Accelerating Storage

Darren J. Harkins – Staff Systems Engineer

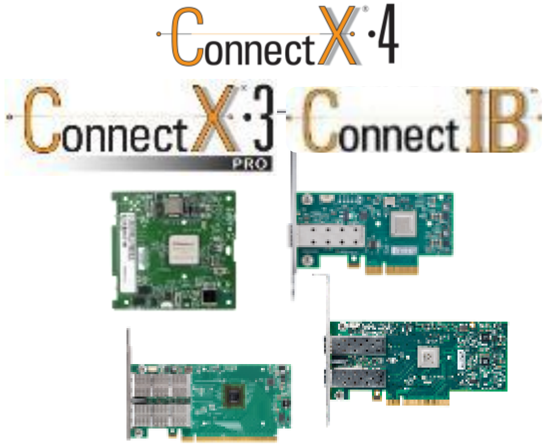
May 2016

Mellanox® TECHNOLOGIES
Connect. Accelerate. Outperform.™

Interconnect Solutions Leadership – Software and Hardware

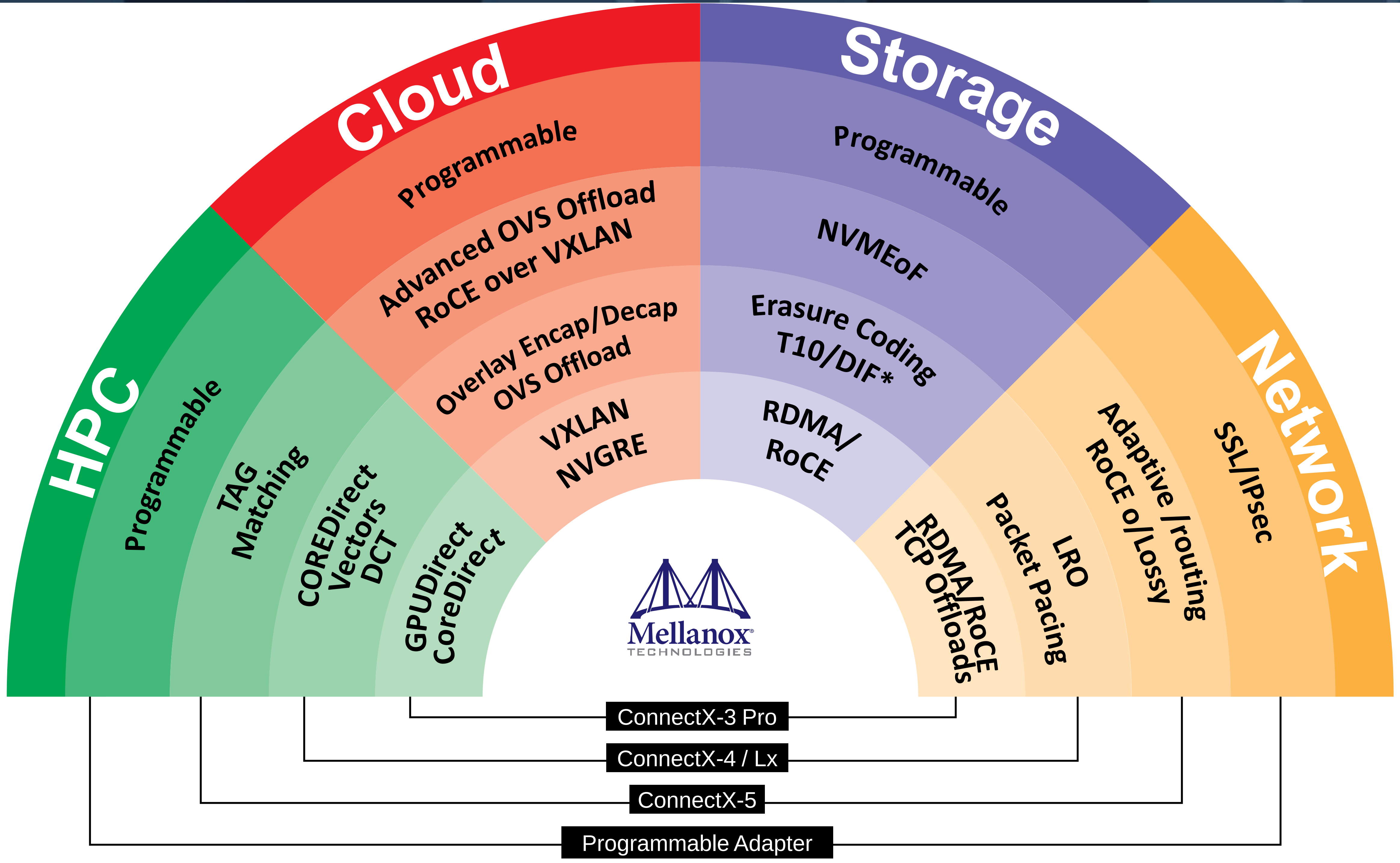
Comprehensive End-to-End Interconnect Software Products

Comprehensive End-to-End InfiniBand and Ethernet Hardware Products

ICs	Adapter Cards	Switches/Gateways	Metro / WAN	Cables/Modules
				

Mellanox Adapters Deliver the Intelligence in the Network

- Unique Set of Application Offloads in the Industry



* ConnectX-4 only

RDMA: Remote Direct Memory Access (RDMA)

What is RDMA?

Direct memory access from the memory of one computer to that of another without involving either one's operating system. This permits high-throughput, low-latency networking, omitting the OS and freeing the Processor to other tasks.

✓Higher ***performance*** and lower latency by offloading CPU transport processing.

✓Remote storage at the ***speed*** of direct attached storage (Including 100Gb/s InfiniBand and RoCE*)

- **Enabling Mobility, Scalability & Serviceability**
  - **More User, Scalability & Simplified Management**
  - **Dramatically Lowers CPU Overhead & Reduces Cloud Application Cost**
  - **Highest Throughput (10/40/56/100GbE), SR-IOV & PCIe Gen3/4**

* RDMA Over Converged Ethernet

Clustered File Systems

- The Original "Software-Defined Storage"
- Distributes File System Across Servers
  - Most popular for HPC
  - Add performance and capacity simultaneously
- Requires High-Speed Network
  - Data distribution and redundancy
  - Metadata and monitor traffic
  - Rapid data access across the cluster
- Support Mellanox RDMA
  - IBM GPFS (Spectrum Scale)

IBM GPFS (Spectrum Scale)

Growth of data, transactions, and digitally-aware devices are straining IT infrastructure and operations; storage costs and user expectations are increasing. As users are added and more data is stored, file-level data availability becomes more difficult to achieve and management can become more complex.

