Eusocial Storage



CROSS | CENTER FOR RESEARCH IN OPEN SOURCE SOFTWARE

FAST'18

Baskin Engineering
UC SANTA CRUZ

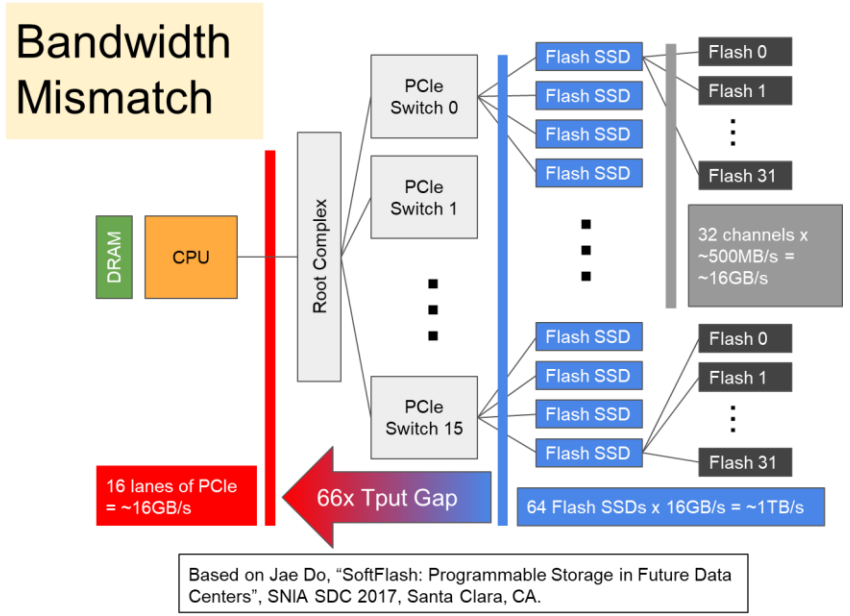
Problem

Server Bandwidth Mismatch

- Flash and Network BW catching up and will surpass DDR BW

Storage IO

- Application Requirements
  - North and South (N-S)
    - Real work: Primary data path
- Media Management Requirements
  - North and South
    - Translation, Compaction, Deduplication, Scrubbing
  - East and West (E-W)
    - Redundancy, Recovery, Rebalancing, Tiering and Caching



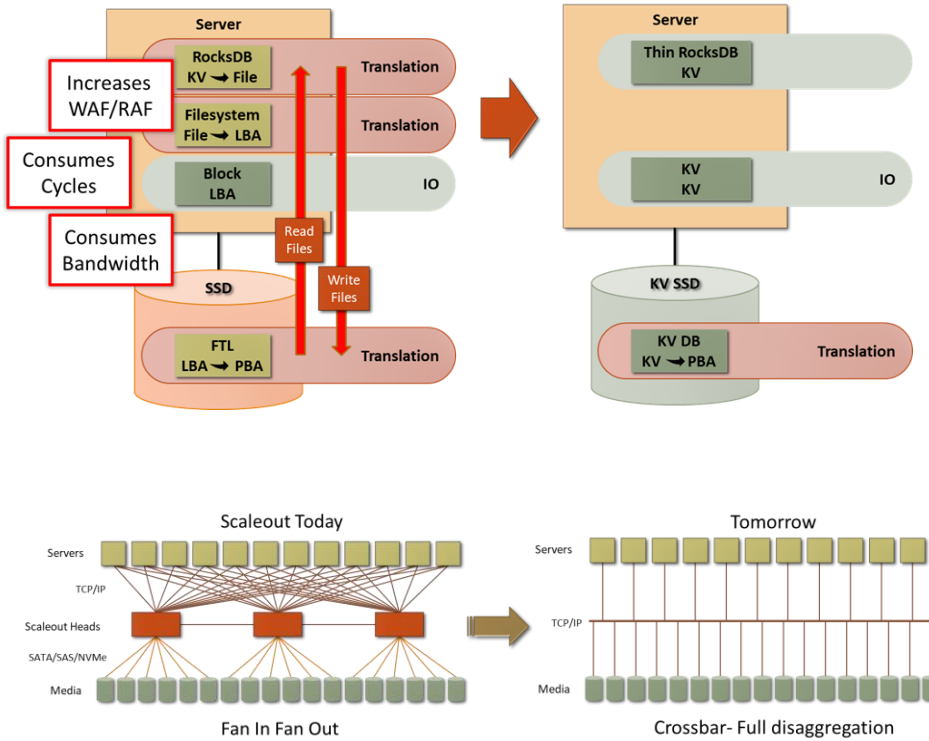
CPU Bandwidth – The Worrisome 2020 Trend, Fritz Kruger, Mar 2016

Offloading

North and south media management
- Push translations, compaction, deduplication, scrubbing down
- Efforts to offload already occurring, e.g. DevFS, NVMe KV
- Still leaves E-W media management

Moving towards a disaggregated world
- From “fan in fan out” to full crossbar
- Permits the removal of servers from IO path
- Need smarter devices
  - Head towards in store compute capabilities

