SDC 18
SNIA INDIA
May 24-25, 2018
Bangalore, India
STORAGE DEVELOPER CONFERENCE

Distributed System with Erasure Coding and NVMeoF

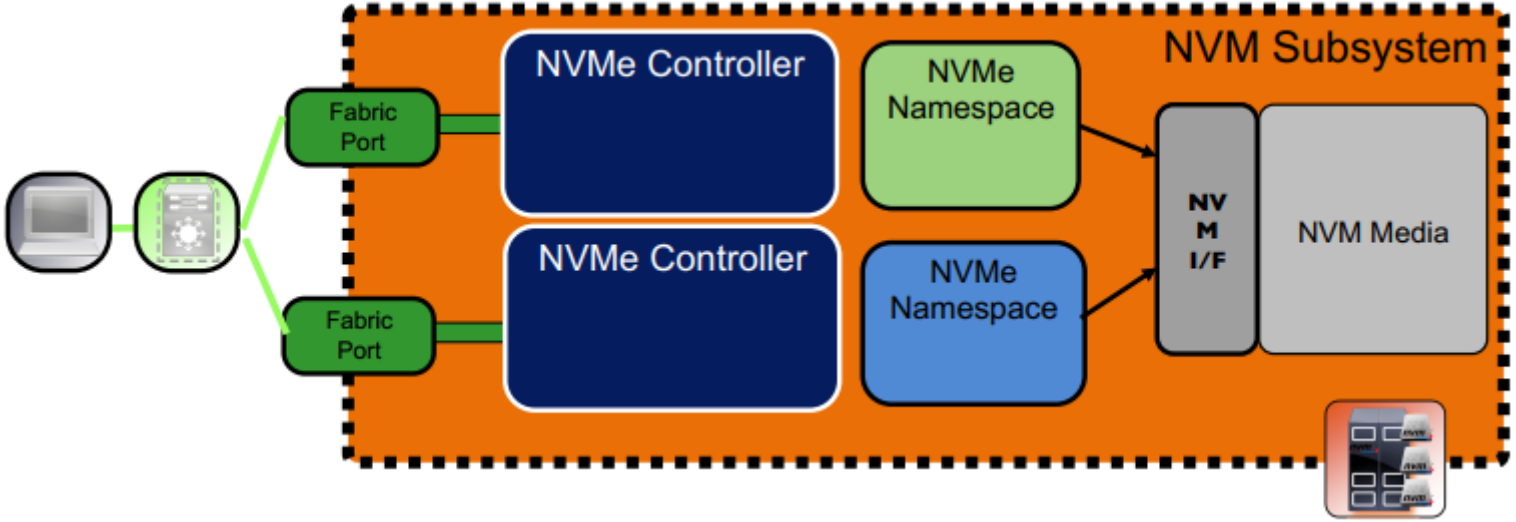
Suman Debnath
Toshiba

Agenda

- Introduction to NVMe Over Fabrics
- NVMe Over Fabrics Architecture
- Introduction to Erasure Coding
- NVMeOF with Erasure Coding for Distributed Systems

Introduction to NVMe Over Fabrics

- *An NVM subsystem includes one or more controllers, one or more namespaces, one or more PCI Express ports, a non-volatile memory storage medium, and an interface between the controller(s) and non-volatile memory storage medium*."

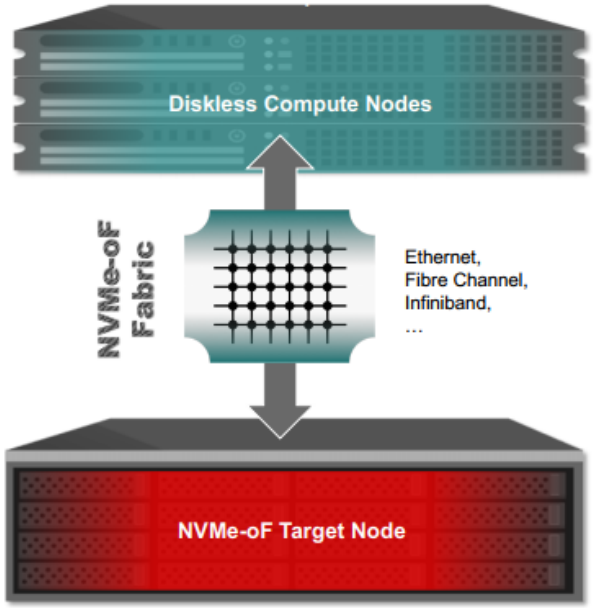


What is NVMeoF ?