To address these data management challenges, you need a cost-effective alternative that can help you move beyond simply adding storage to optimizing your data management. A single file server does not scale and multiple file servers are not flexible enough to provide dynamic 24x7 data access needed in today's competitive digital marketplace.

The IBM General Parallel File System (GPFS – Spectrum Scale), which is a high-performance enterprise file, can help you move beyond simply adding storage to optimizing data management. GPFS – Spectrum Scale is a high-performance shared-disk file management solution that provides fast, reliable access to a common set of file data, online storage management, scalable access and tightly integrated information life cycle tools capable of managing petabytes of data and billion of files.

IBM GPFS (Spectrum Scale)

IBM GPFS – Spectrum Scale – currently powers many of the world's largest scientific supercomputers and commercial applications that require high-speed access to large volumes of data such as the following examples:

- **Digital media**
- **Engineering design**
- **Business intelligence**
- **Financial analytics**
- **Seismic data processing**
- **Geographic information systems**
- **Scalable file serving**

IBM GPFS (Spectrum Scale)

Mellanox TECHNOLOGIES

There are several advantages of using InfiniBand network:

- **Is a low latency network.**
- **Is a fast network, The latest hardware can provide 100 Gbps.**
- **Can transport IP layer and storage layer on the same hardware infrastructure by combining both communication technology (IP and storage) on the same network layer.**
- **Supports RDMA protocol, which has several major advantages:**
- **Offers zero-copy technology that permits transferring data between the memory of separate nodes.**
- **Lowers CPU utilization**

Storage Acceleration – Next steps?!

NVMe Over Fabrics

NVMe standard maintained by NVM Express, Inc.

- 1.0 in March 2011; 1.1 in October 2012
- 1.2 in final ratification

Need for NVMe Over Fabrics

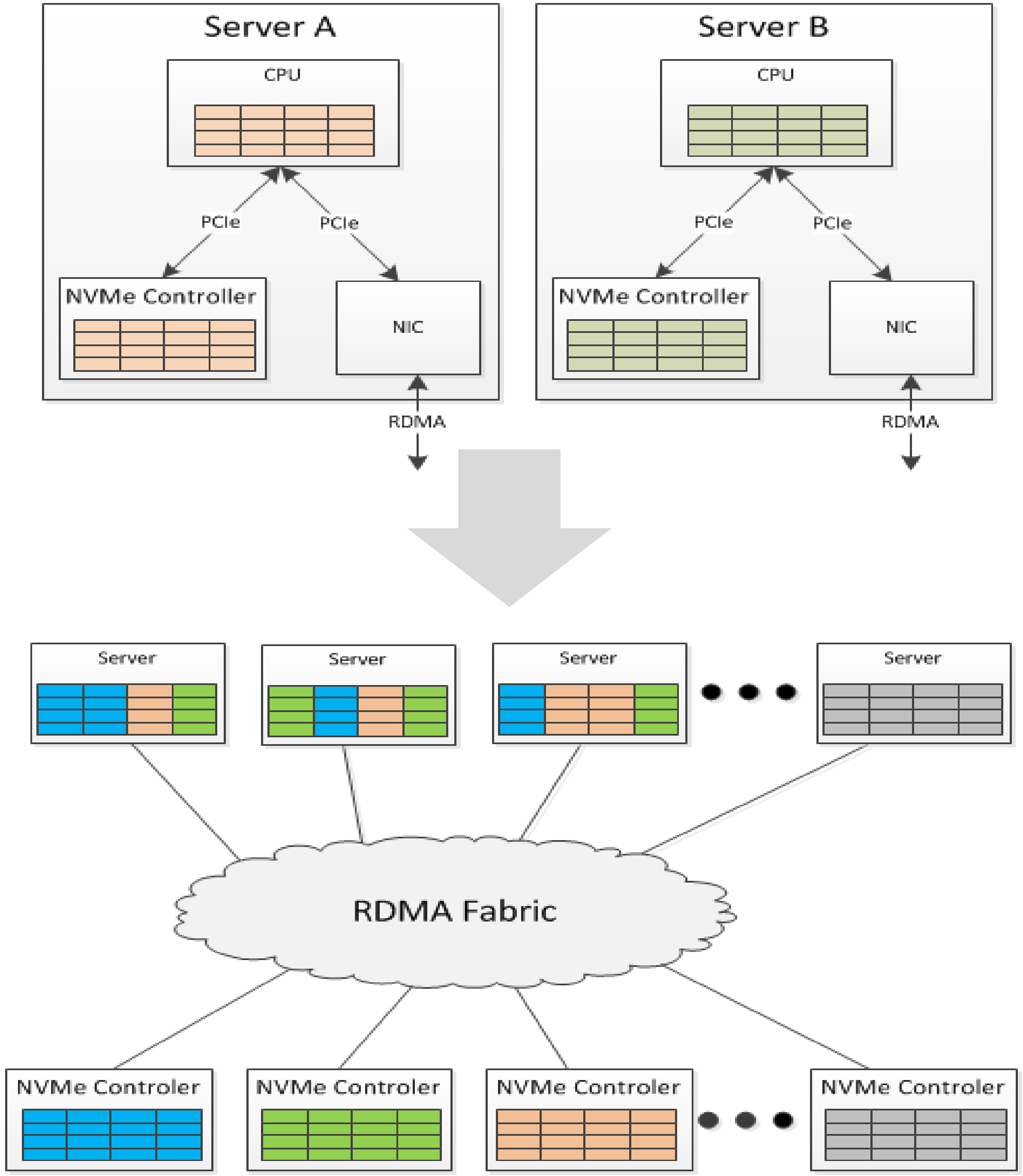
- Scalability, Distance
- Availability / Failover / Flexibility

