

A Lightweight Layered Compressed File System with Hardware Acceleration

**Shirish H. Phatak
Altior, Inc.**

- Altior Introduction
- Market Requirements and Design Goals
- CeDeFS Design and Implementation
- Future Directions
- Q&A

- **Altior Introduction**
- Market Requirements and Design Goals
- CeDeFS Design and Implementation
- Future Directions
- Q&A

Altior Corporate Overview

● Well Capitalized Start-UP

- Founded in 2004
- Privately held, VC backed
- Headquarters in New Jersey
- Design Centers of Excellence in Texas, and in India
- Sales & Biz Development in California
- All Design, Manufacturing, and Support in the USA



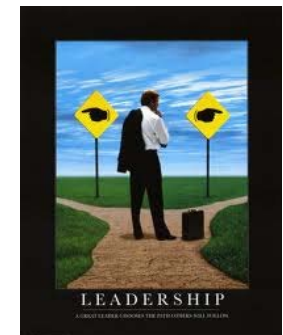
● Engineering Through Capital

- 40 Engineers - Most of the Engineers have Masters in EE or CS
- Average Work Experience - 15 Years
- Domain Expertise - Networking, Storage, & Wireless



● World-Class Leadership

- Management:
 - Ramana Jampala (President & CEO)
 - Rob Zecha (VP of Engineering/Operations)
- Board of Directors:
 - Dr. Adam Drobot (Ex-President & CTO, Telcordia Tech)
 - Morton Meyerson (*Ex President & Vice Chairman, EDS Ex Chairman and CEO of Perot Systems*)
 - Ian Trumpower (Financial Analyst)



AltraFlex Accelerator Cards

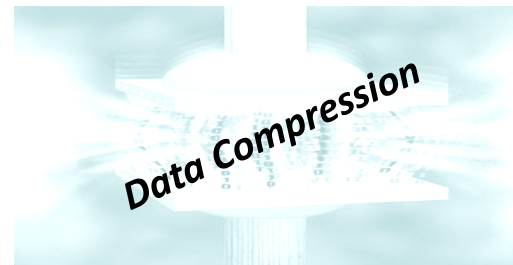
AltraFlex Platform



Specifications

- Altera FPGAs: Arria & Stratix Families
- PCIe Gen 1 x8; Gen 2 Optional
- Half-length, Full-height, OR Small form-factor
- 8 GB RAM; Optional Extension
- 2 Gigabit Ethernet ports (Optional)
- Less than 10W Power Consumption

AltraFlex Acceleration Cards



Throughput	Up to 12 Gbps
Comp Ratio	3:1
SW Stack	File System Agnostic
Drivers	Linux , Windows



Throughput	Up to 12 Gbps
SW Stack	File System Agnostic
Drivers	Linux, Windows



Throughput	Up to 8 Gbps
Dedupe Ratio	Up to 5:1
SW Stack	File System Agnostic
Drivers	Linux , Windows

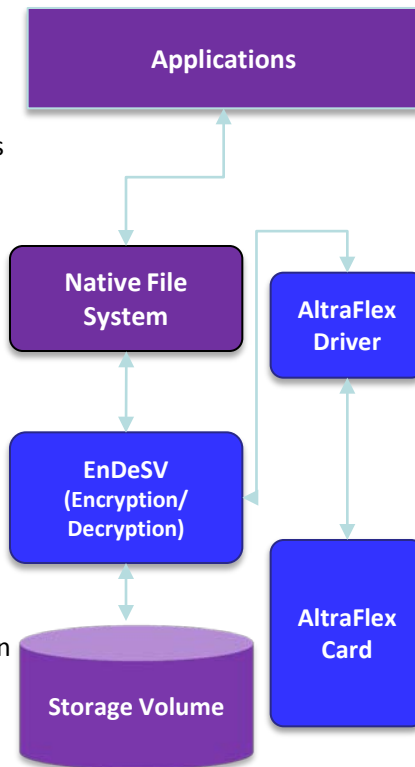
** Deduplication Cards Available in Q4 2011*

