Time Series Analytics at Scale

Record setting STAC-M3 performance with largest dataset (50TB) ever benchmarked

Anand Bisen, Principal Solutions Architect
STAC Summit NYC (Nov 2016)

DELL EMC

Many great results to talk about…

Record breaking performance for Tick Data Analytics with Kdb+

Industry's first STAC M3 benchmark to test with 50TB of data

Previous leading benchmark tested with 30TB of data

Distributed DSSD architecture can continue to scale

With only 33% storage performance used and 50% storage capacity used

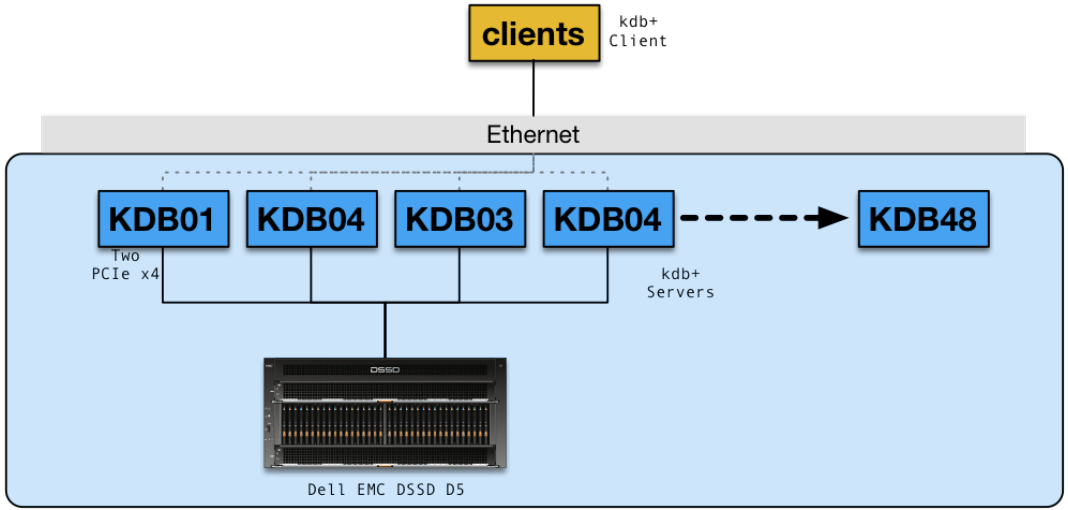
Kanaga

- **Fastest results in 16 of the 19** benchmarks reported previously
- First solution to be benchmarked with 5 Years of data (**47.6TB**)
- Achieving 22GB/s in the 3-year high bid benchmark (1T.3YRHIBID)

Antuco

- **Faster in 14 of 17** benchmarks compared to direct attached SSD storage (XTR141111)
- **Faster in 12 of 17** benchmarks compared to shared flash storage (XTR160413)
- **Faster in 16 of 17** benchmarks compared to a distributed system with 22 database servers (KDB150528)

Scalable Kx kdb+ platform for tick analytics



Consolidate Real-Time and Historical Tick Silos and share across the organization…

Dense and Shared Flash

DSSD D5 – 5U rack scale flash platform

FLASH AND CMs

36 Flash Modules (FMs)
18 Flash Modules when Half Populated

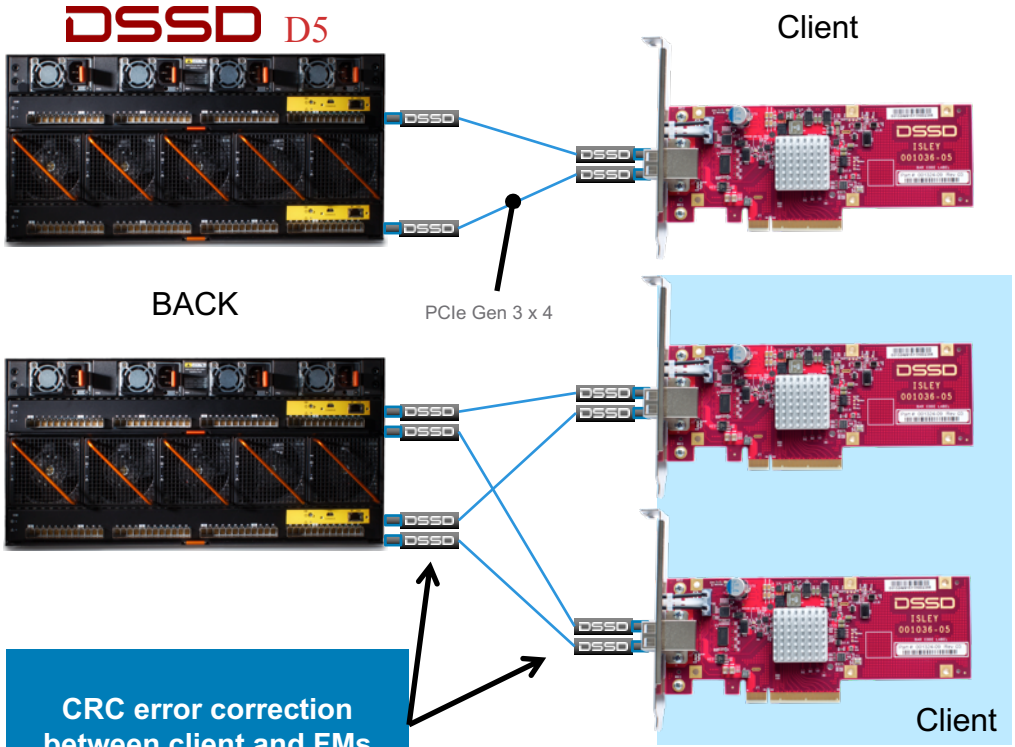
2T/4TB Flash Modules today
Larger FMs on the roadmap

Dual Ported PCIe Gen3 x4 Per FM

Dual-Redundant Control Modules (CMs)
PCIe Gen 3 Connected

Client connectivity

Up to 48 hosts dual connected to one D5



Key Features

- 1-2 DSSD Client Cards per host
- DSSD I/O Cable (PCIe Gen 3 x4) connects Client Card ports to DSSD D5 I/O Module ports
- CRC Data Protection
- Always-On Multipathing
- Client connects to single D5

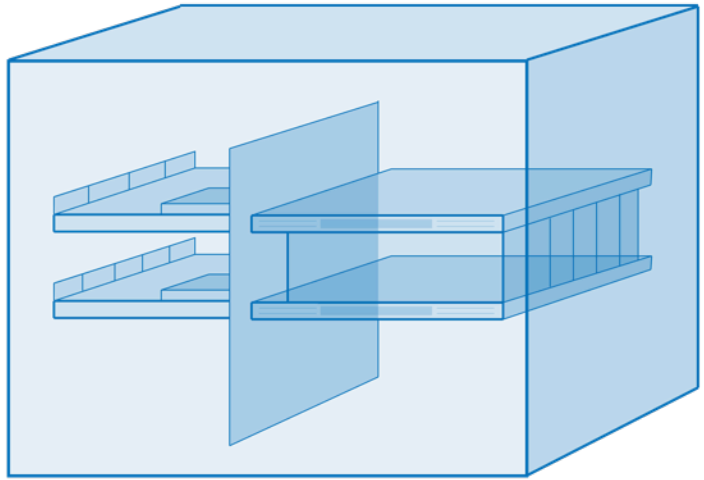
DELL EMC

5

Appendix:
D5 Technical Overview

DSSD Hardware Overview

OBJECTIVE:
Maximize flash performance in a space-efficient and highly available fashion



DSSD

D5

Hardware

Maximizes flash performance in a space-efficient and highly available fashion

Software

Maximizes flash performance with flash management stack in a way useful to applications - Flood