NVMe over Fabrics standard proposed in September 2014

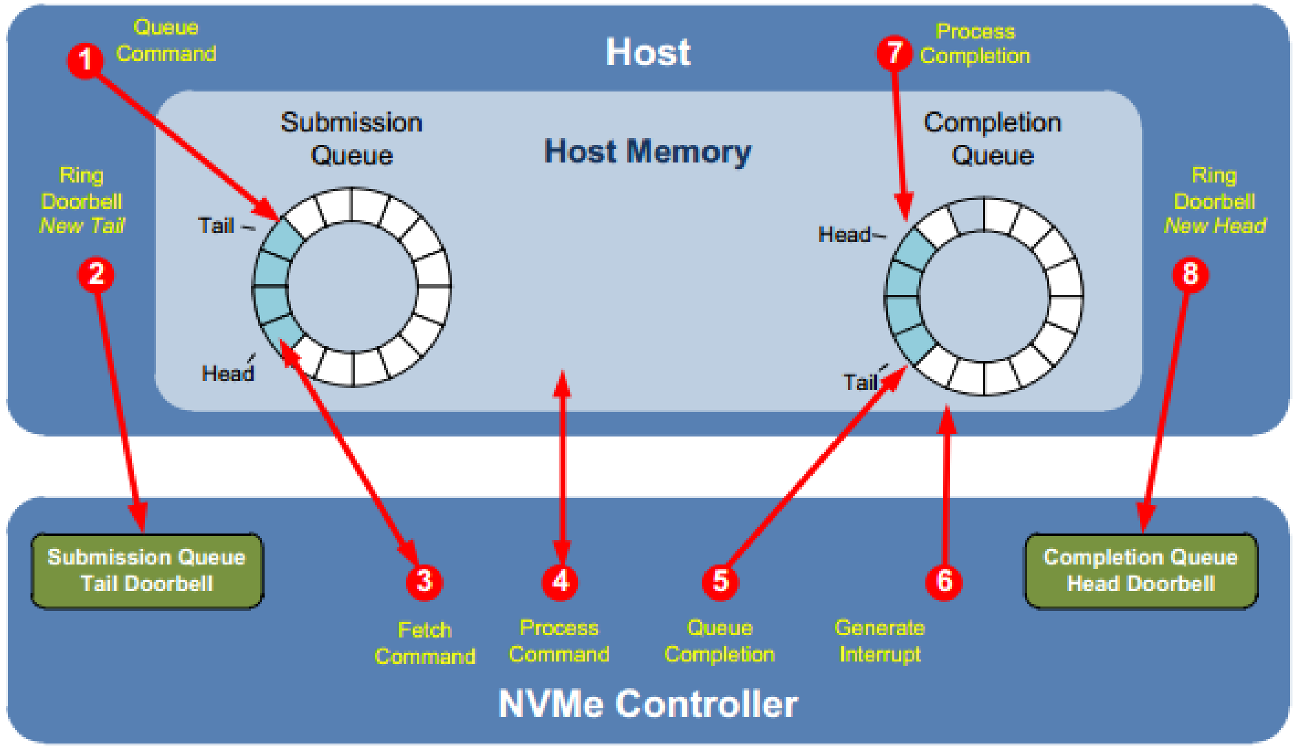
- Intend to support InfiniBand, Ethernet and others
- Mellanox active in the working group

Supported by current Mellanox adapters

- Fastest connections available today
- Most popular RDMA technologies
- NBDX demonstrates today NVMe over fabrics capability



NVMe Operation Flow



Command Submission

1. Host writes command to submission queue
2. Host writes updated submission queue tail pointer to doorbell

Command Processing

3. Controller fetches command
4. Controller processes command

Command Completion

5. Controller writes completion to completion queue
6. Controller generates MSI-X interrupt
7. Host processes completion
8. Host writes updated completion queue head pointer to doorbell

NVME/Fabrics and NBDX

Collaboration

- NVME/fabrics spec work-group
- Spec + Coding work-group

NBDX

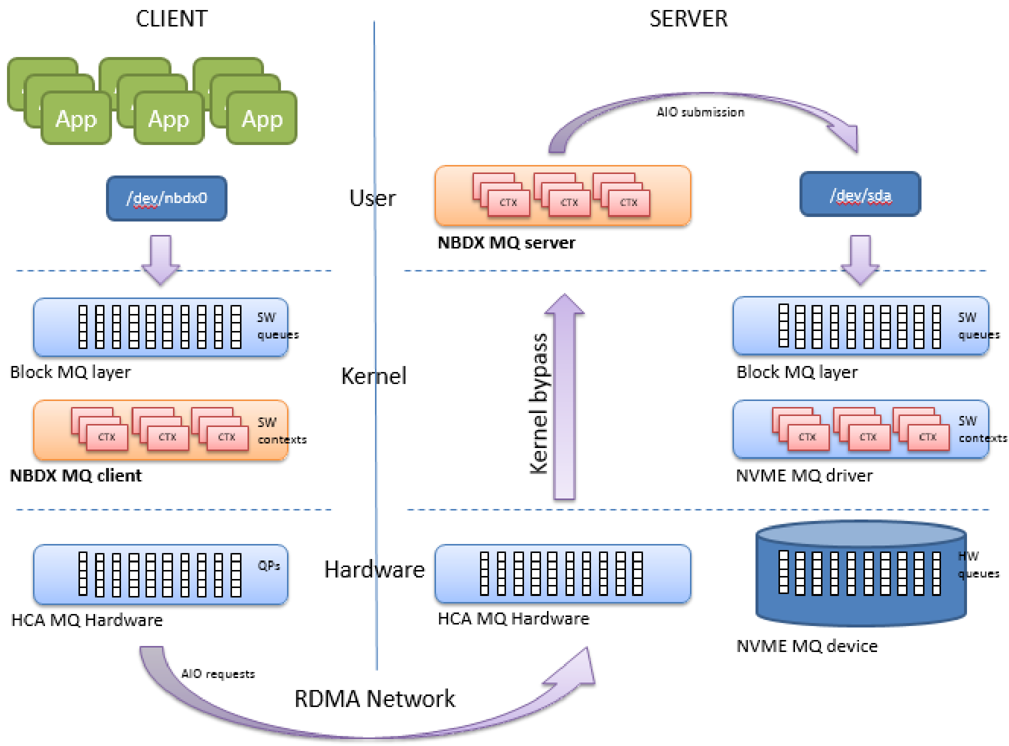
- Pre-standard demo
- Powered by Accelio (See GitHub)
- +20us Lat on NULL target, +50 real NVME

Demo

- Proprietary NVME/RDMA
- +6us Latency on NVME
- Using Mellanox ConnectX-3

Future

- Participate in initiator and target coding
- End-to-end solution on ConnectX-3 and 4
- Target offload in ConnectX-5



