# AIX Kernel CAPI Flash Support

## Introduction

AIX 7.2 introduces a CAPI (Coherently Accelerator Processor Interface) infrastructure and support for CAPI attached flash storage. AIX 7.2 provides a framework to view CAPI attached devices as a new coherent accelerator device class. The initial CAPI flash solution attaches an IBM Flash Systems model 900 with a CAPI adapter.

CAPI flash allows application I/O requests to be serviced in user-mode with a new AIX CAPI block library. The user-mode library greatly reduces I/O overhead, but requires application changes for exploitation.

CAPI flash can also be exploited with traditional storage APIs (AIO, filesystem, LVM). AIX includes an optimized kernel CAPI flash adapter driver. The kernel device driver exploits CAPI technology to reduce kernel path length and enhance I/O scalability. This article provides details about the AIX kernel CAPI flash driver including requirements, implementation details, and performance.

## Requirements

- IBM Flash Systems model 900
- CAPI Flash adapter: low profile feature code EJ17, high profile feature code EJ18
- AIX 7.2 TL0 SP1
- Firmware 840
- HMC V8 R840
- Power Systems S814, S822, and S824
    - CAPI activation feature code EC2A
- Power Systems E870 and E880
    - CAPI activation feature codes EC18, EC19

CAPI flash only supports direct attachment of the IBM Flash Systems model 900. Up to 16 8Gb FC ports can be ordered with the Flash Systems 900. Flash LUNs must be configured with 4K sector size. LUNs must be zoned to either FC HBA WWPNs (World Wide Port Names) or CAP Flash HBA WWPNs, since AIX will not support sharing/MPIO between traditional FC stack and CAPI flash stack. AIX does support multipathing for CAPI flash disks, so LUNs can be cabled and zoned to multiple CAPI flash adapter ports.

CAPI flash adapters can only be assigned to a LPAR in dedicated mode. Virtualization with the VIOS is not supported. Features requiring virtual I/O support such a Logical Partition Migration and Live Kernel Updated are not supported with CAPI flash.

## Implementation Details

CAPI provides a new efficient method to attach accelerators to a system. Each Power8 chip contains a CAPP (coherent accelerator processor proxy) that enables accelerators to participate in the system coherency protocols. IBM provides a PSL (power service layer) that reside in an FPGA along with an accelerator functional unit (AFU). The PSL and CAPP units provide an accelerator with coherent memory access using a host virtual address. The result is a reduction in software overhead required to communicate with an accelerator.

AIX manages CAPI attached accelerators with a coherent accelerator subsystem. AIX configuration assigns a CAPI slot a unique bus (capi0, capi1) and a special file (/dev/capi0, /dev/capi1). CAPI flash adapters are configured as children of the CAPI bus with a special file /dev/cflashX and an ODM subclass of “capi”. The CAPI flash driver functions as a typical storage adapter in the AIX storage stack.  CAPI flash disks are configured as hdisks.  MPIO commands such as lspath/lsmpio support CAPI attached disks.  AIX diagnostics including error log analysis and concurrent firmware download are also supported.

Significant optimizations have been made to the kernel CAPI framework and the kernel CAPI flash driver.  The kernel CAPI bus driver DMA setup services exploit CAPI virtual memory addressing.  Traditional PCI adapters requires a hypervisor TCE (translation control entry) to be setup for each page in an I/O operation.  With CAPI, the system virtual address can be used avoiding the per page I/O translation path length.  The significant per page I/O overhead is avoided providing more efficient I/O interface.

The CAPI flash kernel stack(Figure 1) also contains additional optimizations.  First it implements multiple-queues and multiple locking domains for enhanced scalability.  A CAPI flash adapter is able to maintain up to 508  contexts.  Each context contains an address space, work queue, and interrupt resources providing independent channels to issue I/O requests.  The kernel driver uses 8 (default, configurable up to 32) contexts to allow 8 separate kernel I/O channels.  Queues are assigned to hdisks allowing I/Os from different disks to be processed in parallel.  The remaining CAPI contexts are used for the user-mode block library.  Kernel driver I/O and user-mode I/O to be issued independently.  The kernel multi-context/multi-queue design allows good scaling on today’s large multi-core/multi-thread configurations.  A second optimization combines the fibre channel protocol module and adapter module into a single kernel module.  A single module improves efficiencies further reducing path length.