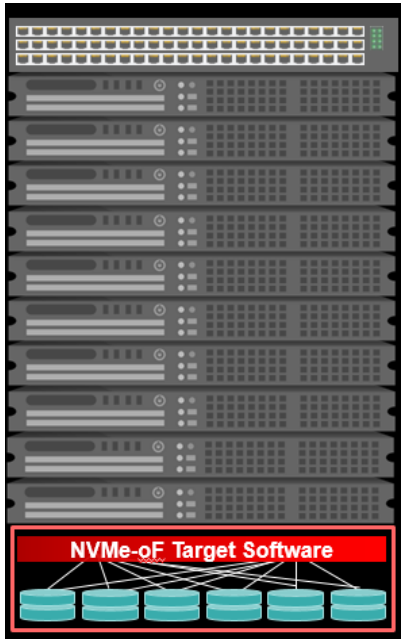
Connects compute nodes to NVMe storage across the datacenter network

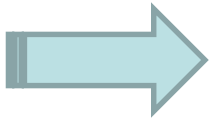
Preserves the performance and low latency of *native* NVMe

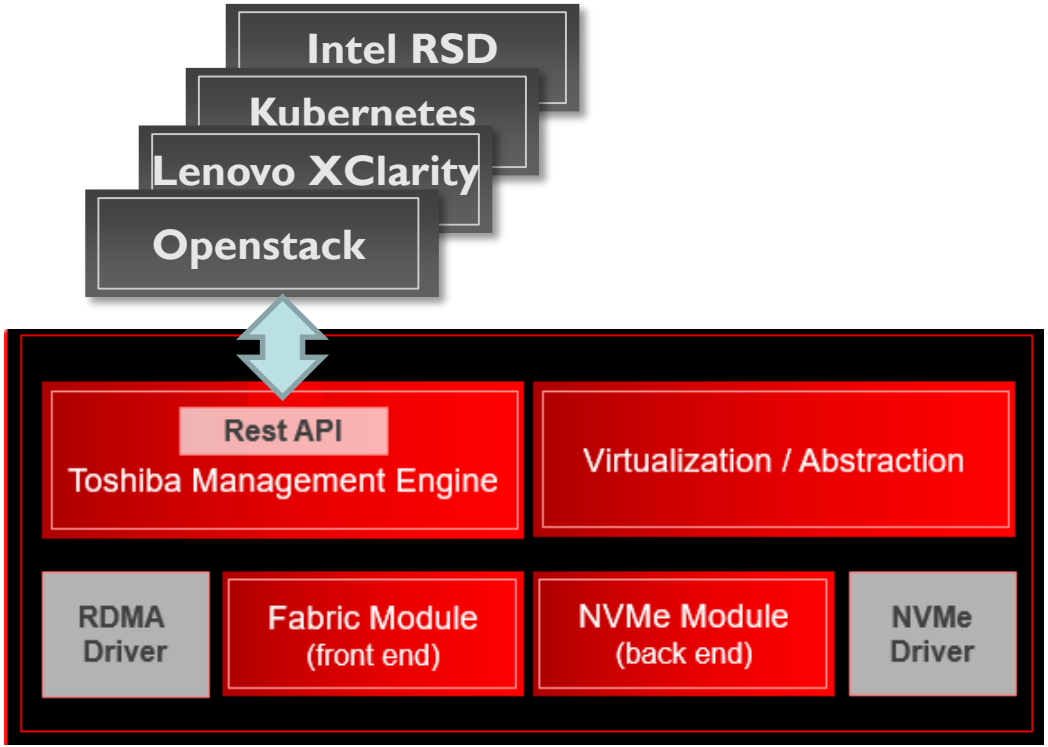
Uses remote direct memory access (RDMA), with bindings for several transport protocols