NBDX

Features

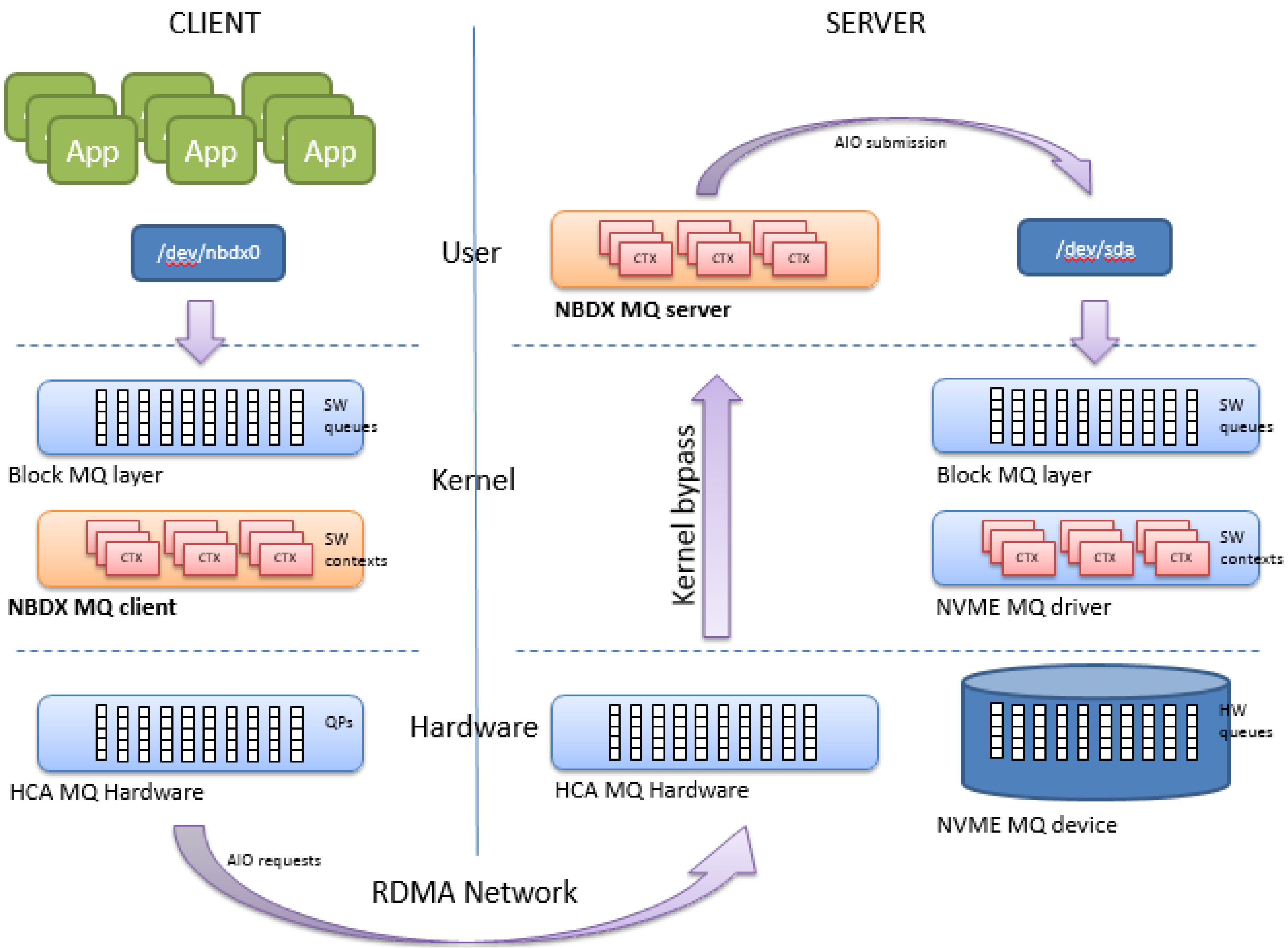
- Block-RDMA driver
- End-to-end multi queue design
- Userspace server
- Encapsulating AIO commands
- Serve any block/file from server
- InfiniBand or RoCE (RDMA)

Pre-standard example

- Encapsulate AIO instead of NVME descriptor
- No cut-through to NVME device on server side

Performance

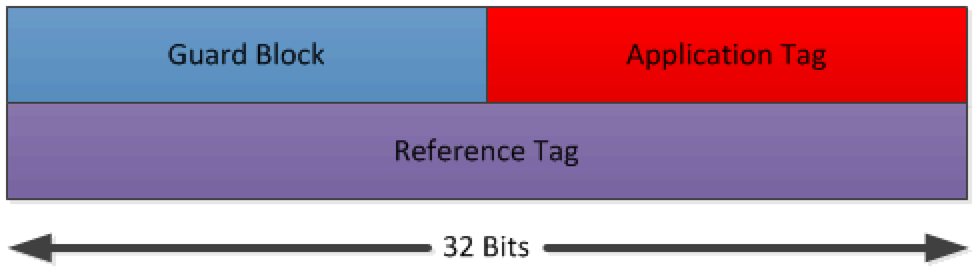
- 2M IOPS for a single LUN
- <18us latency
- Wire speed @16K blocks



T10-DIF / T10-PI

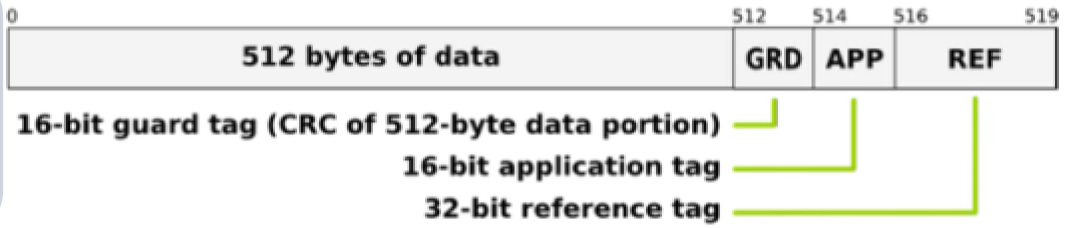
Definition

- By T10 – Storage protocols commitee
- Data Integrity Field (DIF) or Protection Information (PI)
- **Protect each block from data corruption and misplaced write**



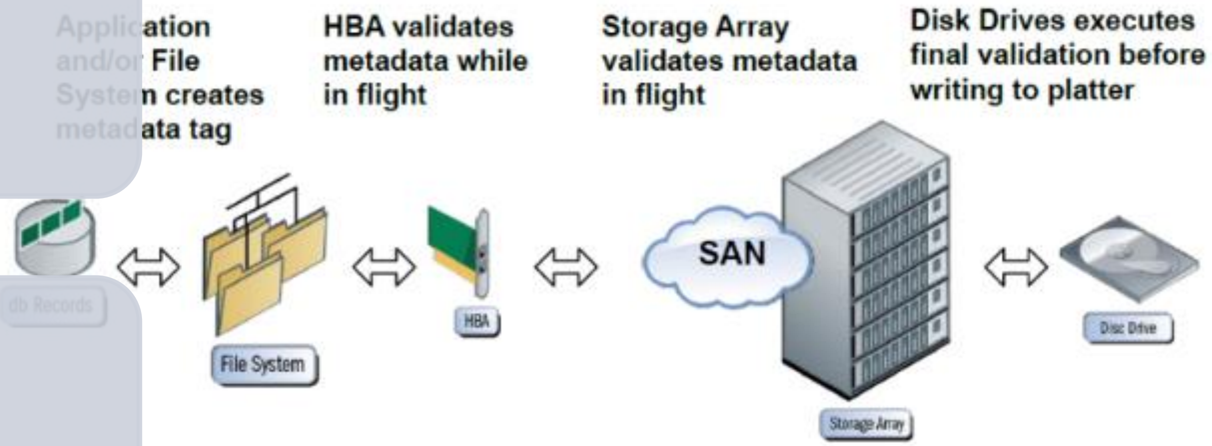
End-to-End

- From application (OS) to disk
- Every point along the path can verify
- Unrelated to per-packet CRC (not end-to-end…)



