panasas
Beyond Storage

Architecture of a Next-Generation Object Storage Device in the Panasas Filesystem

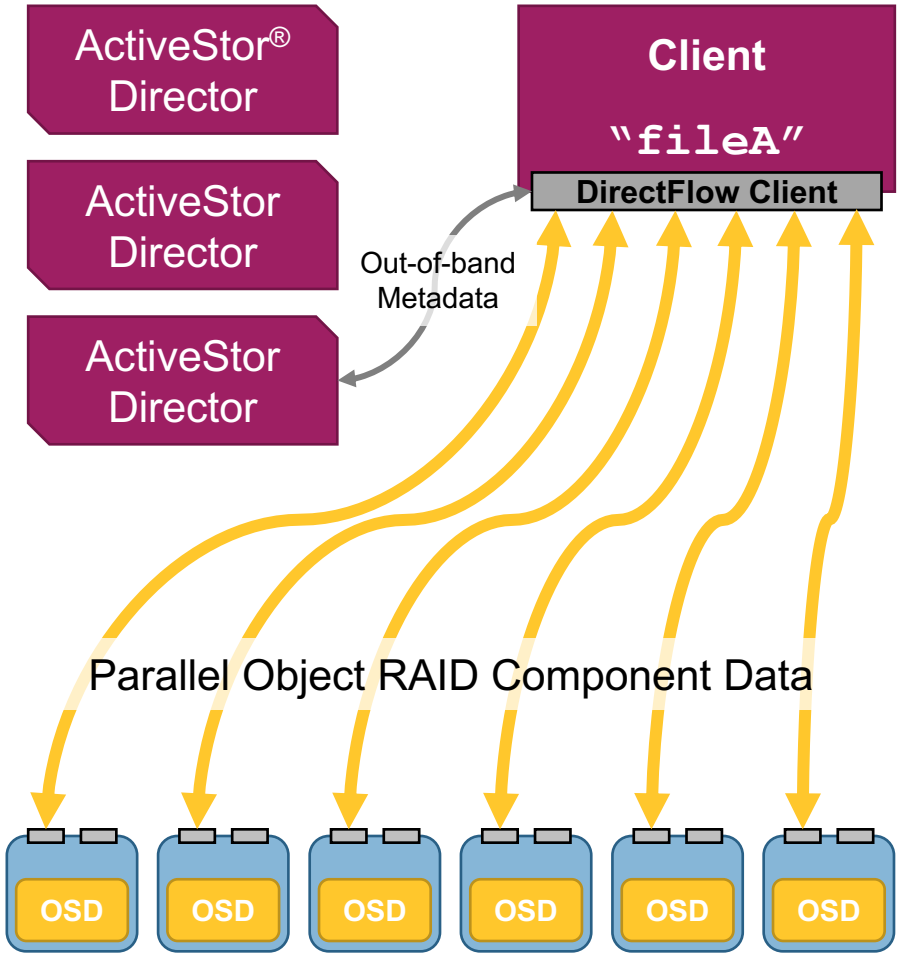
Computing Insight UK 2018
December 13, 2018 - Manchester, UK

Curtis Anderson
Software Architect - Panasas, Inc.