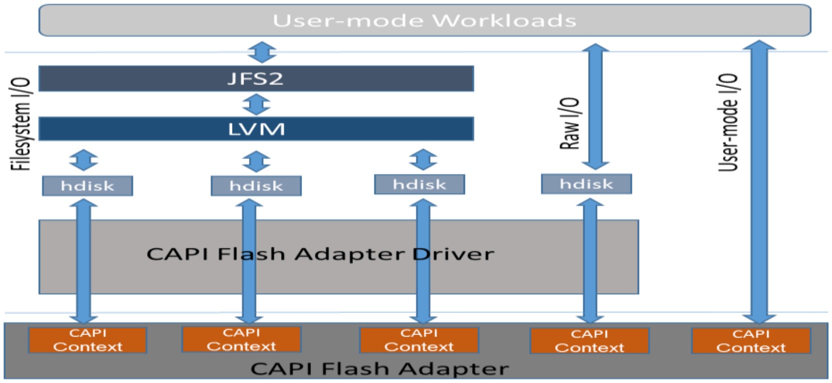

Figure 1

### Performance Comparison

To demonstrate enhanced scalability and efficiency in the kernel CAPI flash I/O stack an IBM FlashSystem 900 with 12 x 2.9 TB Flash modules was directly attached to a Power S824.  Two traditional fibre channel adapters are used as baseline results to compare against two CAPI flash adapters. Each adapter (fibre channel/CAPI) used 2 - 8Gb ports.  The AIX LPAR contained 24 dedicated cores and 256 GB memory.

The first test measured instructions to drive a single synchronous I/O using the AIX raw device interface.  The raw device interface bypasses AIX filesystem and kernel disk caching thus isolates the AIX storage stack.  A graphical comparison is shown in Figure 2.

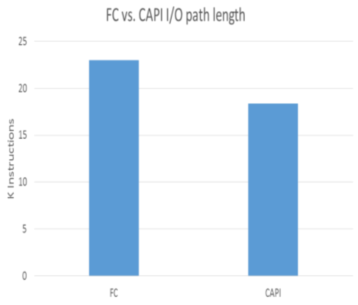

Figure 2

The instructions required to drive a single synchronous CAPI I/O are reduced by 25% compared to the fibre channel baseline.  CAPI flash path length improvements are due to reduced I/O setup time and CAPI interface efficiencies.  Path length reduction applies to all AIX kernel storage I/O methods including filesystem and AIX asynchronous I/O.

The second test compares maximum IOPs (I/O operations) between fibre channel and CAPI.  Internal I/O benchmarks are used to simulate workload I/O stress.  Figure 3 shows Maximum 4K IOP throughput achieved, while Figure 4 shows IOPs per core.

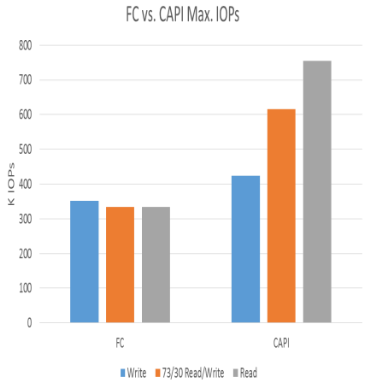

Figure 3

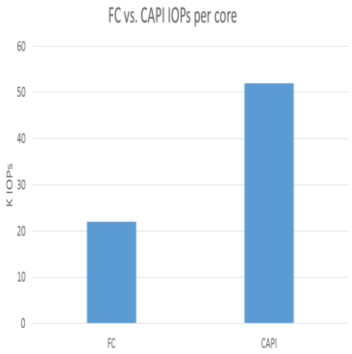

Figure 4

Figure 3 shows that the CAPI flash stack is able to deliver a significantly higher maximum IOP rate than current AIX fibre channel adapters.  CAPI flash is able to deliver 113% improvement in maximum read IOPs.  CAPI flash write IOPs are improved by 20%.  A 70%/30% mixture of read/write IOPs is used to show simulate a common application load and achieves an 84% improvement.  These large improvements are primarily due to enhanced scaling.  The multi-queue/multi-CAPI-context design used by the CAPI flash kernel driver reduces kernel lock contention on large core count / high IOP tests.  The path length reductions shown in Figure 2 further improve scalability by reducing lock hold time.  Reduced lock contention and pathlength improvements provide better per core IOP rates as seen in Figure 4.  At maximum IOP rates the CAPI flash per core IOP rate is improved 150%.

The kernel CAPI flash stack provides significant improvements in IOP scalability and efficiency when compared to current AIX fibre channel stack.  A much larger improvement can be achieved by exploiting user-mode library APIs.  A user-mode I/O benchmark run against the configuration shown in Figures 4 improves per core IOPs by over 12 times compared to the fibre channel baseline. The user-mode APIs require application adoption, but the kernel CAPI flash driver is able to deliver efficiencies with traditional I/O APIs benefiting existing I/O intensive workloads.

## References

CAPI flash user-mode APIs:

http://www-01.ibm.com/support/knowledgecenter/ssw_aix_72/com.ibm.aix.capi/capi_flash_adapter.htm?lang=en

PCIe3 CAPI Fibre Channel Flash Accelerator Adapters:

http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hcd/fcej17.htm
http://www-01.ibm.com/support/knowledgecenter/POWER8/p8hcd/fcej18.htm