Components

- Guard tag (CRC or IP-Checksum)
- Reference tag (block address)
- Application tag (free usage)



Offload (ConnextX-4)

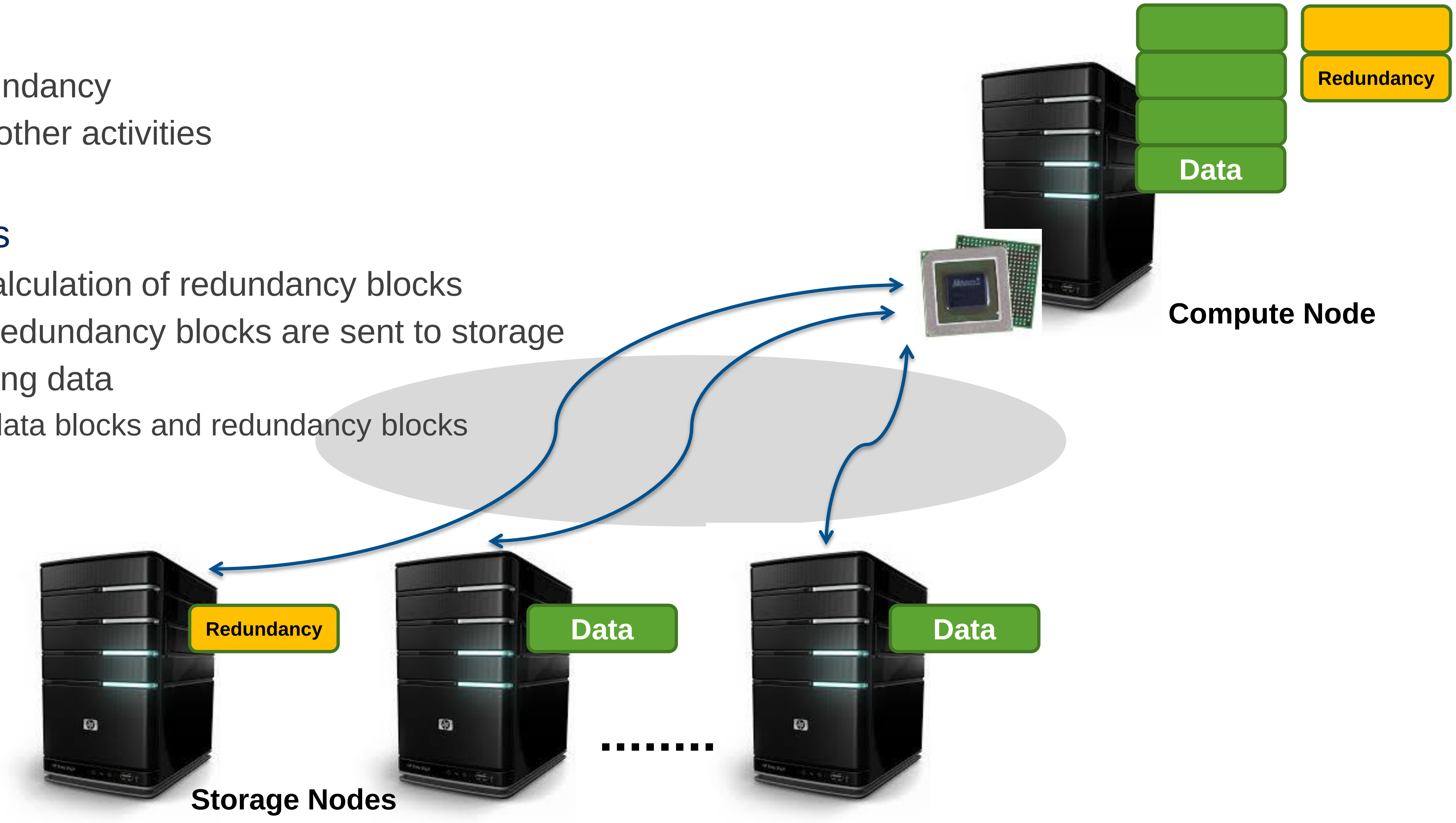
- Add
- Verify and pass
- Verify and strip
- Convert CRC/Checksum, block sizes

The Value of Erasure Coding

- **EC = More Efficient Data Protection**
  - Only 30-50% overhead
  - Withstands multiple failures
  - Can include geographical dispersion
  - High level of data protection
- **Who Uses Erasure Coding Today?**
  - Object storage
  - Customers with big content, >1PB
  - Solutions with large hard drives
  - Applications that tolerate high latency
- **Limitations of Traditional RAID**
  - Only protects against 1 (or 2) failures
  - Fast or efficient or reliable – pick any one
  - High risk of data loss with large drives
  - Often requires RAID + remote mirroring
- **Why doesn’t Everyone use Erasure Coding?**
  - CPU-intensive (slow) on writes or reconstruction
  - Cost-effective only if >300TB of content
  - Smaller deployments okay to pay RAID tax
  - Common RAID-1/4/5/10 hardware offloads