Data Protection

Maximize data protection with less overhead at flash speeds

Dense and Shared Flash

DSSD D5 – 5U rack scale flash platform

FLASH AND CMs

36 Flash Modules (FMs)
18 Flash Modules when Half Populated

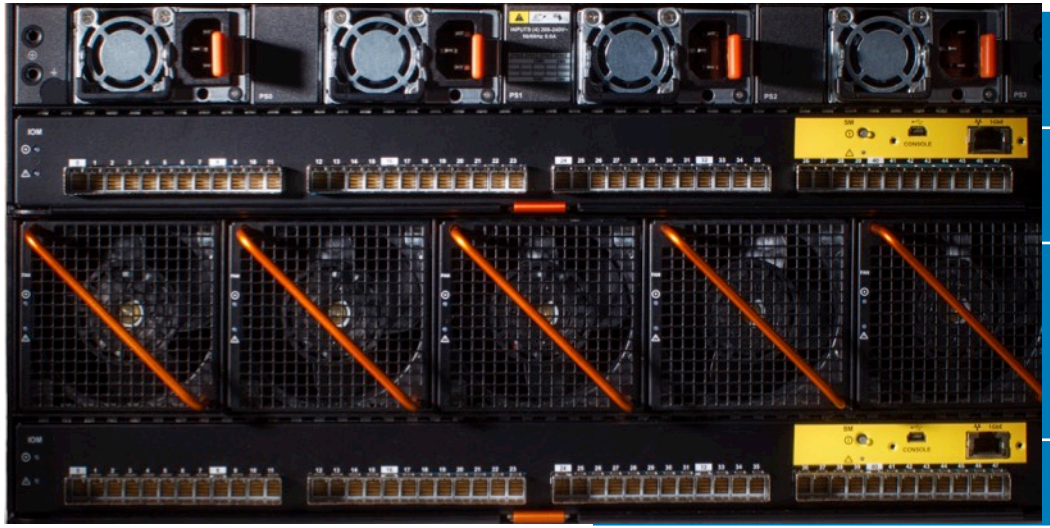
2T/4TB Flash Modules today
Larger FMs on the roadmap

Dual Ported PCIe Gen3 x4 Per FM

Dual-Redundant Control Modules (CMs)
PCIe Gen 3 Connected

Dense and Shared Flash

DSSD D5 – 5U rack scale flash platform



IOMs, Fans, Power Supplies

Redundant Power Supplies x4

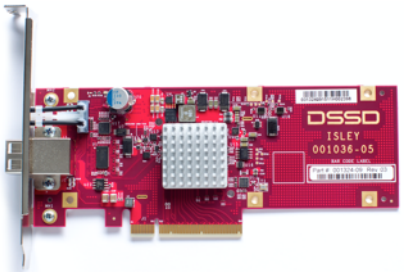
Dual-Redundant IO Modules (IOMs)
PCIe Gen 3 Connected

48 PCIe Gen 3 x4 Client Ports Per IOM
Total of 96 PCIe Gen 3 x4 Client Port Connections per D5

Redundant Fan Modules x5

Dense Client Card and DSSD I/O Cable

Card creates a non-transparent bridge to the attached DSSD appliance



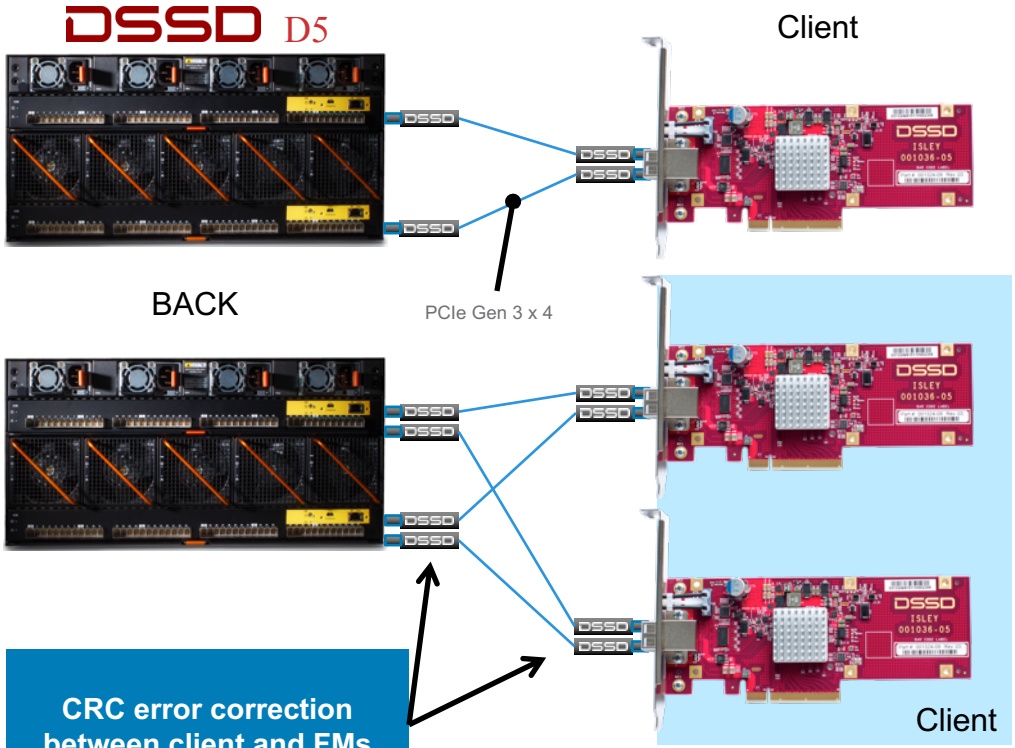
DSSD
CLIENT
CARD

DSSD
I/O
Cable

- DSSD Client Card fits in qualified X86-based servers
- Each with 2 x PCIe Gen 3 x 4 lane ports
- Hosts must run supported version of Linux
- DSSD I/O Cables based on new type of connector
- One or two cards per client, must connect to single D5
- Lengths of 1m, 2m, 3m & 4M copper

Client connectivity

Up to 48 hosts dual connected to one D5



Key Features

- 1-2 DSSD Client Cards per host
- DSSD I/O Cable (PCIe Gen 3 x4) connects Client Card ports to DSSD D5 I/O Module ports
- CRC Data Protection
- Always-On Multipathing
- Client connects to single D5

DELL EMC

12

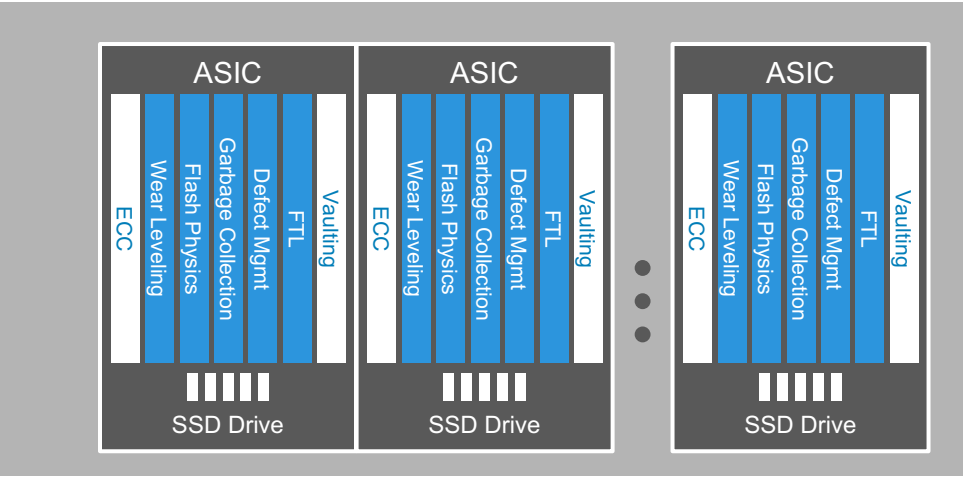
DSSD Flash Module (FM)

- 2TB, 4TB
- Future FMs will offer greater capacities and support 3D NAND and NGNVM
- Industry's first Hot-Swappable PCIe Gen 3 x4 based FM
- Industry's first Dual Connected PCIe Gen 3 x4 FM
- Industry's most reliable FM
- Power delivery to flash bursts up to 60 Watts