Storage Solutions

Data Compression, Encryption & Deduplication Storage

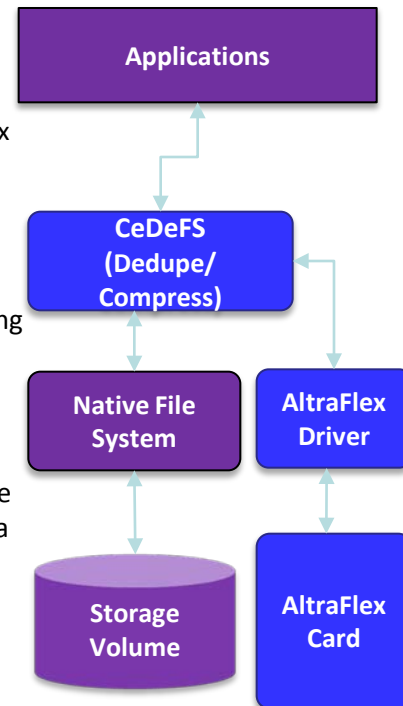
EnDeSV – Encryption Storage

- Full system solution consisting of:
 - AltraFlex hardware accelerator and drivers
 - EnDeSV storage volume filter
- AltraFlex ensures host CPU/memory is not used for Encryption/Decryption
- Standards compliant AES-XTS
- EnDeSV designed to be a light-weight high-performance filter layer
- EnDeSV can layer over any storage volume
- Data is encrypted as it is written to the disk and decrypted as it is read from the disk







CeDeFS – Compression/Dedupe File System

- Full system solution consisting of:
 - AltraFlex hardware accelerator and Linux drivers
 - AltraFlex CeDeFS file system on Linux
- Data is deduped and compressed as it is written to the disk and reassembled as it is read from the disk
- AltraFlex solution does compression, hashing and hash-match/locate functions
- Can be offered as a Compression only solution
- Enhanced solution for conserving disk space using full Dedupe and Compression for data at rest
- CeDeFS designed to be a light-weight high-performance filter layer
- CeDeFS can layer over any standard file system



- Altior Introduction
- **Market Requirements and Design Goals**
- CeDeFS Design and Implementation
- Future Directions
- Q&A

Market Requirements

- ❑ Reduction in amount of raw storage required  Compression File System
- ❑ Minimum impact on existing applications and storage stack  Lightweight
- ❑ Interoperability with existing applications and storage stack  Layered
- ❑ 1+ GBps Performance with Minimum CPU utilization  Hardware Acceleration

Additional Product Requirements

- ❑ Focus on “Captive Storage” or DAS
- ❑ Support for Large Files
- ❑ Efficient Random Access IO
- ❑ Support for Sparse Files
- ❑ Efficient Handling of Uncompressible Data