NVMeoF Software Abstraction

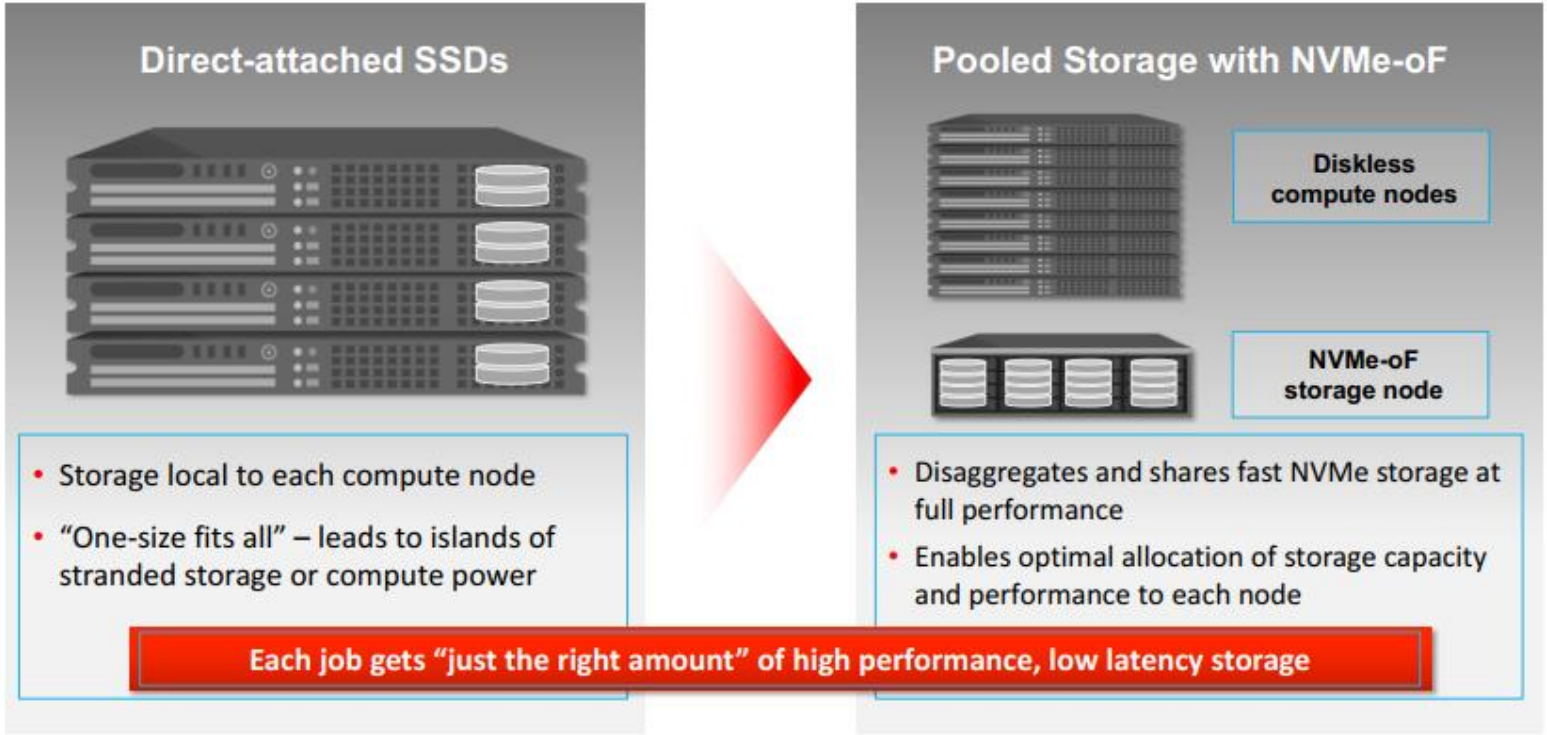






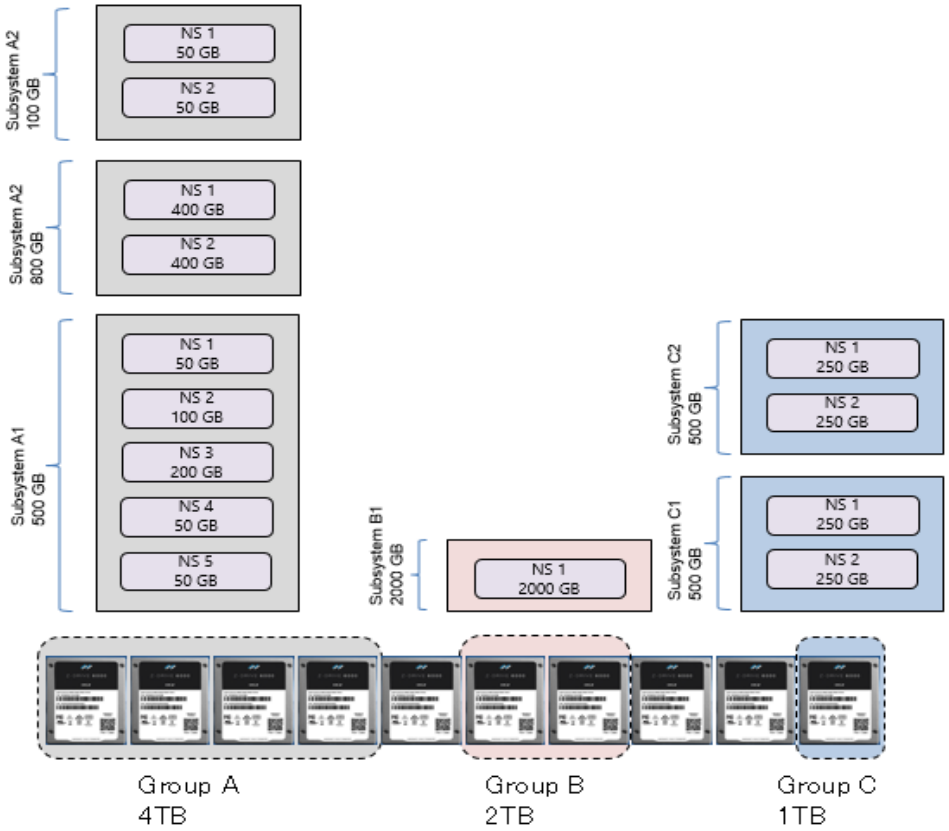
SDC 18
SNIA INDIA

NVMef Storage Abstraction



SDC 18
SNIA INDIA

Virtual Subsystem and Namespaces



So where we stand now?

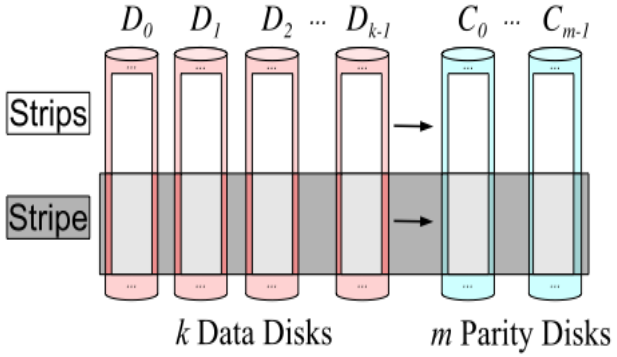
- *We can have a centralized super fast NVMe based storage*
- *Software abstraction makes the utilization of storage more efficient*
- *The storage can be accessed over the fabrics via RoCE v2 (RDMA) over a high speed transport (100Gbps and even more)*
- *Storage provisioning and management can be done via REST API*
- *So, now that the storage is provisioned to Hosts, we can leverage technology like Erasure Coding which may of the Distributed OS supports these days to make the best use of backend performance over the network as well as the goodness of fault tolerance*