DSSD FM vs. other flash storage

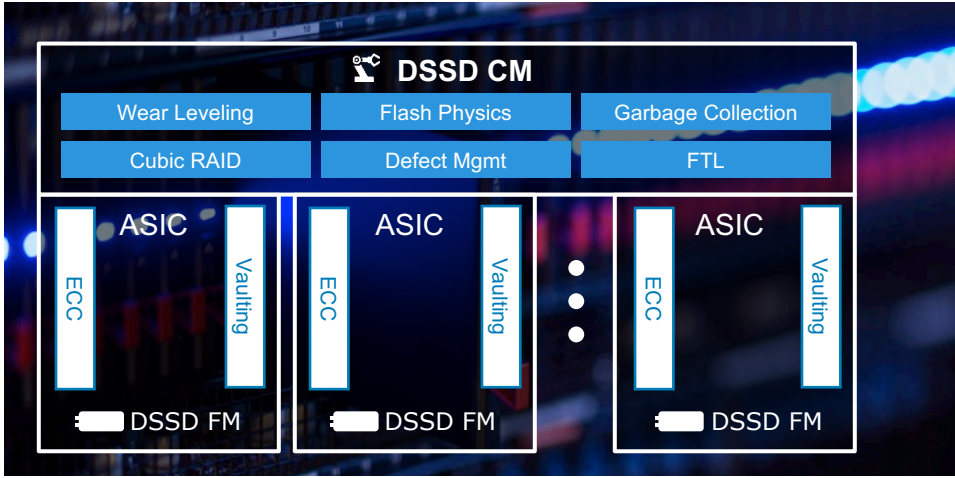
Simpler and faster flash modules

Standard Flash Devices



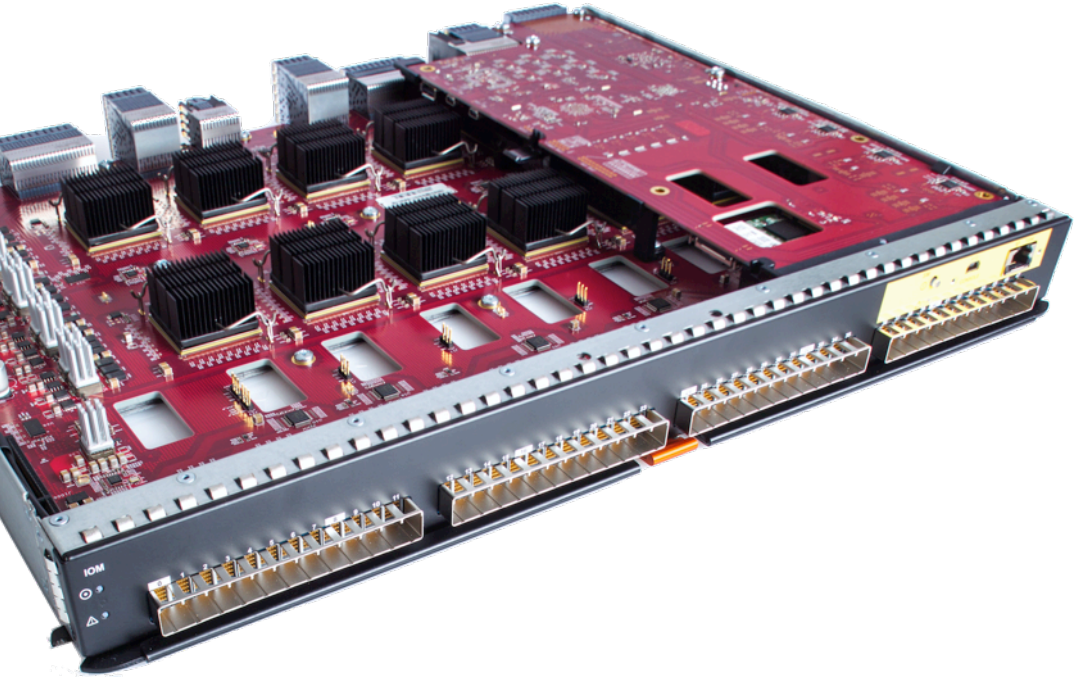
- Complex firmware, limited power
- Independently managed media

DSSD D5



- DSSD has simple, fast Flash Modules
- Control Module with rich resources implements advanced global algorithms

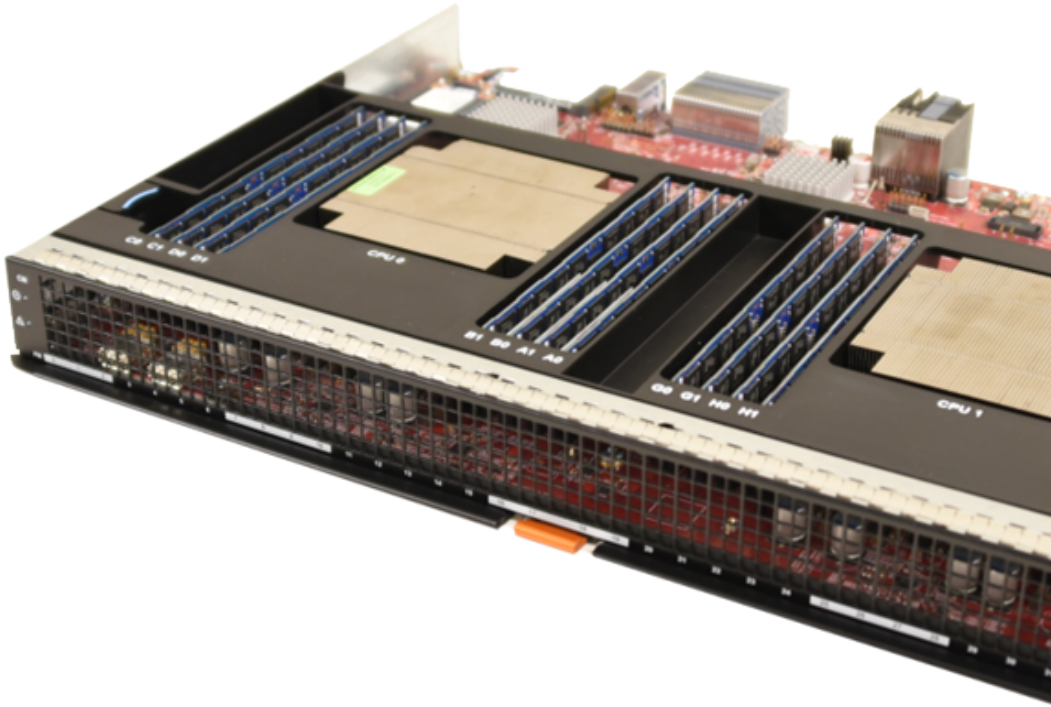
DSSD I/O Module (IOM)



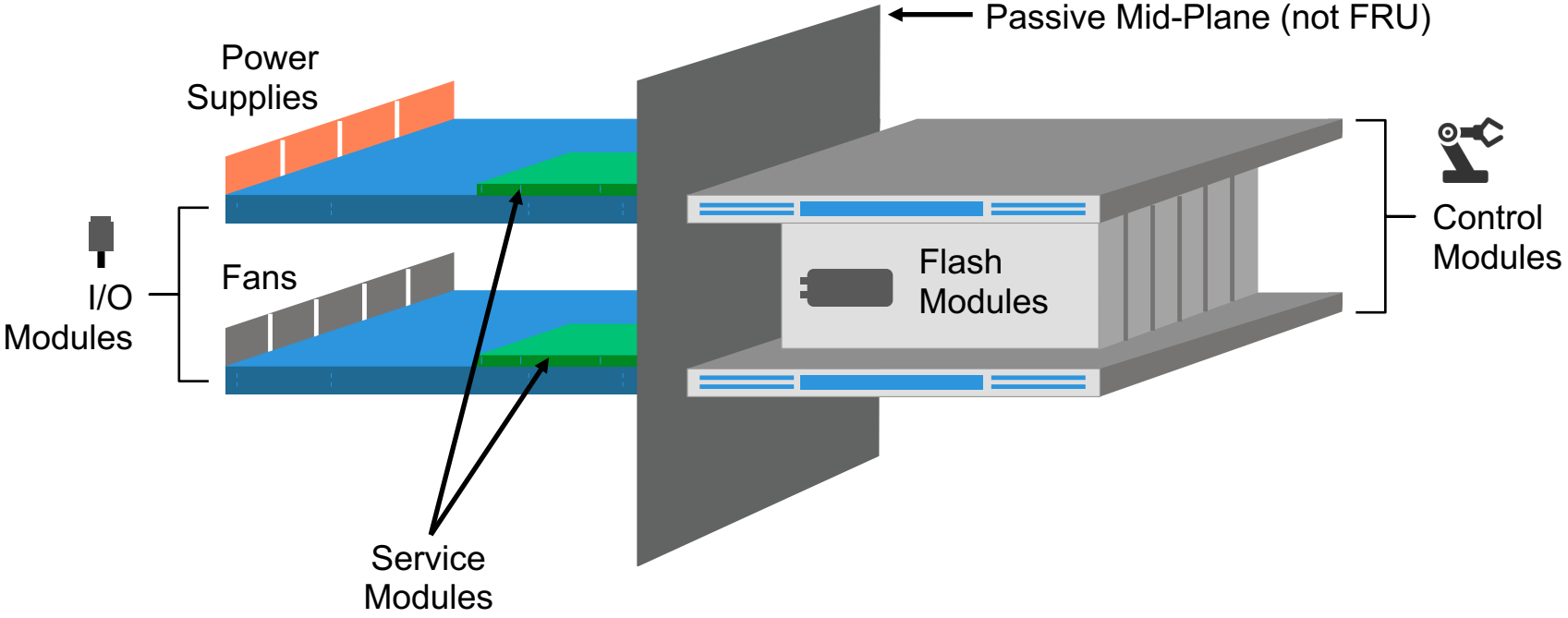
- The IOM interconnects Client Ports, CMs and FMs
- 12x PCIe Hubs
- 48 PCIe Gen 3 x4 connections to client
- 36 PCIe Gen 3 x4 links to Flash Modules
- Dual-Redundant FRU
- Service Module Daughter-Board