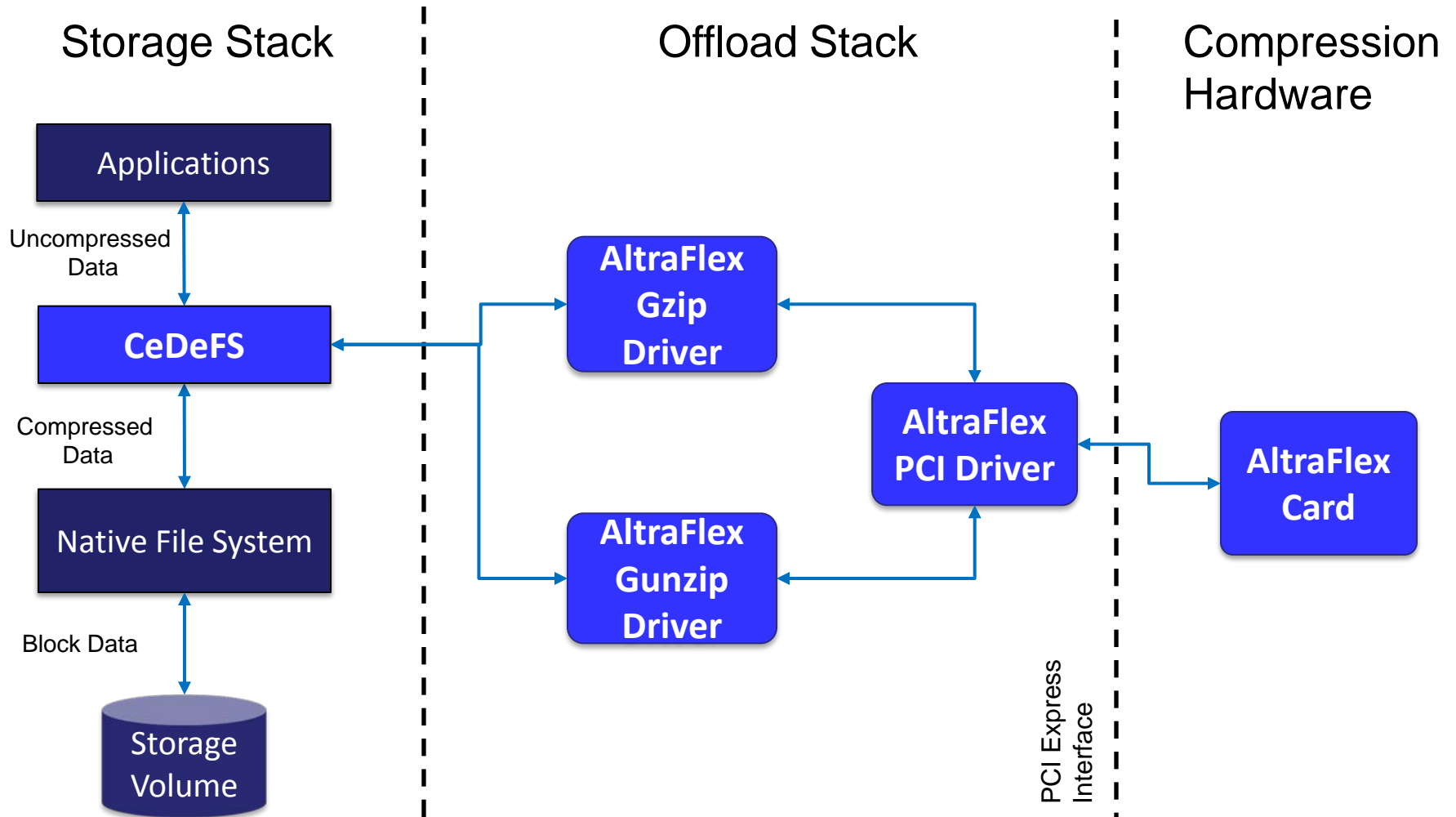
General Value Proposition

- ❑ Minimal Change to Existing Software
- ❑ Reduce Number of Disks/Nodes Required
- ❑ Reduce Power/Cooling Requirements
- ❑ Increase Available Compute Power
- ❑ Increase Data Center Efficiency

- ❑ Big Data
- ❑ Cloud Computing
- ❑ Map/Reduce Type Applications
- ❑ Traditional File Servers and NAS

- Altior Introduction
- Market Requirements and Design Goals
- **CeDeFS Design and Implementation**
- Future Directions
- Q&A