www.Panasas.com

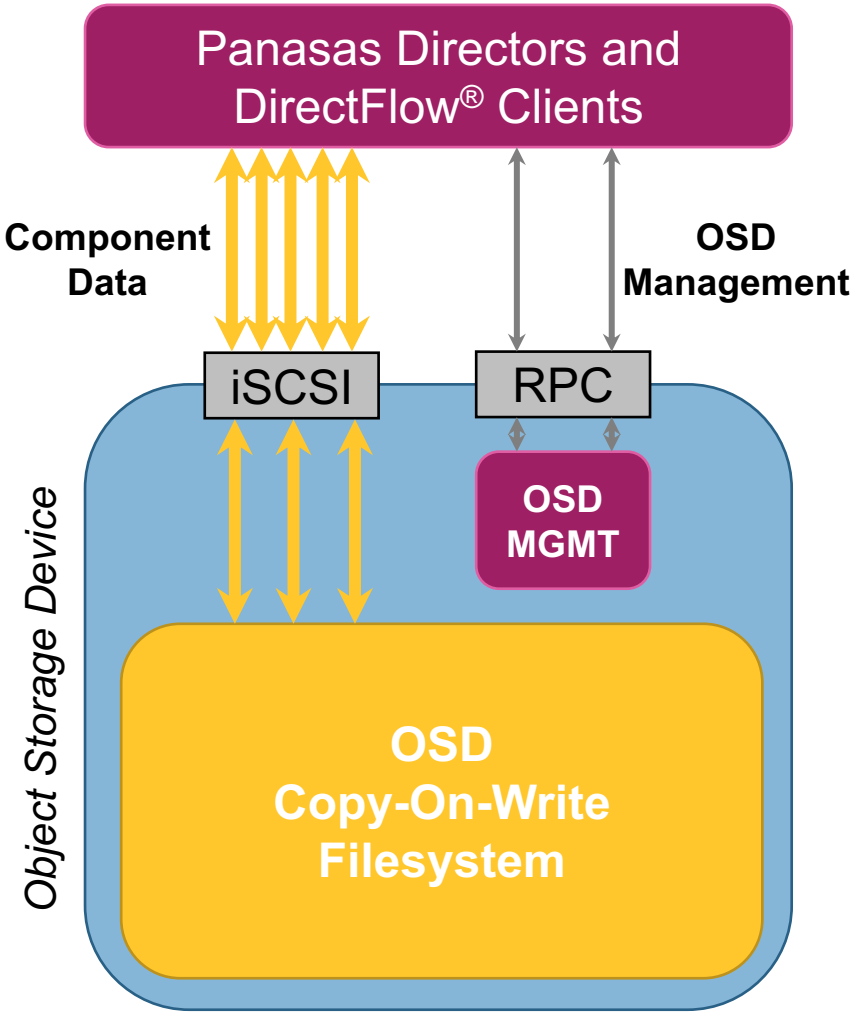
What's an Object Storage Device (OSD)?

- Out-of-band metadata management on Directors
  - B/W and IOPs scales linearly with the number of OSDs
- OSDs are our main data storage targets
  - Clients transfer data directly to/from OSDs in parallel
    - After communicating with Director(s) for metadata
  - Per Realm: 3-100s of Directors, 10s-1000s of OSDs
- OSDs enable Erasure Coded RAID per File
  - Each file is striped across Component Objects (COs)
  - N+2 Erasure Codes are just additional COs
  - At most one CO from a file on any given OSD
- OSDs all help in scale-out reconstruction
  - All *N+1* OSDs are reading and writing during rebuild
    - Traditional RAID limited by B/W of replacement drive
  - Faster recon times result in higher data reliability