DSSD Control Module (CM)

- Manages the control plane, FMs, IOMs
- Receives copy of all data writes but only for Cubic RAID calculations
- Dual Redundant FRU
- Appliance-wide view of all activity so more advanced flash management algorithms
- Flood, the DSSD D5 software runs on the CM



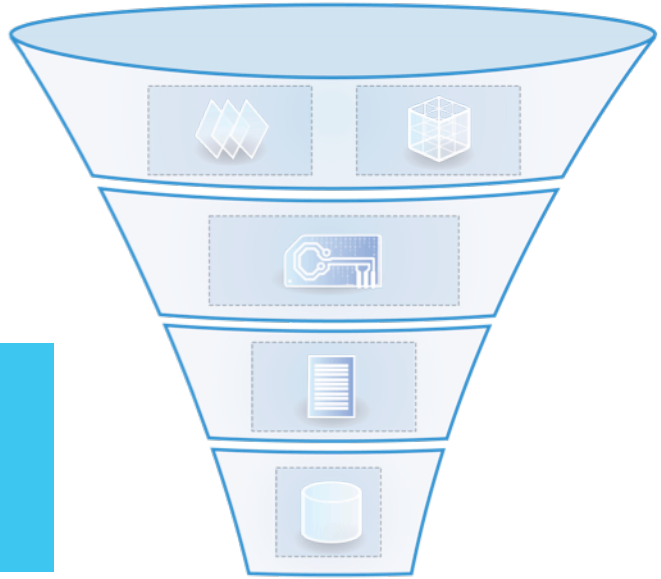
D5 hardware block diagram



DSSD

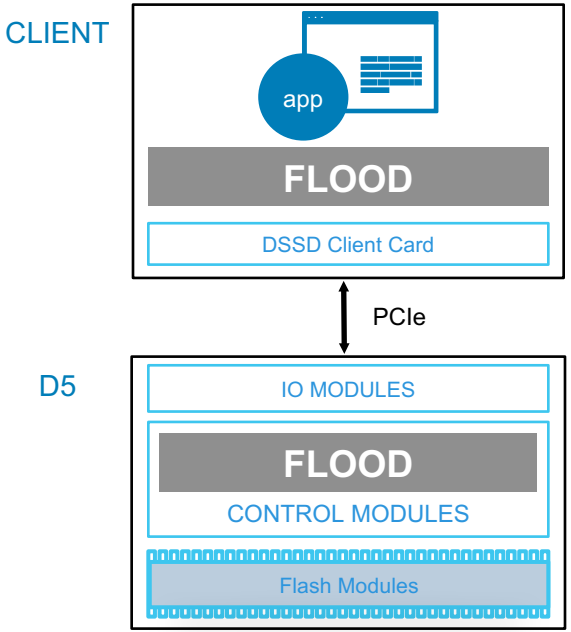
SOFTWARE OVERVIEW

OBJECTIVE:
Maximize flash performance with flash management stack in a way useful to applications - Flood



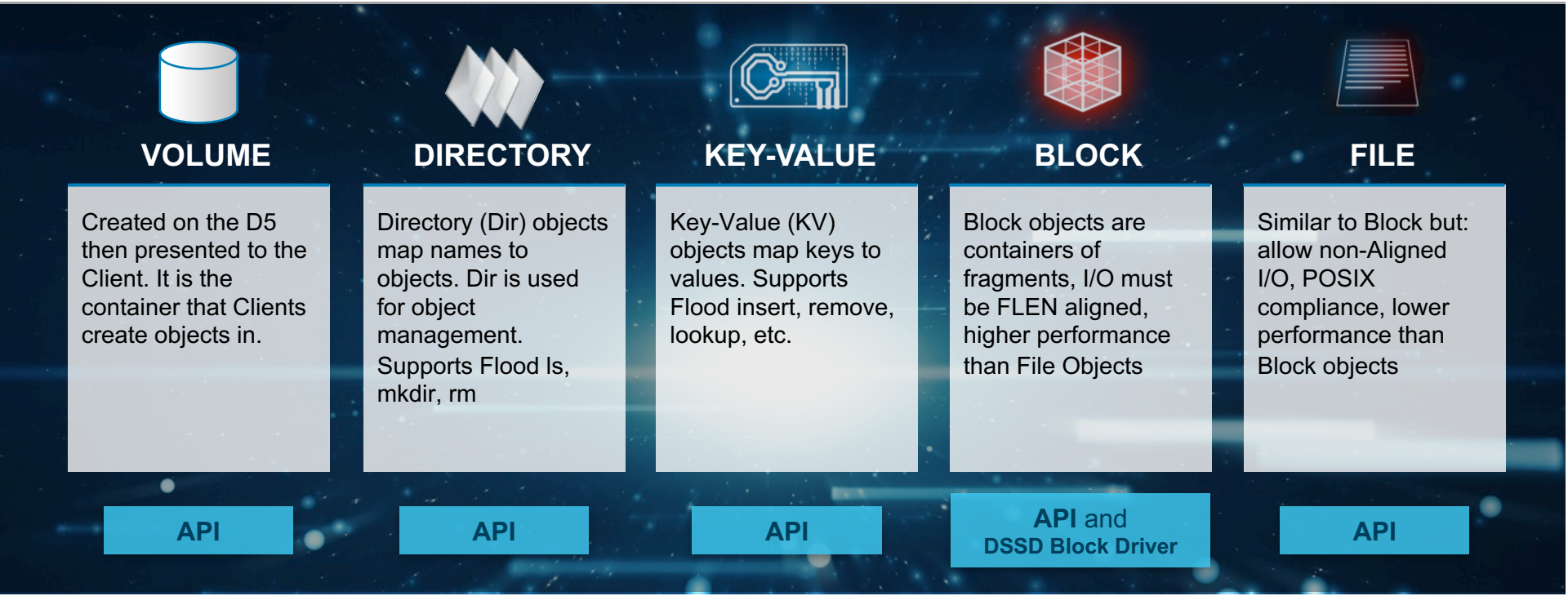
What is Flood?

Deliver the full performance of a pool of flash in a way that is useful to applications



- Flood is the Client Software, CLI, BUI
- Flood is a Direct Memory API
- Flood is the Object Store supporting Key-Value, Directory, Block and File Object types
- Flood is the Data Protector
- Flood is the Data Manager
- Flood is the Appliance Manager, Appliance CLI

Types of DSSD objects



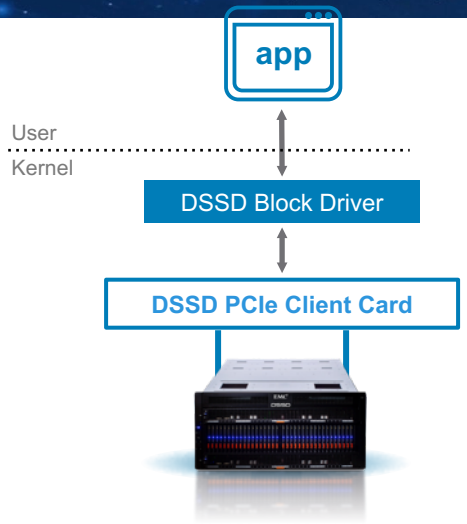
VOLUME	DIRECTORY	KEY-VALUE	BLOCK	FILE
Created on the D5 then presented to the Client. It is the container that Clients create objects in.	Directory (Dir) objects map names to objects. Dir is used for object management. Supports Flood ls, mkdir, rm	Key-Value (KV) objects map keys to values. Supports Flood insert, remove, lookup, etc.	Block objects are containers of fragments, I/O must be FLEN aligned, higher performance than File Objects	Similar to Block but: allow non-Aligned I/O, POSIX compliance, lower performance than Block objects
API	API	API	API and DSSD Block Driver	API