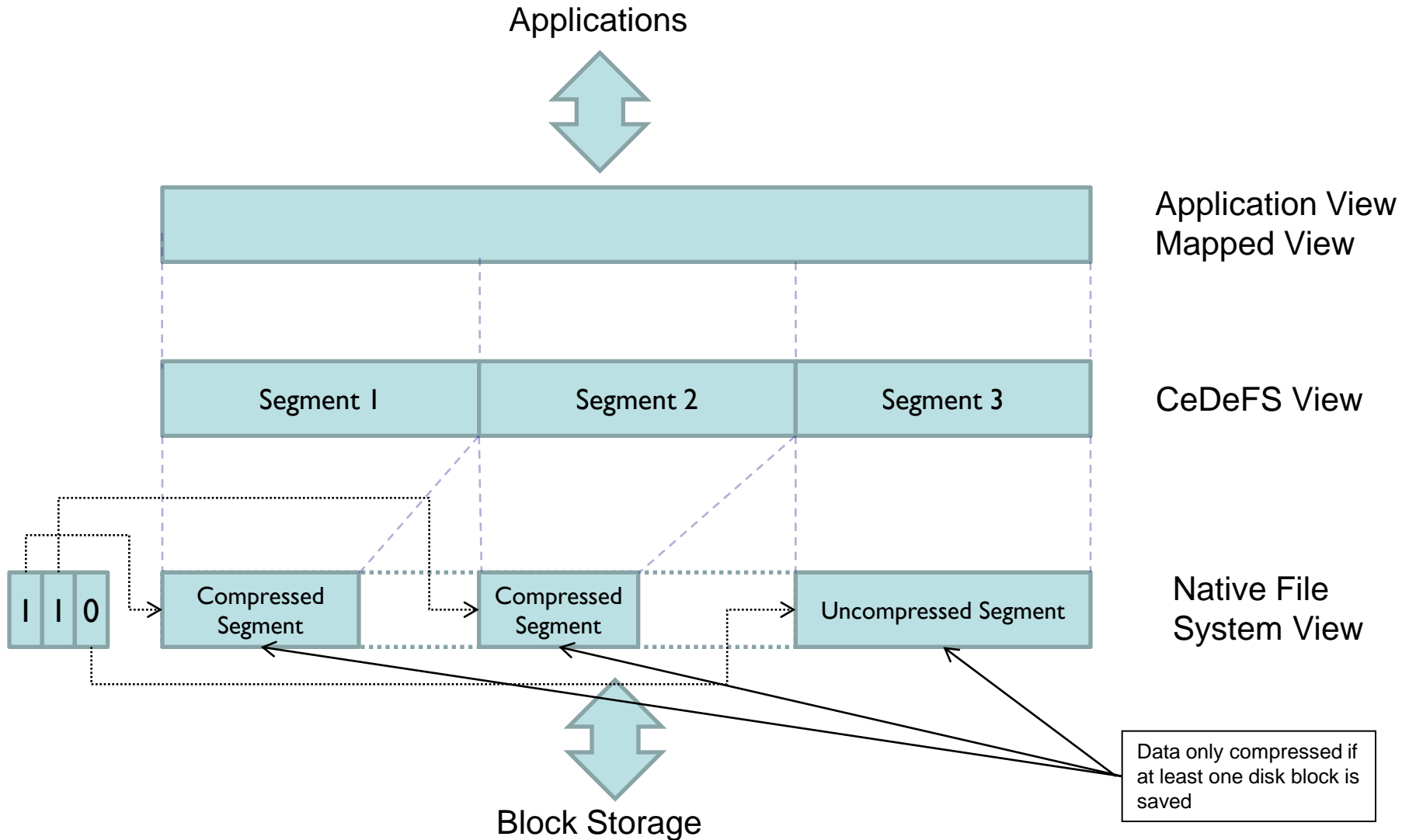
CeDeFS Block Diagram



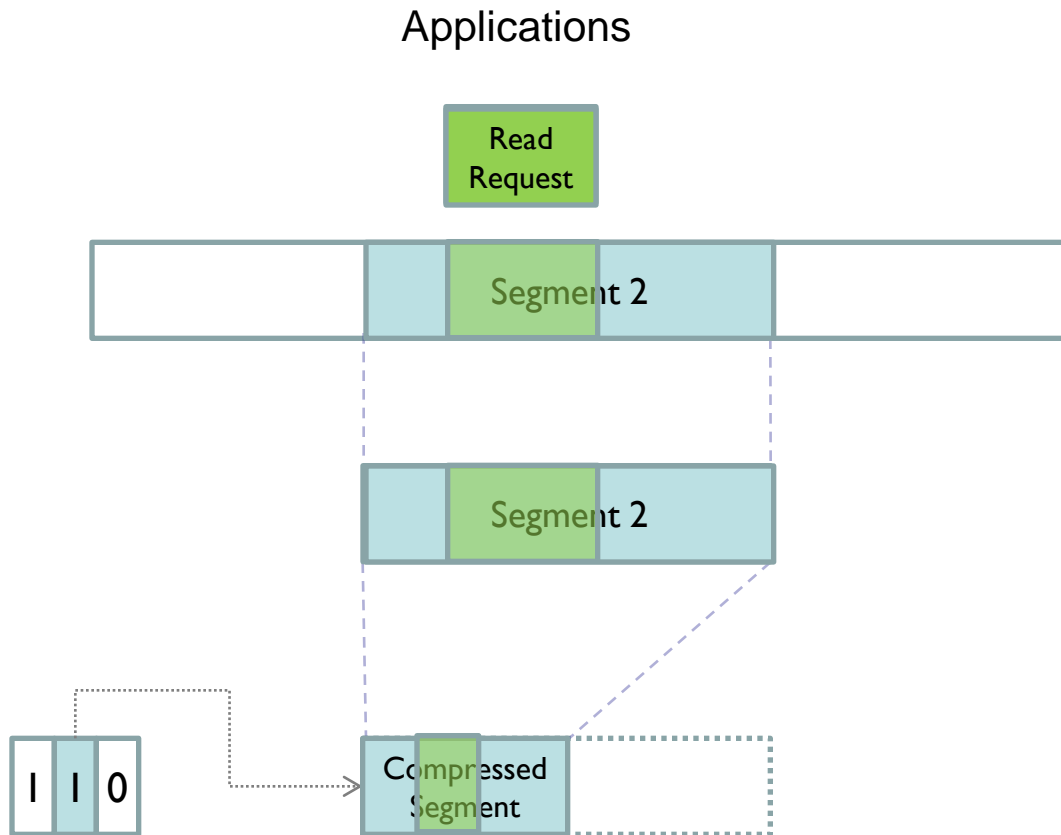
- ❑ Layered File System
 - ❑ Stacked (FSD over FSD) on Linux
 - ❑ Intercepts only read/write IO, all other operations are pass through

- ❑ Read/Write operations performed through mapped IO
 - ❑ Fully integrated into page cache/memory mapped IO
 - ❑ Reads are done via page cache
 - ❑ Uses native page write back (BDI threads) to update lower file system

CeDeFS File Layout



