Abstract

1. Problem: Almost an order of magnitude bandwidth gap between storage devices and host
2. Approach: Mitigate gap by offloading storage services to storage devices
3. North-south offloading: includes translation, compaction, scrubbing, but not redundancy. Redundancy performance (including failure recovery) is sensitive to host bottleneck
4. East-west offloading: includes redundancy as a service of a collection of devices
5. Open-source research, software, and standards effort: Based on experience with peering in Ceph RADOS, investigate storage device designs that provide useful behaviors collectively.

Definition

**eu·so·cial** /ˈjoo-sō-kəl/
noun
1. Biological feature where the group is composed of more than one genotype, as a result of reproductive castes or inbreeding within a single genotype

**eu·so·cial stor·age** /ˈjoo-sō-kəl stōr-ij/
noun
1. A software abstraction

A software abstraction

- Standardized Object Protocol
  - Network/Fabric based
  - Disaggregates
  - Mechanism based
  - Policy is configured
  - Cluster Operations
  - Client/Peer Operations
  - Peer 2 Peer Operations
  - Control Operations
  - Configuration aware
  - Data Integrity Mechanisms
  - Tiering Mechanisms
  - Improved Failure Domains
  - Improved Placement/Rebalancing
  - Could support In-Store Compute

Hardware

- Highly optimized for media type
  - Can be any combination of HW
  - Could be:
    - An Ethernet-connected SSD
    - Small servers and HDDs
    - Gateway to S3
  - Must be network attached media
  - Must support bi-directional comm
  - Public/private paths
  - No restrictions
  - Media type
  - Form factor
  - Capacity
  - Components
  - Fabric type

Roadmap

1. North-south Media Mgmt Offloading
   - Example: replication/encryption, coding, failure management, makes host media management obsolete
   - Requires East-West communication

2. East-west Media Mgmt Offloading
   - With Disaggregation
   - Example: NVMe key/value abstraction, SSD level NVMe over fabrics

3. In-store Compute
   - Could support In-store Compute

Bandwidth Mismatch

**Storage Media**

- Network Object Protocol
- Client Object OPs
- Network Object Protocol
- Peer 2 Peer OPs

**Media management**

- Data Integrity
- Caste Hopping
- In-Store Compute

**Eusocial Storage Devices**

Carlos Maltzahn (UC Santa Cruz), Shingo Tanaka (Toshiba Memory), Christine Green (Seagate), Grant Mackey (Western Digital), Philip Kufeldt (Independent), Timothy Feldman (Seagate), Christine Green (Seagate), Grant Mackey (Western Digital), Carlos Maltzahn (UC Santa Cruz), Shingo Tanaka (Toshiba Memory)

Contact:

- Philip Kufeldt: phil@protium.com
- Carlos Maltzahn: cm@baskin.ucsc.edu

Center for Research in Open Source Software
cross.ucsc.edu