Erasure Coding (First Blush) – Traditionally

Putting it Together

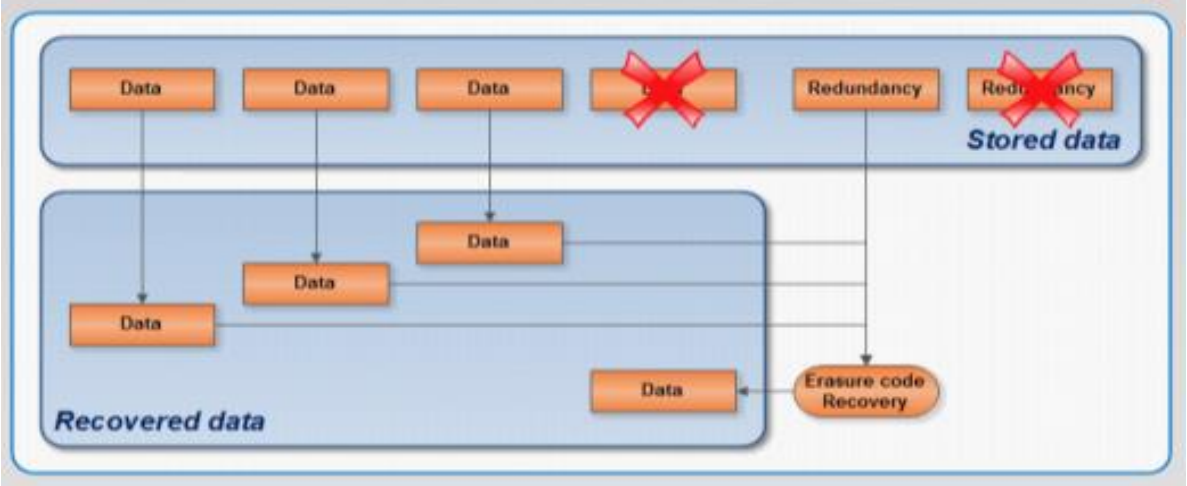
How it works

- Erasure Code
  - Store “k” disk worth of data in “n” total required disk
    - (n > k)
- “m” redundant disk
  - Pick “m” to choose the failure tolerance
    - (n = k + m)
- A generalization of RAID6
- Distributed across the nodes



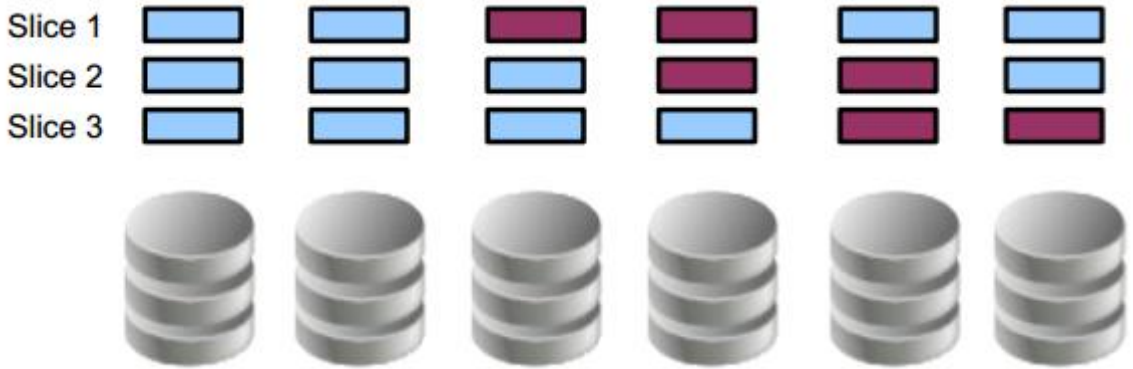
Erasure Code Algorithm

- Based on erasure codes
  - Fast implementation of the Rabin IDA (Information Dispersal Algorithm)
  - m additional fragments are computed from a set of n– k data fragments
  - Any data fragment can be recovered from any subset of n – k fragments (data or redundancy)