Lifecycle of a Read Request



- ❑ Application issues Read
- ❑ CeDeFS determines corresponding segment(s)
- ❑ CeDeFS reads entire (compressed) segment from native FS
- ❑ CeDeFS decompresses segment and updates mapped view
- ❑ CeDeFS completes Read

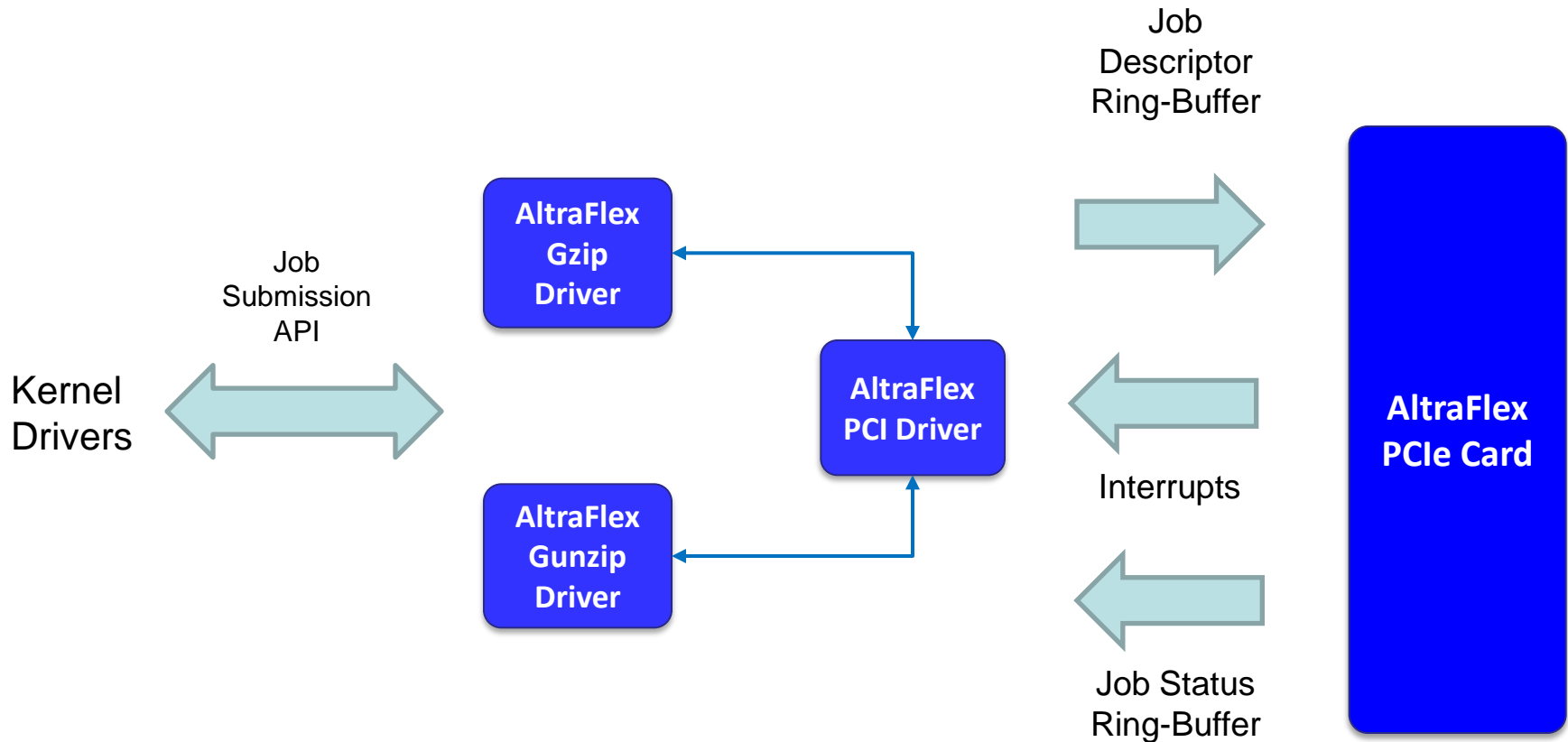
- ❑ Analogous to Read, reverses the operation flow
- ❑ Highly optimized for extended writes and writes that overwrite an entire segment
- ❑ Needs read-modify-write for partial segment updates if data exists in the segment