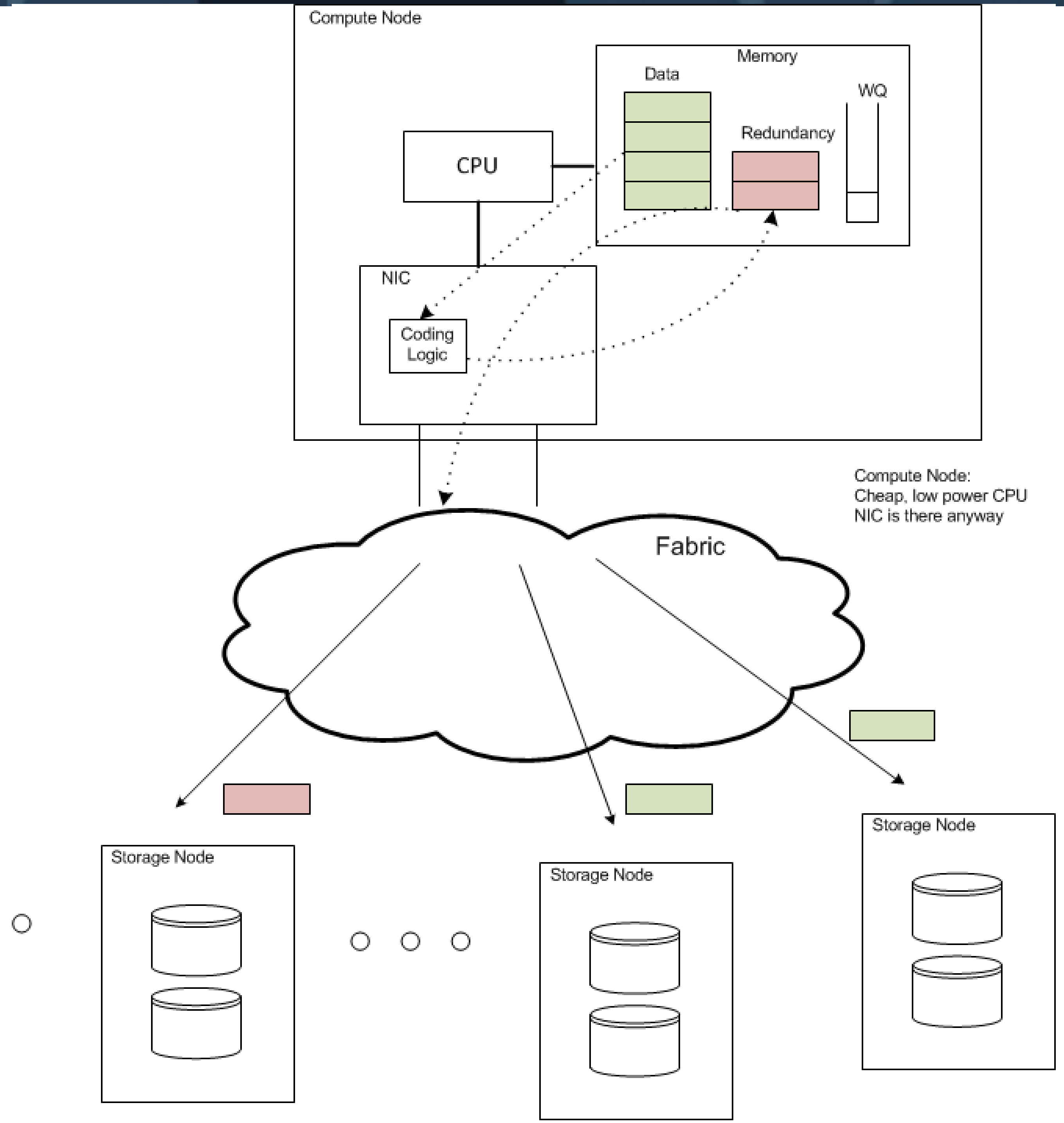
Erasure Coding Offload

- Advantages
  - Cluster-level redundancy
  - Free-up CPU for other activities
- Hardware offloads
  - Reed-Solomon calculation of redundancy blocks
  - Data blocks and redundancy blocks are sent to storage
  - Recalculate missing data
    - Based on other data blocks and redundancy blocks



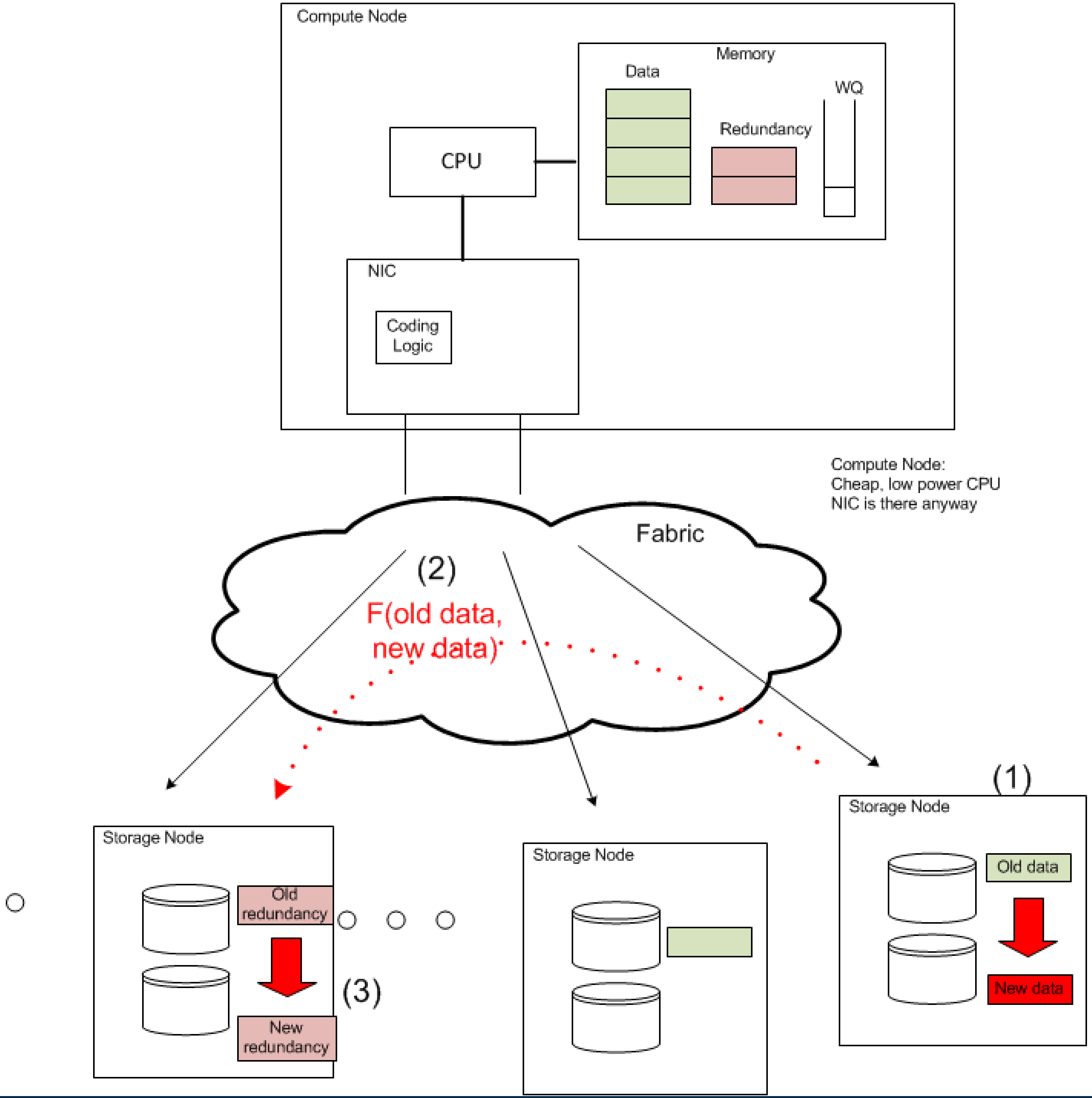
RAID implementation in Cloud – Encoding

- Data exists in compute node
- Redundancy is calculated in the compute node
  - Offloaded to the NIC
- Data and redundancy are sent to the storage nodes