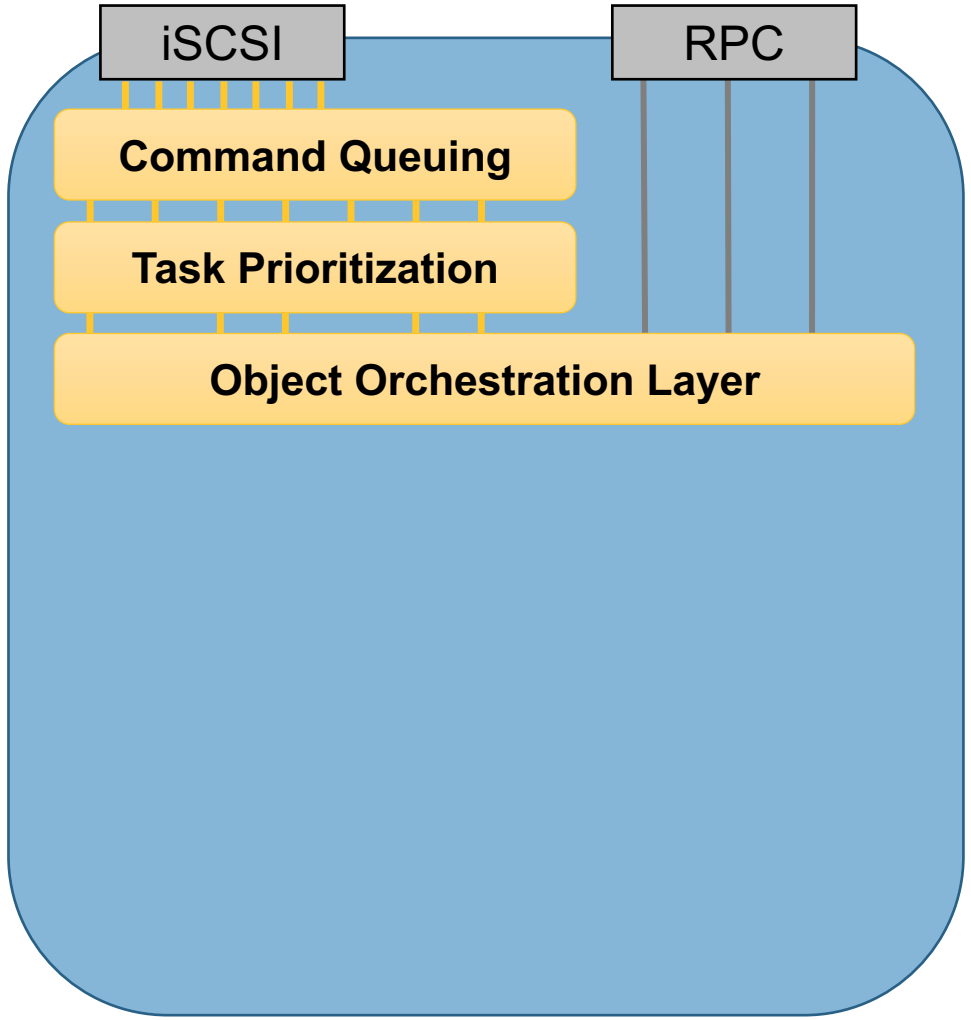
What's Inside an OSD?

- OSDs expose iSCSI and RPC interfaces
  - iSCSI for data transport to/from our client S/W & Director
  - RPC for health and mgmt. to/from our Director
- OSDs have an internal F/S to manage devices
  - On-disk formats, head scheduling, transactional updates
  - Internal details transparent to the rest of the architecture
- OSDs each do their own used/free space mgmt.
  - Only the OSD knows logical-to-physical placement
  - Internal details transparent to the rest of the architecture
- OSDs each do their own COW-based snapshots
  - Only the OSD knows which bytes have been COWed
  - Internal details transparent to the rest of the architecture



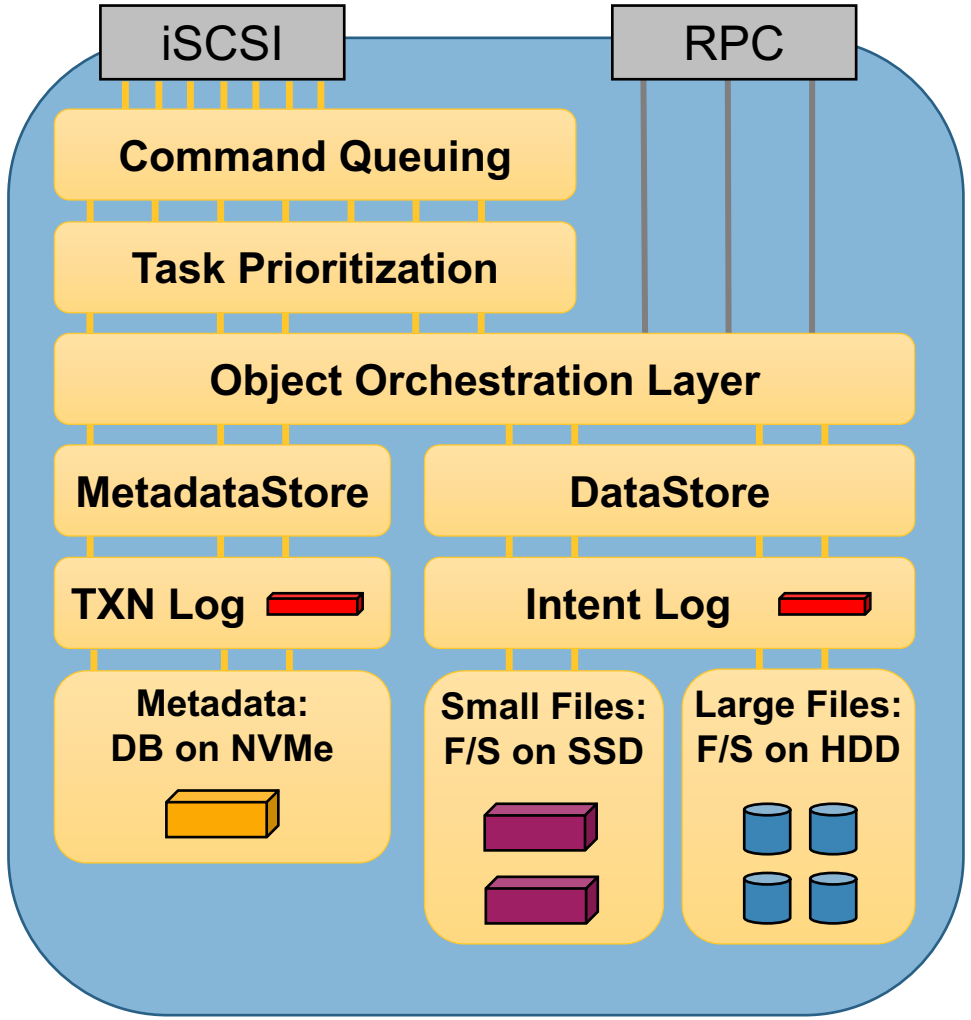
What is ActiveStor® Ultra?