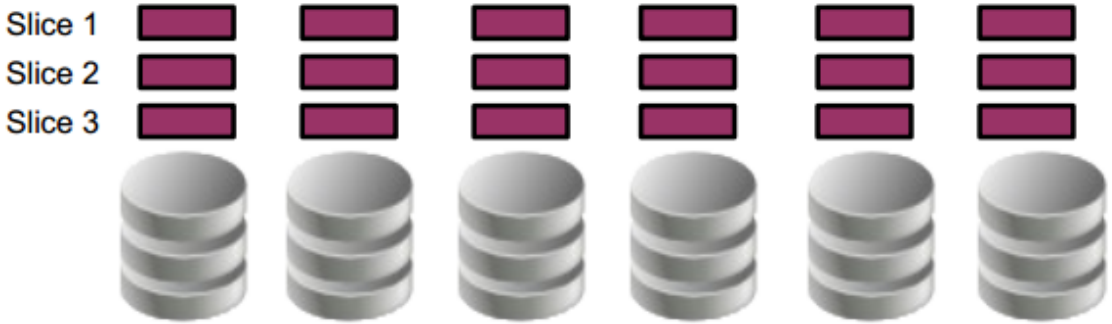
Systematic

- “k” data chunks, “m” coding chunks
  - Can stripe parity and data on the same disk
- Reads are simple, only decode in case of failure/repair



Non-Systematic

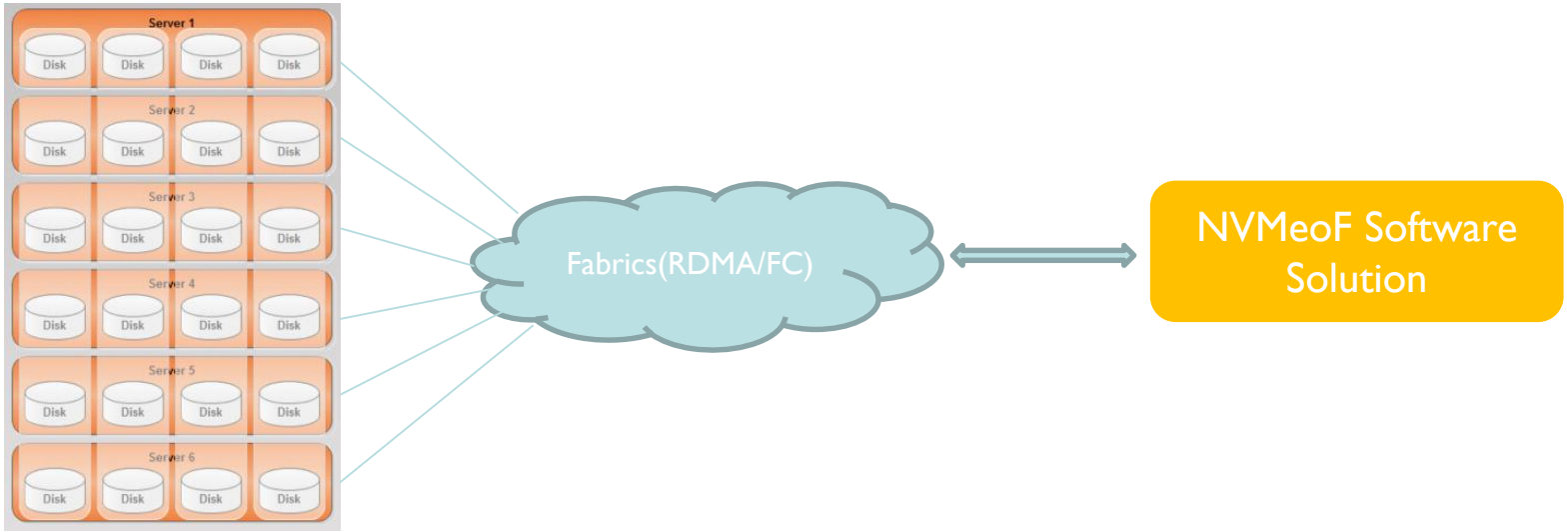
- All “n” chunks in a stripe are coded
- Encode/Decode for each read/write



Implementation of Erasure Coding

- Distributed Operating System
  - HDFS
  - Ceph
  - GlusterFS (Dispersed Volume)
  - Etc.

NVMeoF with Erasure Coding



Advantage of NVMeOF with EC

- <Need to add the pointers>