ACCESS METHODS

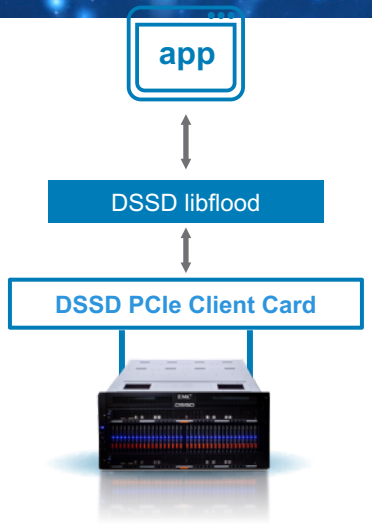
Native and flexible data access

APPLICATION CENTRIC & PROVIDES MAXIMUM CHOICE

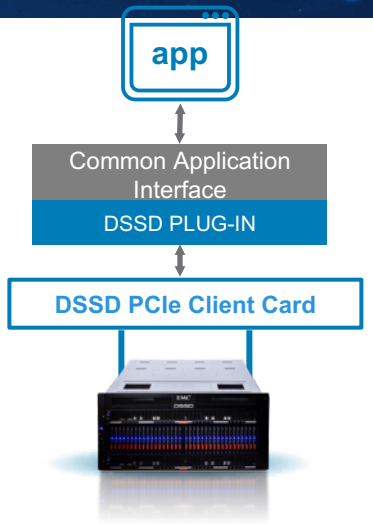
DSSD BLOCK DRIVER



FLOOD DIRECT MEMORY API

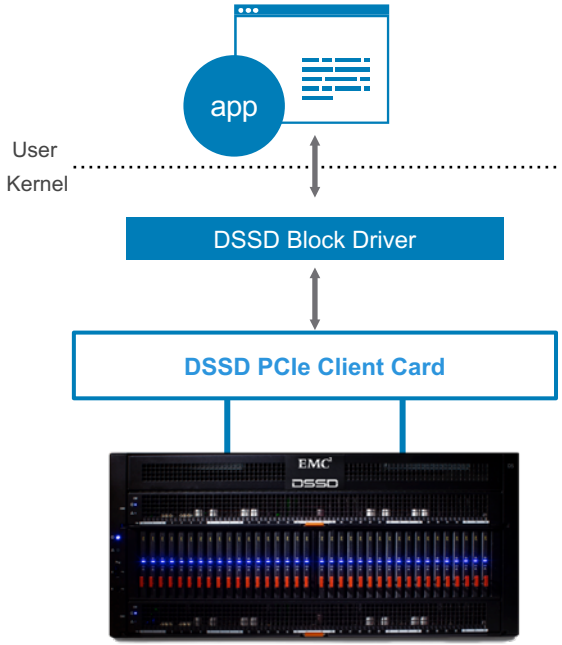


DSSD PLUG-INS



DSSD DELL EMC

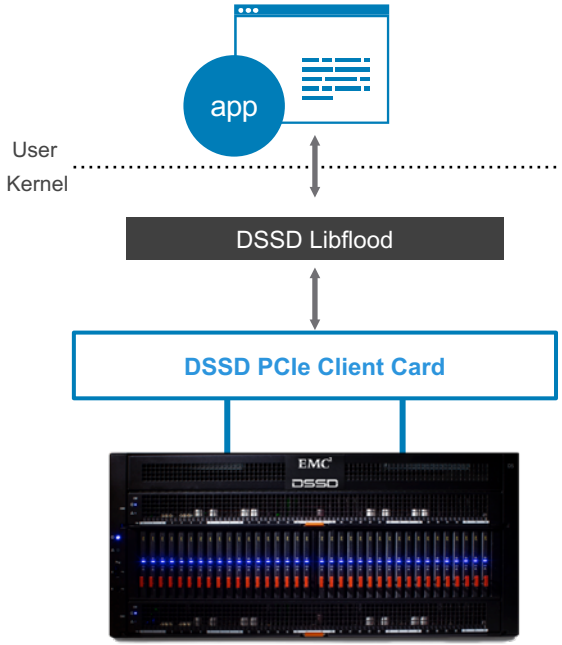
Access method: DSSD Block Driver



ACCELERATED BLOCK I/O

- New devices appear under client's device tree
- Two Fragment Lengths allowed by Linux: 512B and 4KB (4KB strongly recommended)
- Usable by unmodified client applications
- Performance decrease due to kernel overhead
- Using the DSSD Block Driver is the only access method that introduces software into the data path

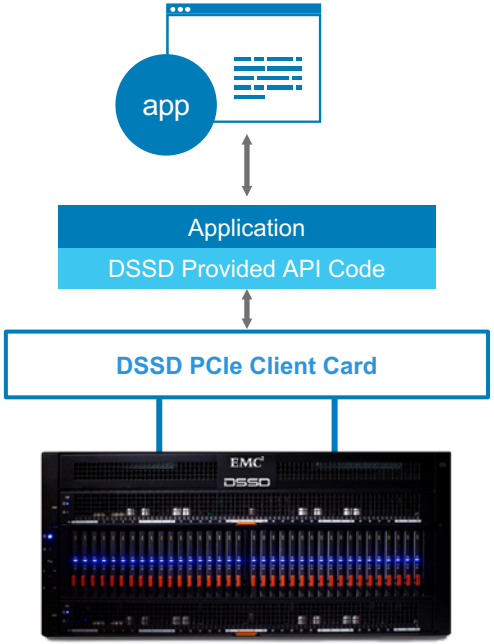
Access method: Flood direct memory API



ACCELERATED APPLICATIONS

- The DSSD libflood “C” library allows applications to utilize the Flood Direct Memory API
- All I/O is hardware triggered PCIe memory mapped DMA to and from user space
- Objects can be viewed as a flat namespace or as a directory hierarchy
- Object types and features beyond POSIX I/O are provided for advanced applications

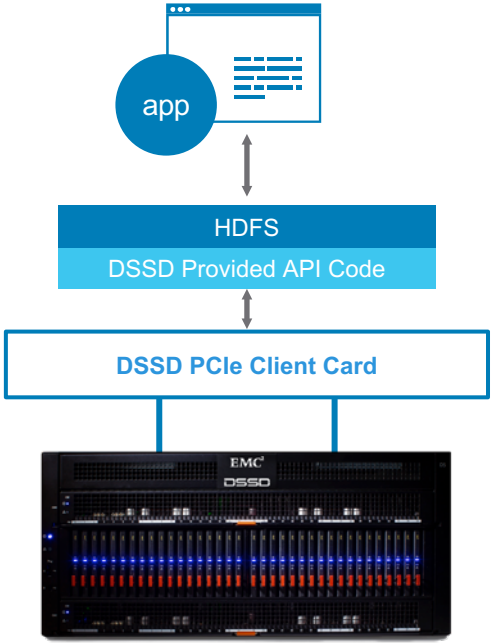
Access method: DSSD Plug-In



ACCELERATED APPLICATION I/O

- Apps with plug-in architectures can use the DSSD Plug-In feature
- Apps that can be modified with shared libraries can also benefit from DSSD created application specific shared libraries
- Apps get improved performance without any change