- Unchanged OSD external API (in 1st release)
  - Risk-reduction to “change one thing at a time”
  - Looks & acts the same, but runs much faster
  - PanFS® taught to use new OSD capabilities in future
- Uses a high-performance COTS platform
  - NVDIMMs for power-fail instead of built-in UPS
  - Choose dual 25GbE or InfiniBand ports
- Enables ‘wider’ and ‘taller’ OSDs
  - Adapt to more HDDs, and more performance tiers
  - Adapt to different ratios in the tiers for new workloads
- Modular S/W design entirely in user space
  - Running on Linux, leveraging Open Source packages



What is ActiveStor Ultra?

- Optimize storage/handling for metadata/data
  - **Transaction Logs:** in NVDIMM
    - Very fast and power-safe transaction completion
  - **DRAM Cache:** access to unmodified data/metadata
  - **Metadata:** in database or KVS on NVMe SSD
    - Fast transactions, consistent performance, intelligent queries
    - DB may be used for Map/Reduce data analytics in the future
  - **Small Files:** in COW F/S on SATA SSD
    - Cost-effective high-IOPs, consistent performance
  - **Large Files:** in COW F/S on SATA HDD
    - HDDs are good at delivering B/W if they only store large files
- Full data stability with fully async performance
  - NVDIMM is intent-log for data & metadata operations
    - Intent-log is layered above the COW F/S and the DB
    - Will re-execute operations in the event of an interruption
  - Allows COW F/S & DB to run full async for best perf
    - e.g. coalesce writes into contiguous runs for later read-back perf