Hardware Integration: API

- ❑ Software uses a Uniform Transform API
 - ❑ Same generalized API works for AES, GZIP/GUN, SHA, etc.
 - ❑ Driver and hardware internally multiplexes different functions
 - ❑ API is usable from user space (e.g. zlib integration)

- ❑ Each individual data transformation request is a “Job”
 - ❑ Job ID, transform function (AES, GZIP, GUNZIP, etc), input buffer and output buffer length/addresses and asynchronous callback information
 - ❑ Buffers lengths are limited; bigger transform requests can be split across jobs

Hardware Integration: Block Diagram



Performance Considerations

- ❑ Asynchronous transform processing
 - ❑ Hardware introduces latency, so synchronous processing is undesirable; use callback based processing

- ❑ Zero Copy and Scatter Gather
 - ❑ Copying adds to CPU load, to the extent possible directly map input and output pages for DMA

- ❑ Tuning segment sizes
 - ❑ Best results if the segment size fits one job

- ❑ File System Check (FSCK)
 - ❑ Segment and meta-data integrity checks
 - ❑ Online vs Offline

- ❑ Software decompression/compression
 - ❑ For skewed workloads
 - ❑ Decompression in software generally makes sense

- ❑ NFS Integration
 - ❑ File Handle Mappings

- Altior Introduction
- Market Requirements and Design Goals
- CeDeFS Design and Implementation
- **Future Directions**
- Q&A

- ❑ Deduplication
 - ❑ Use on chip SHA engines for hashing
 - ❑ Non-dedupped data can be compressed
 - ❑ Backward compatible with CeDeFS

- ❑ Encryption
 - ❑ Integrated along with compression and de-duplication
 - ❑ Key management

- ❑ Windows
 - ❑ Mini-filter?

- ❑ Network optimization
 - ❑ Run CeDeFS over NFS or SAN
 - ❑ Reduce data center network bandwidth requirements

- ❑ Offloads for Native File Systems
 - ❑ BTRFS
 - ❑ ZFS

- Altior Introduction
- Market Requirements and Design Goals
- CeDeFS Design and Implementation
- Future Directions
- **Q&A**

Q&A

Shirish H. Phatak

shirish.phatak@altior.com

+1-650-516-7471