Access method: HDFS via Plug-In

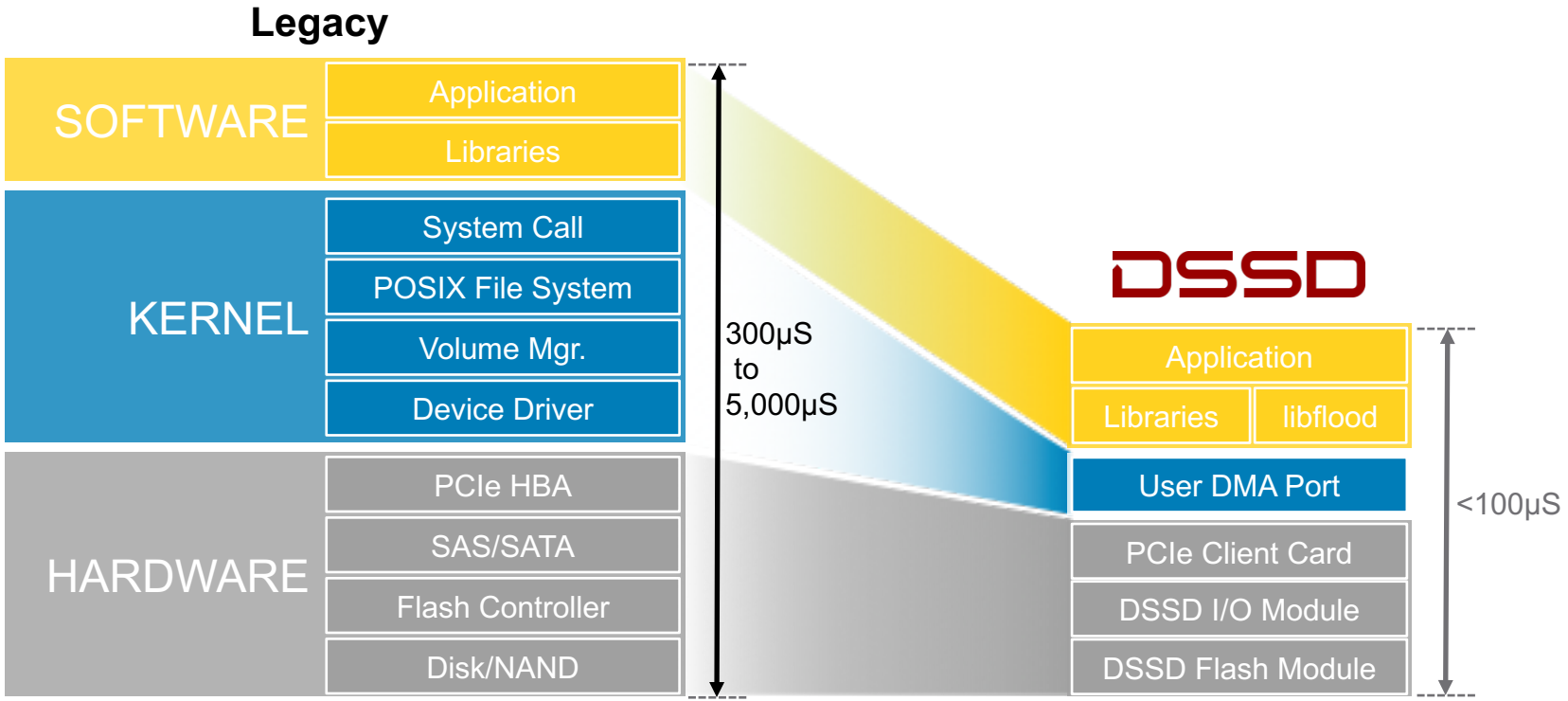


ACCELERATED HDFS I/O

- HDFS supports a Plug-In Architecture
- DSSD Hadoop Plug-In will be the first DSSD supplied API Code released
- The first Hadoop Distribution to be certified will be Cloudera
- Other HDFS distributions such as Pivotal, HortonWorks etc. will be released as they are certified

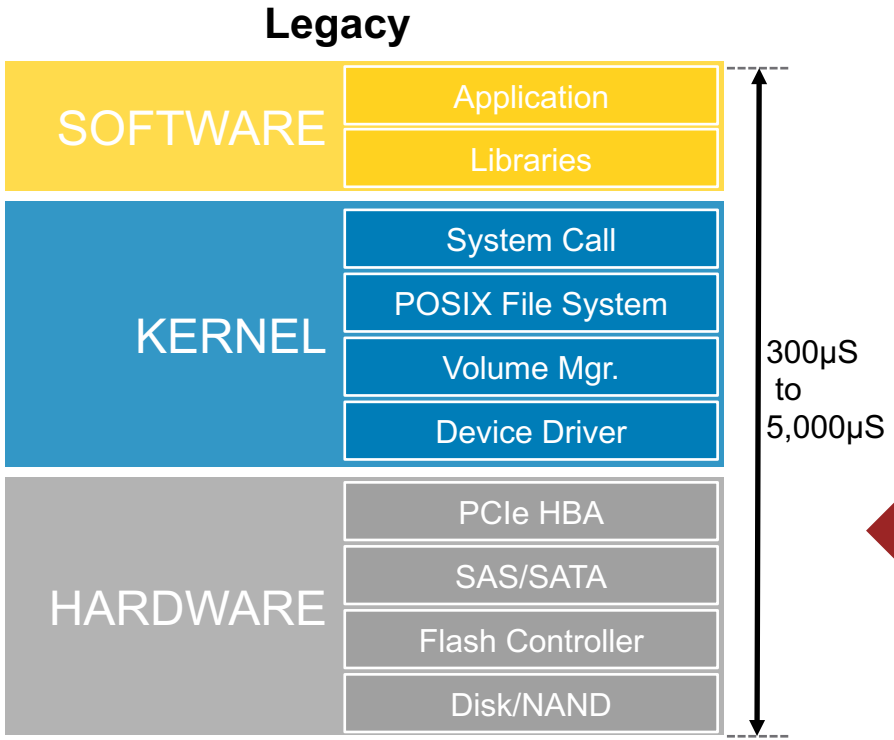
Software performance architecture

Measure performance to the application



Even more latency removed

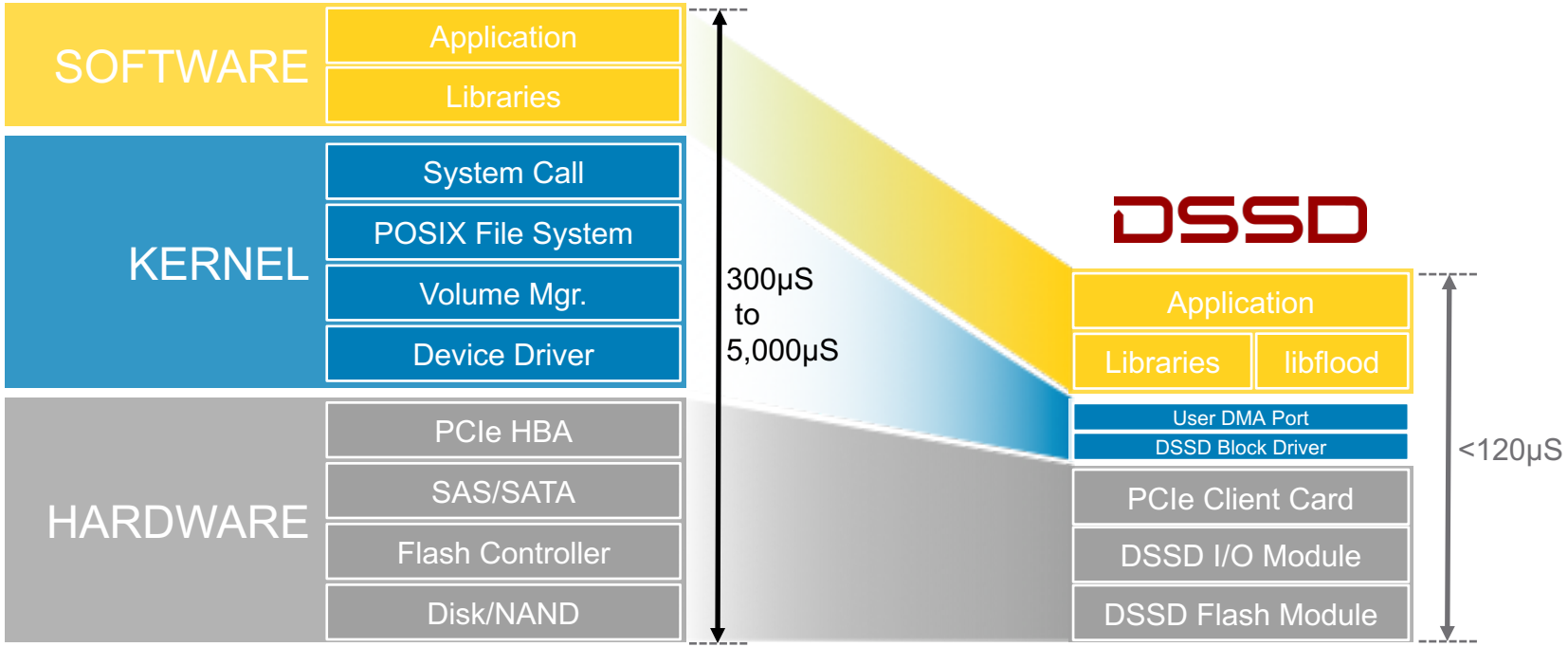
Legacy network connections reduce performance