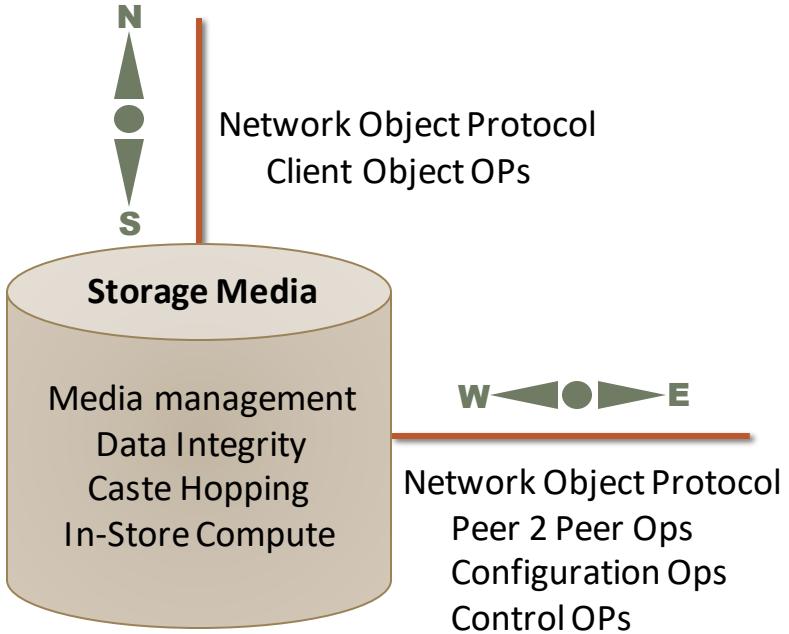
East and West media management
- Leverage disaggregated storage elements to do E-W
  - Redundancy, Recovery, Rebalancing, Tiering and Caching



What is Eusocial Storage

A software abstraction

- Standardized Object Protocol
  - Network/Fabric based
    - Disaggregates
  - Mechanism based
    - Policy is configured
  - Cluster Operations
  - Client Object Operations
  - Peer 2 Peer Operations
  - Control Operations
- Configuration aware
- Data Integrity Mechanisms
- Tiering/Caching 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 a
    - An ethernet connected SSD
    - Small server and HDDs
    - Gateway to S3
- Must be network attached media
  - Must support bi-directional communications
  - Public/private paths recommended
- No restrictions
  - Media type
  - Form factor
  - Capacity
  - Components
  - Fabric type

Eusocial Storage

Storage Media

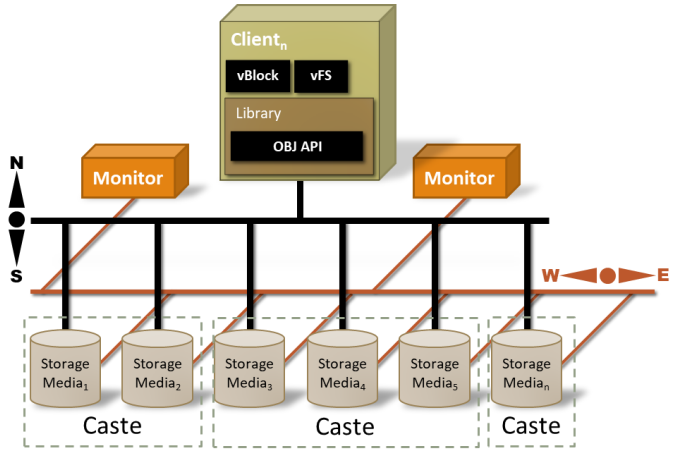
- Highly optimized, autonomous units of storage
- Define lines of service
  - Throughput, latency, media type, compute availability

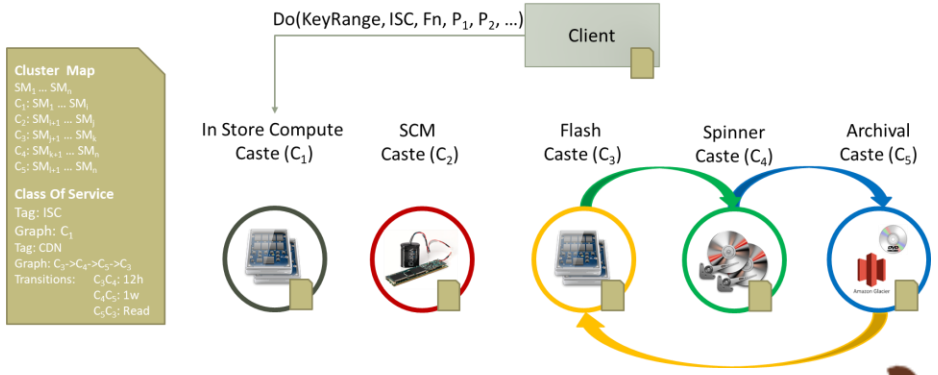
Castes

- Groups of Storage Media that provide similar lines of service
  - Similar media, throughput, latency
  - Similar functionality, each member is a replacement for the others
- Permit the scaling of a line of service
- Defines availability
  - Eraser coded caste, replicated caste

Cluster

- Defines all Storage Media and Castes
- Manages events and cluster configuration (cluster map)
- Can be hierarchically managed on global and caste levels





https://cross.ucsc.edu/

Work with Us – Join

UCSC Center for Research in Open Source Software

THANK YOU