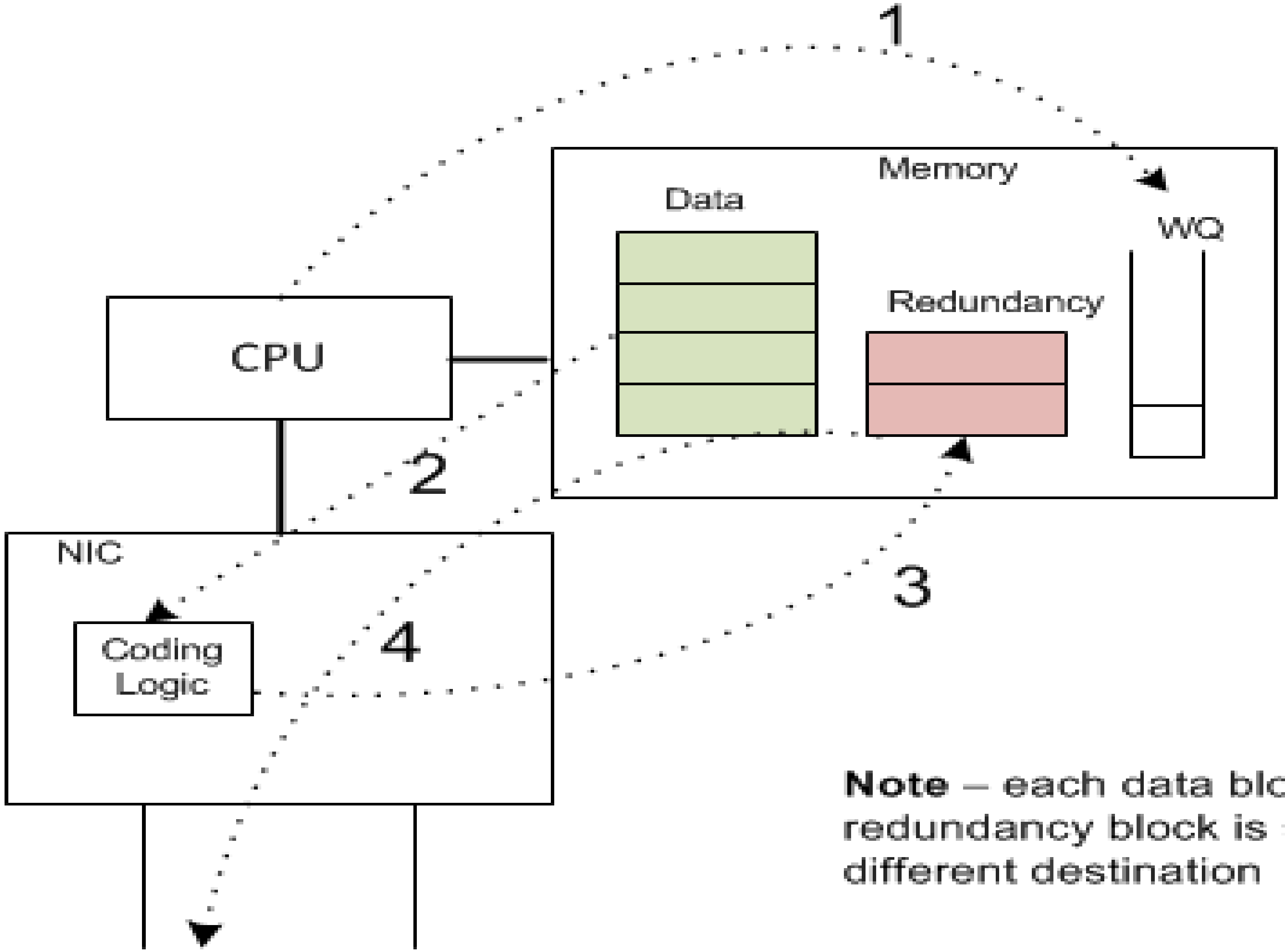
RAID implementation in Cloud – Data Update Flow

- (1) data is changed in storage node
- (2) Storage node calculate F(old data and new data)
  - Send this to location of the redundancy
- (3) Storage node that hold the redundancy block calculate the new redundant block
  - Function of data from one and old redundant block