The performance is even worse if an Ethernet, IB or FC network is in the middle

DSSD block device access to DSSD

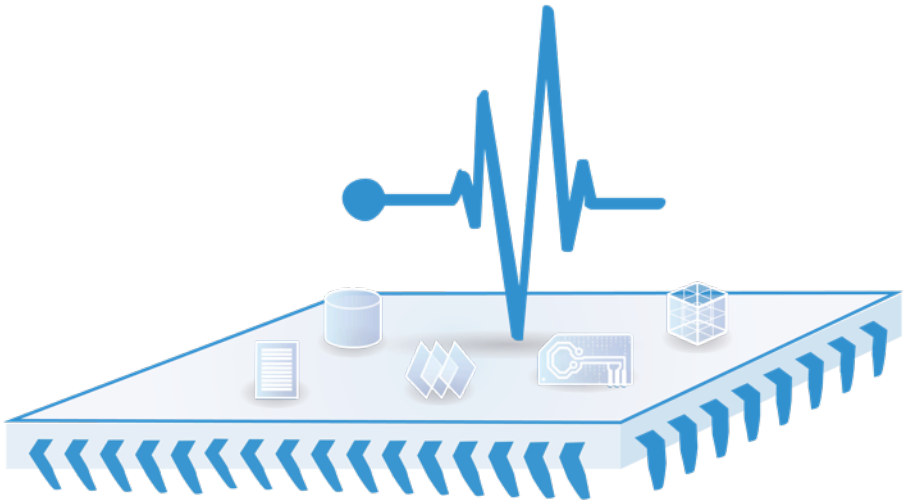
A bit more latency due to kernel overhead



DSSD

HardWare + Software
Data Protection

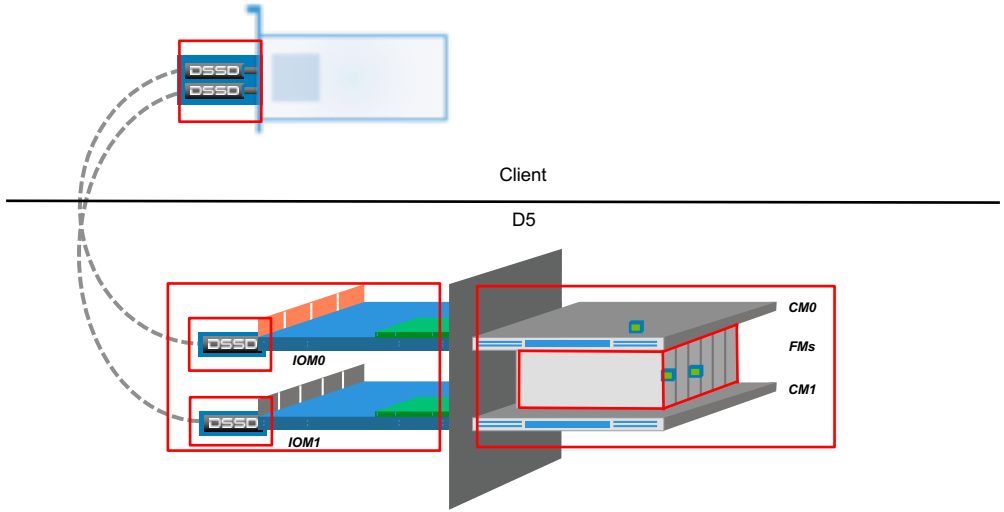
OBJECTIVE:
Provide better protection with less overhead at flash speeds



DSSD hardware + software data protection

A system of resiliency features from the client to the flash

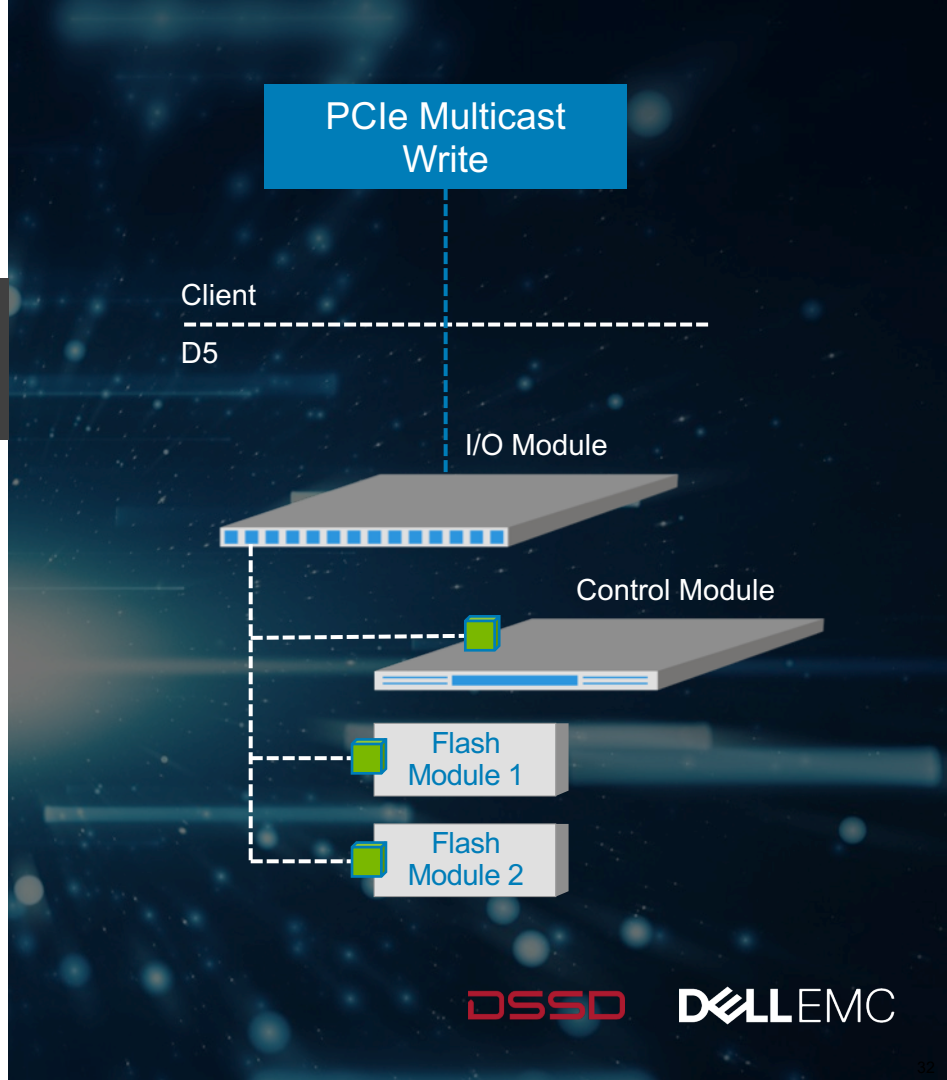
1. PCIe multicast write
2. Always-on multipathing
3. CRC for data in flight
4. Dual redundant hardware FRUs
5. Flash physics control
6. Space-time GC
7. Defect avoidance
8. Enterprise ECC
9. Always-on Cubic RAID
10. Vaulting
11. Dynamic overprovisioning
12. Resilvering



Hardware + Software Resilience

PCIe multicast write

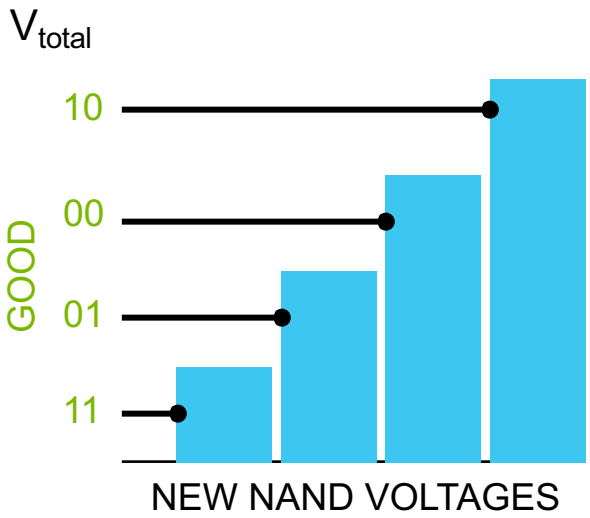
- A copy of each write ends up in 2 FMs and the CM DRAM
- A single write from the client-side application triggers a PCIe multicast write to 3 separate locations on the D5

