 1Gbps/10Gbps Ethernet and SAS/SATA I/O controllers.

The software bundle can be downloaded from DENX free of charge by following the appropriate link at AMCC web site (***http://www.amcc.com/Embedded/Downloads/440SPe***). The low cost 440SPe platform (Katmai) facilitates evaluation of the complete software bundle

The bundle includes cross development tools for code development and debug. Industry-standard benchmarks are supplied to support customers' performance analysis of the RAID stack. These include Iometer which can be run either from an iSCSI initiator or locally via a TCP connection, as well as xdd which can be run locally.

AMCC Linux RAID Bundle



DENX provides free of charge the Linux bundle that includes the components shown in the figure. This bundle has been tested on AMCC PowerPC 440SPe evaluation board (Katmai). The bundle was created using the DENX cross development tools.

The components of the bundle are:

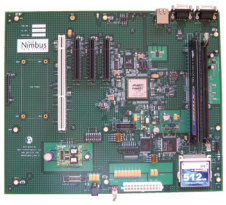
- Linux 2.6.21 and associated U-Boot for the Katmai board.
- Modified Linux MD driver for RAID. The modifications include the recent standard ADMA interface implementing the async_tx_api to asynchronously submit tasks to DMA engines. PowerPC 440SPe specific code implements the async_tx_api functions in the storage processor's RAID specific DMA engines. These functions include RAID 5 and 6 computations and memory copy.
- MDADM RAID utility
- Linux Logical Volume Manager (LVM)
- Drivers for SAS I/O controllers from major vendors
- Drivers for 10 Gbps Ethernet controllers from industry-leading suppliers
- iSCSI server
- Ext3 file system
- Network File System (NFS) server

Development Tools

The Katmai evaluation kit includes a CD from DENX. This CD contains the ELDK (Embedded Linux Software Development Kit). It includes cross development tools and a Linux host based NFS file system for use either with the customer's target system or with the Katmai board. It also includes full source code of U-Boot, Linux kernel trees for the RAID bundle, and a small Linux root file system image for standalone operations.

Also available from DENX is a Windows-hosted (coLinux) version of ELDK. This provides a Linux shell within a Windows environment as well as well as the complete ELDK described above. Both CDs may be downloaded directly from the DENX web site.

440SPe Katmai Evaluation Board



With an approximately 10" x 12" ATX form factor, the Katmai board is a custom-designed platform developed by Nimbus Data Systems for evaluating the 440SPe processor. For more details please download the Katmai evaluation board product brief from the AMCC web site (***http://www.amcc.com/Embedded/Downloads/440SPe***).

Availability

System level demonstrations are provided by AMCC on request. The demonstration system comprises a 2U enclosure complete with 12 Seagate 250GB SATA disk drives, a Katmai board with 1GB of SDRAM, the Linux RAID bundle described above, 2 LSI SAS I/O controllers, 10Gbps Ethernet + TCP/IP controllers, optical or CX4 cables. These units are ready to connect to a network for testing and evaluation. Please consult your local AMCC sales office for more details.

Linux Asychronous DMA (ADMA) Interface

This interface is a recent development. Its purpose is to provide a standard interface to DMA engines. The MD driver in Linux implements a RAID device. It exposes a number of hard disk drives (e.g /dev/sd[a-h]) as a single drive (e.g. /dev/md0). The hard disk drives form a RAID set. The MD driver computes the parity and Galois Field operations synchronously in software. This means that during the computation the processor is dedicated to this task. In addition running code for performing block computations is inefficient especially for RAID 6. This software-based RAID 5 operation typically achieves 30 MBps of sequential write performance, although the DMA engines within the PowerPC 440 SPe are capable of over 600 MBps of RAID 5 performance.

At this time the ADMA interface provides the following functions:

- async_memcpy
- async_memset
- async_xor

These functions are exposed to the layer above. The ADMA interface splits the MD driver into two. The layer above is control functions and the layer below is data flow using processor DMA engines. AMCC and DENX are extending this interface to include RAID 6 Galois Field computations.

The layer below is also a standard interface. The function there is device_tx_submit. The implementation of this function is processor specific. AMCC processors support multiple methods of implementing RAID called WXOR, RXOR, and DMA 2.

The word asynchronous refers to the way the RAID computations are now decoupled from the control path. The DMA engines cause an interrupt when the task is completed. This causes a callback to the calling routine that can then update its control structures.