RAID Offload - Encoding

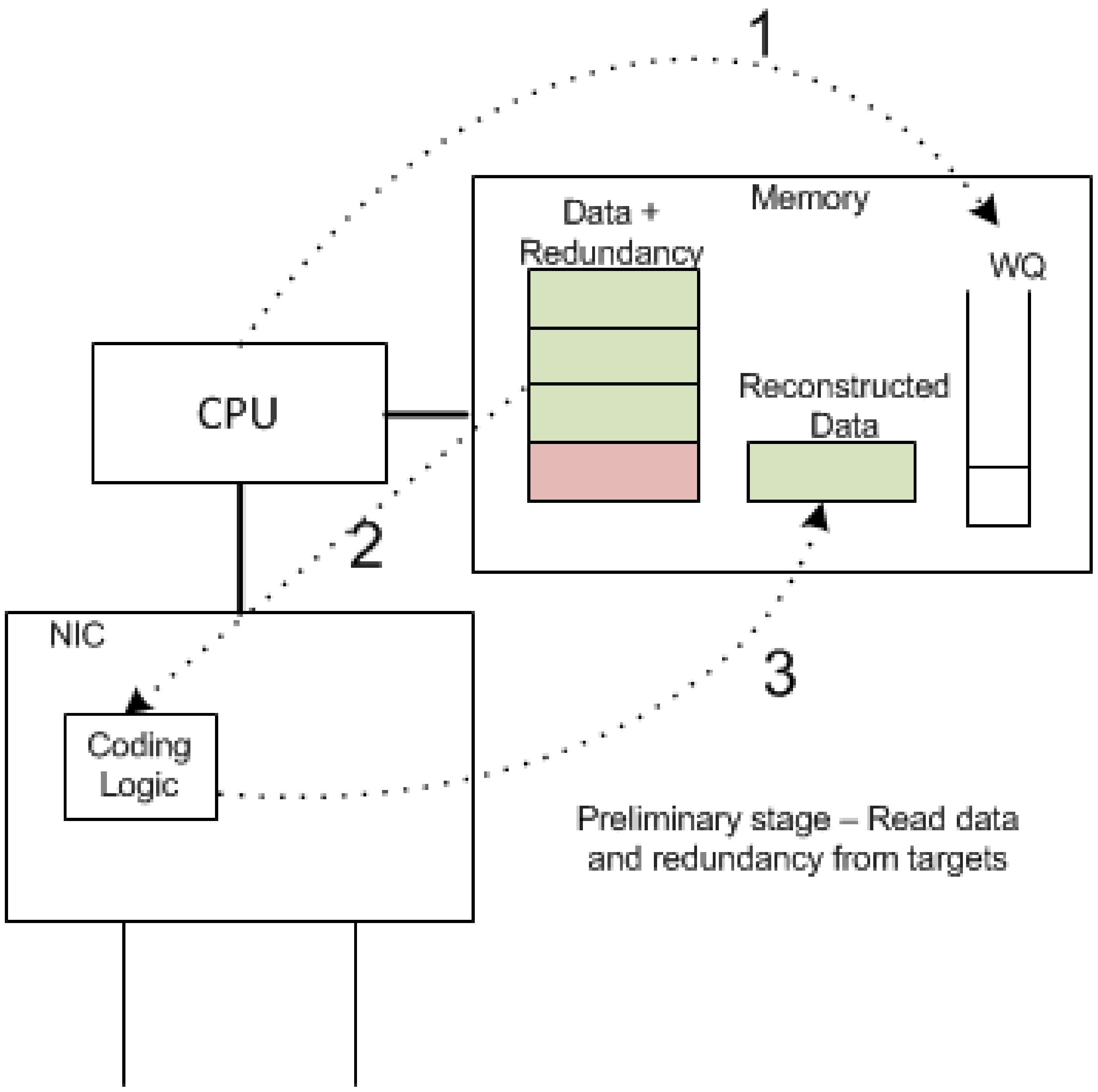
- Use case:
  - Data already exists in the node
  - Need to calculate redundancy
  - Need to send data and redundancy to other nodes.
- Alternative Use case
  - Data does not exist in the node
  - Read data from other nodes
  - Calculate redundancy
  - Not shown in this slide



Note – each data block and redundancy block is sent to a different destination

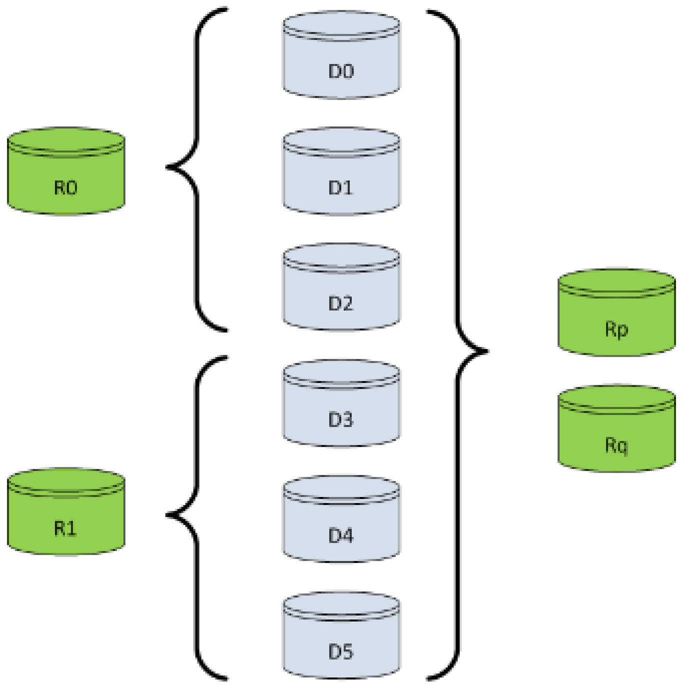
RAID Offload – Decoding

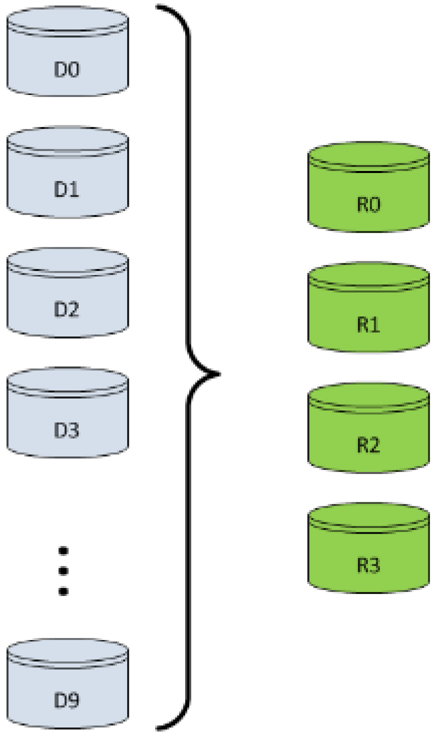
- Use case:
  - 3 blocks of data and redundant block already fetched to this node
  - Need to calculate missing data block



Encoding Schemes:

- Reed Solomon coding
- RS(n,k):
  - n: data blocks
  - k: redundant blocks
- J-Erasure-like implementation
- Implementation
  - m: symbol size
    - m=4 bit
    - Up to 15 data block
    - Up to 15 redundant blocks
- Usage examples:
  - RS(10,4)
  - Microsoft LRC (6,2,2), (12,2,2)
    - Locally Repairable Codes





Erasure Coding Offload FAQs

- Can EC Offload Support RAID-6?
  - Yes Reed-Solomon is commonly used for RAID-6
- How many data and redundancy blocks?
  - User-defined, up to the maximum limit
- Can it offload updates or reconstruction?
  - Yes
- Is EC offload needed to read stored data?
  - Only if some data blocks have been lost
- Does it track where blocks are stored?
  - No, the storage/application must track it
- Must data/redundancy blocks be remote?
  - No, distribute across local disks or remote nodes
- Is EC more efficient than RAID?
  - Yes for larger drives or vs. smaller RAID groups
  - Yes if it replaces RAID+mirroring
- Why do large drives need smaller RAID groups?
  - Reconstruction too slow with large RAID groups
- How fast can Erasure Coding offload run?
  - In most cases at line speed up to 100Gb/s
- Does EC offload add latency to storing data?
  - Yes but much less than EC done in software
- Why don't all storage systems use HW RAID?
  - RAID cards increase cost and reduce flexibility
- Will EC offload work with SSDs? NVMe flash?
  - Yes
- Can EC offload work with IB & Ethernet
  - Yes, with both
- How much data overhead for Erasure Coding?
  - From 10% to 100%; Typically 15-60%

Question Time

Thank You

Connect. Accelerate. Outperform.™