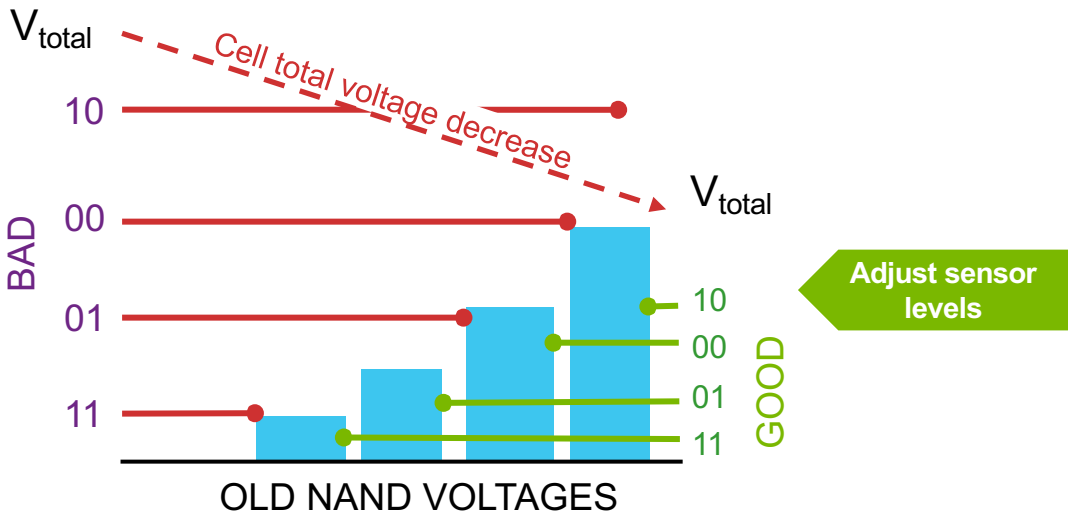


Hardware + software resilience

Flash physics control

DSSD software extends the life of NAND, improves the performance of NAND and increases the efficiency of operations performed on NAND





Hardware + software resilience

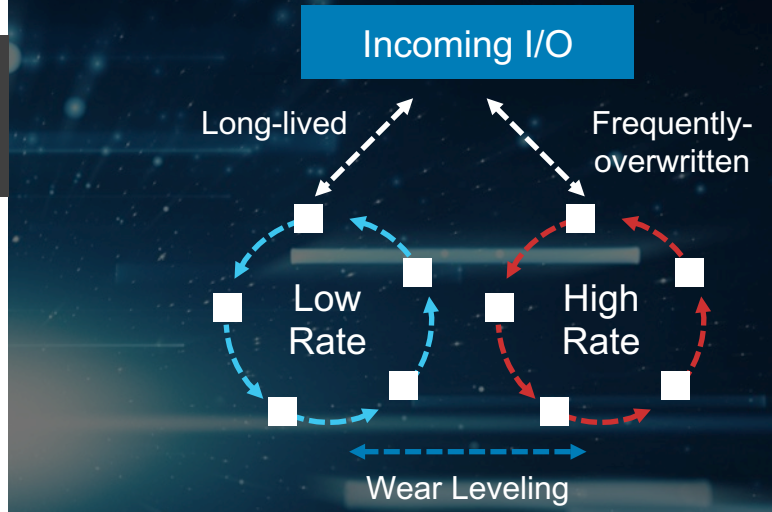
Space-time garbage collection

Space: Improves on generic GC

- Reduced write amplification through fragment level GC analysis
- GC only valid data, maximum space efficiency

Time: Continuously segregates data by observed lifetime:

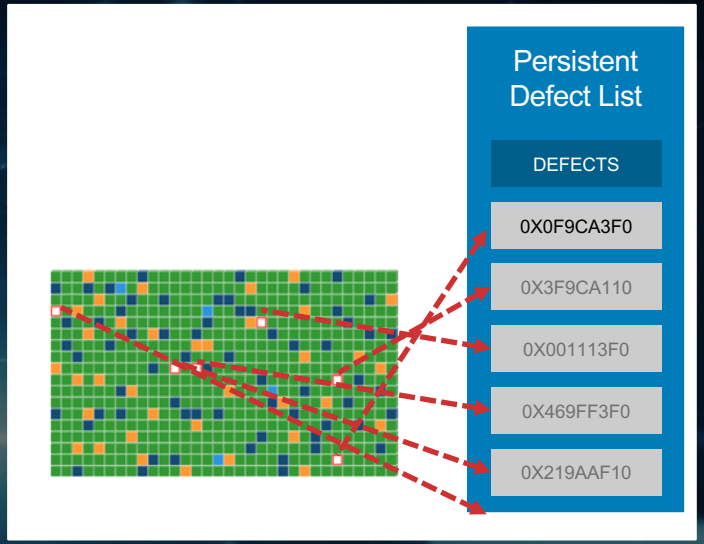
- Active data (frequently-overwritten)
- Stable data (long-lived)
- Active and stable locations swapped over time by GC for data retention and wear leveling



Hardware + software resilience

Defect avoidance

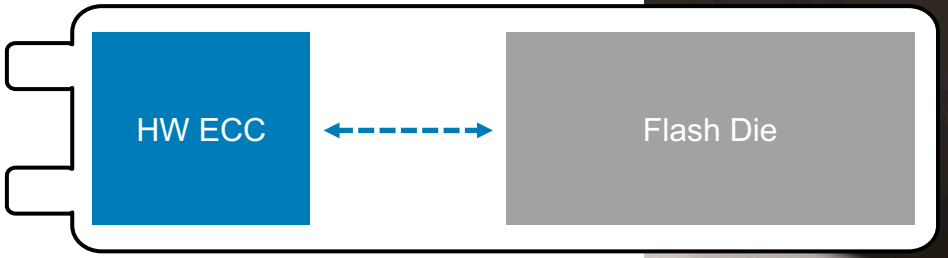
- Flash die found defective are marked and address is stored persistently within Flood metadata structures on the flash itself



Flash hardware resilience

DSSD enterprise ECC

- Flash die found defective are marked and address is stored persistently within Flood metadata structures on the flash itself

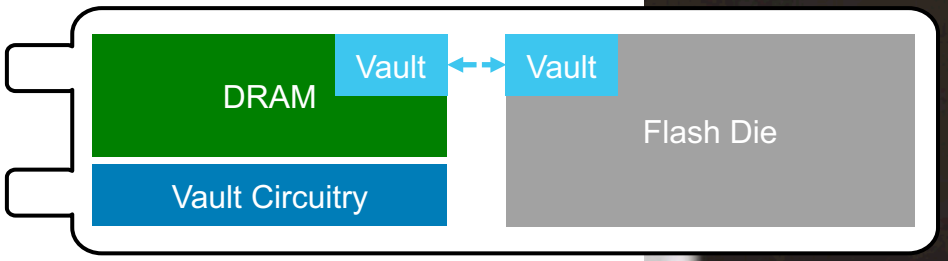


36

Flash hardware resilience

Vaulting

- Flash die set aside as recovery area in case of FM power loss



37

Hardware + software resilience

Dynamic overprovisioning

- Flash die set aside in case of other FM failures and for performance

Hardware + software resilience

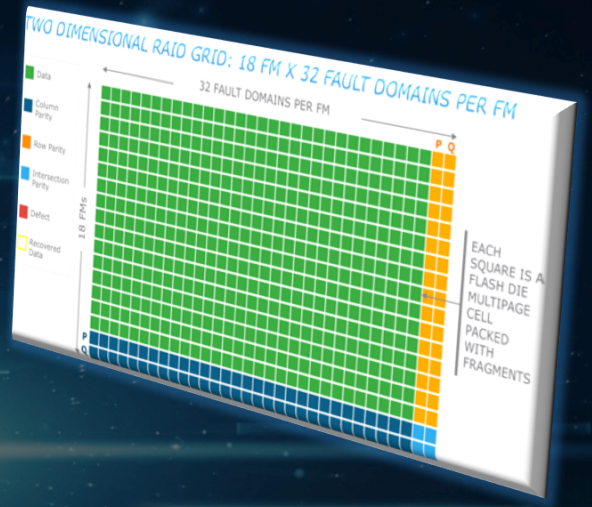
Resilvering

- Automatically moving data onto a new FM after replacement

Hardware + software resilience

Always on Cubic Raid

- Cubic RAID has 2x greater reliability of other RAID but has the same overhead (17%)
- Cubic RAID grid is an interlocked, multi-dimensional array of multi-page “cells” of NAND die
- High performance – always on



System Wide Data Protection


DELL EMC