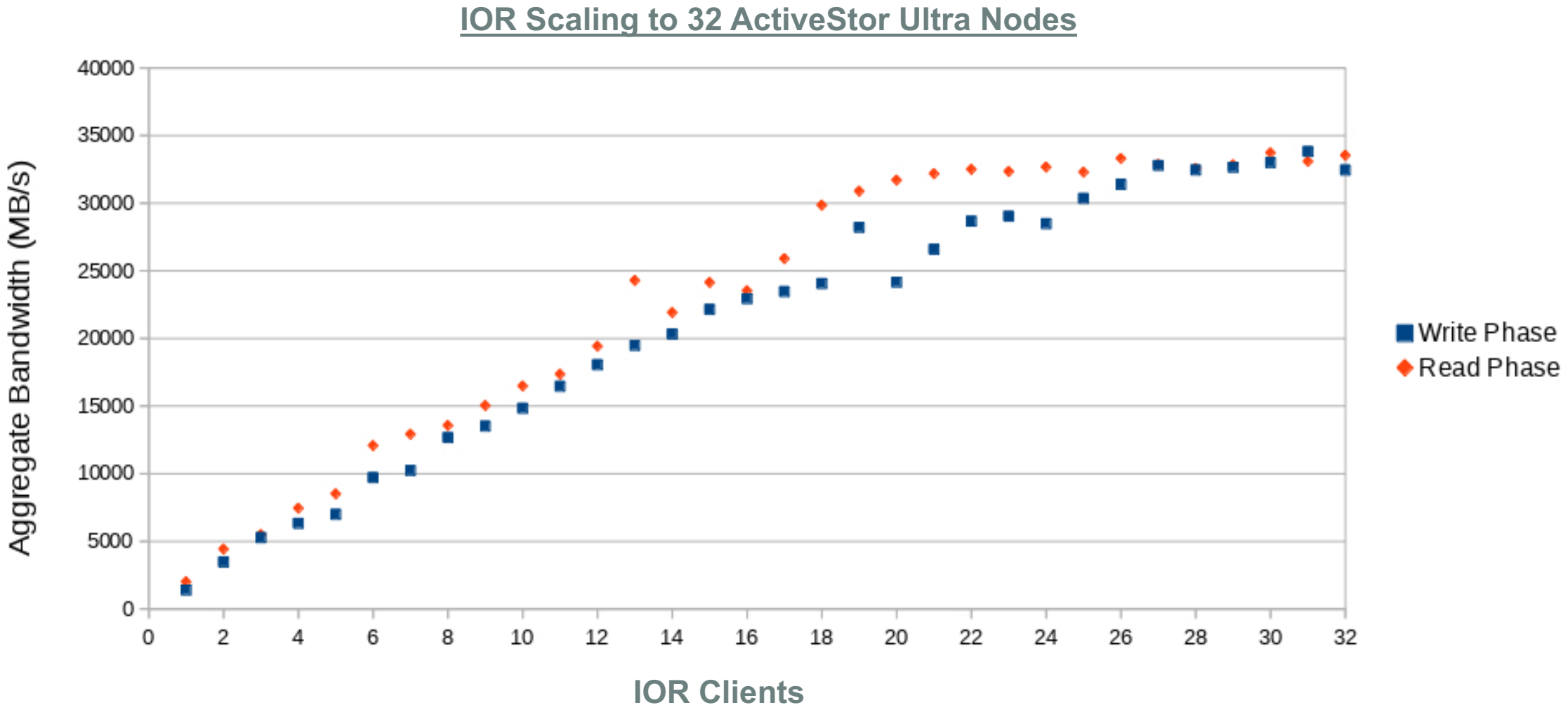


IOR Scaling Experiment Setup

- 1-32 DirectFlow Clients
  - 10 Core / 2x10GbE / 32GB RAM
- 32 OSDs in ActiveStor Ultra Realm
  - Each populated with 8x4TB drives
  - 2x10GbE / 16GB NVDIMM / 32GB RAM
- Single Volume Used
- RAID6+ Erasure Coding

- IOR: Scale Number of Clients
  - 32 tasks executed on every node, each writing or reading their own 10GB file
  - Demonstrates ability to scale I/O effectively without wild fluctuations
  - Full unmounts of DF client performed between write and read – barriers included

Initial Untuned Performance Scaling Results



ActiveStor Ultra Summary

panasas
Beyond Storage

- Starts a new era of innovation and performance for PanFS and Panasas
  - **Higher Performance:** Novel algorithms and intelligent use of the right storage media
  - **Consistent Performance:** Novel algorithms and intelligent use of the right storage media
  - **Wider Choice of Platforms:** Fully decoupled OSD software from the OS and hardware
  - **Latest Networking and Storage Media:** Adopting COTS platforms
  - **Higher Density per Rack:** Adopting COTS platforms
  - **Improved Snapshots:** Instantaneous, scalable snapshots with near-zero cooldown time
  - **Higher Feature Velocity:** Leveraging COTS + Open Source frees up resources
- COTS platforms allows us to focus on what we do best:

High-Performance Parallel Filesystems

Thank You!