AMCC enhancements to Linux RAID

The development of ADMA was a significant step forward in improving the performance of RAID in Linux, decoupling it from data flow. This will impact many other functions such as security, network processing, iSCSI and others. As Linux is confined to control functions, not processing data, it will become an RTOS as well as a general purpose OS where middleware can run simultaneously.

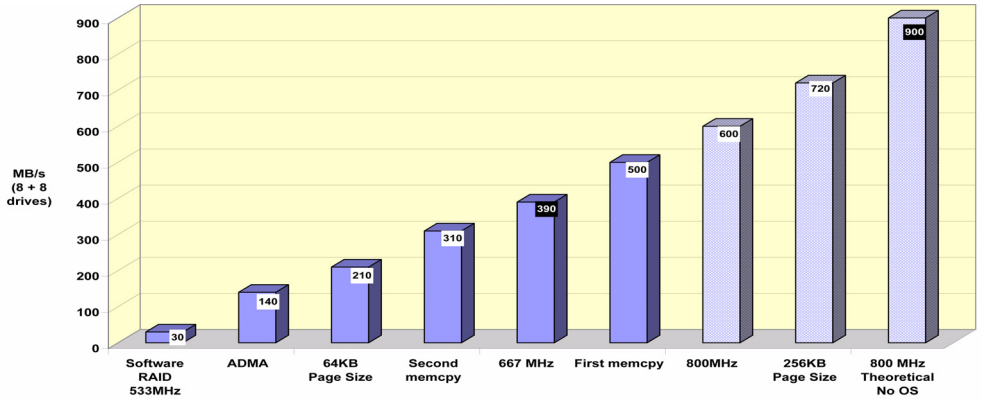
For RAID, the hardware acceleration was not enough. It provided a performance improvement of 3x. AMCC and DENX are collaborating to improve the MD driver. To date two improvements have been identified and implemented for significant benefit. These changes to the MD driver are a part of the Linux bundle that is available from DENX.

The first improvement is to make the page size a flexible quantity. 4KB, 16KB, and 64KB page sizes are supported with a roadmap to 256 KB page sizes. The hard disk drives perform progressively better for sequential writes when the I/O size is increased. RTOS based RAID firmware developers know this and concatenate data in large chunks. The MD driver uses the page size of embedded Linux as the I/O size. There is a ratio of 6 to 1 in hard disk write performance for 64KB and 4KB page sizes. There is no loss of generality from extending the page size. If the appliance is used as an NFS server the local file system can have block size smaller than the page size. If the appliance is an iSCSI target the reassembly of iSCSI packets creates the local page size. If the appliance is directly connected via PCIe the Linux driver on the server will gather the I/O into local page size chunks.

The second improvement is to remove a memory copy that is internal to the MD driver. The MD driver stages strip data ready to be written next to the I/O controller in a page size pre-allocated buffer. It is possible to bypass this memory copy for sequential writes thereby saving SDRAM access cycles.

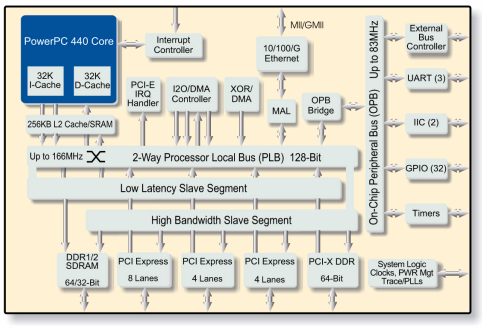
Performance Results and Roadmap

With the improvements described above RAID 5 performance is at 500 MBps for a 667 MHz system and two 8-drive arrays. The software is being ported to an 800 MHz 440SPe platform. With this and two other improvements (256 KB page size and another memory removal) performance is expected to reach 720 MBps for RAID 5 sequential writes. Further study of the MD driver for other possible improvements is in progress. The implementation of RAID 6 is also in progress. The figure below summarizes the series of performance enhancements. AMCC has a uniquely rich storage processor roadmap and future devices will leverage these same RAID enhancements.



Possible Applications

- iSCSI or NFS file servers with RAID
- CAS (content addressed storage) systems
- Direct attached RAID enclosures
- Search engines



PowerPC 440SPe Block Diagram

AMCC

Corporate address:
215 Moffett Park Drive
Sunnyvale, CA 94089
Tel: 408-542-8600
Fax: 408-542-8601
www.amcc.com