Distributed Asynchronous Object Storage (DAOS)

Fault and Performance Domains

Kris Jacque

DAOS User Group 2023

Agenda

- The problem
- DAOS domains explained
- Configuration
- Using fault domains
- Using performance domains
- Future changes

Problem

- How to place objects to maximize:
  - Fault tolerance
    - Distribute data across hardware resources (nodes, racks, etc.)
  - Performance
    - Group certain nodes together based on performance characteristics of network

- DAOS needs to understand organization of the physical nodes

DAOS domain hierarchy

- System-defined layers
  - Rank (daos_engine)
  - Target
- User-defined layers
  - Examples:
    - Node/host (will be system-defined in future)
    - Rack
    - Network switch
    - Power source
    - Room
    - Performance groups

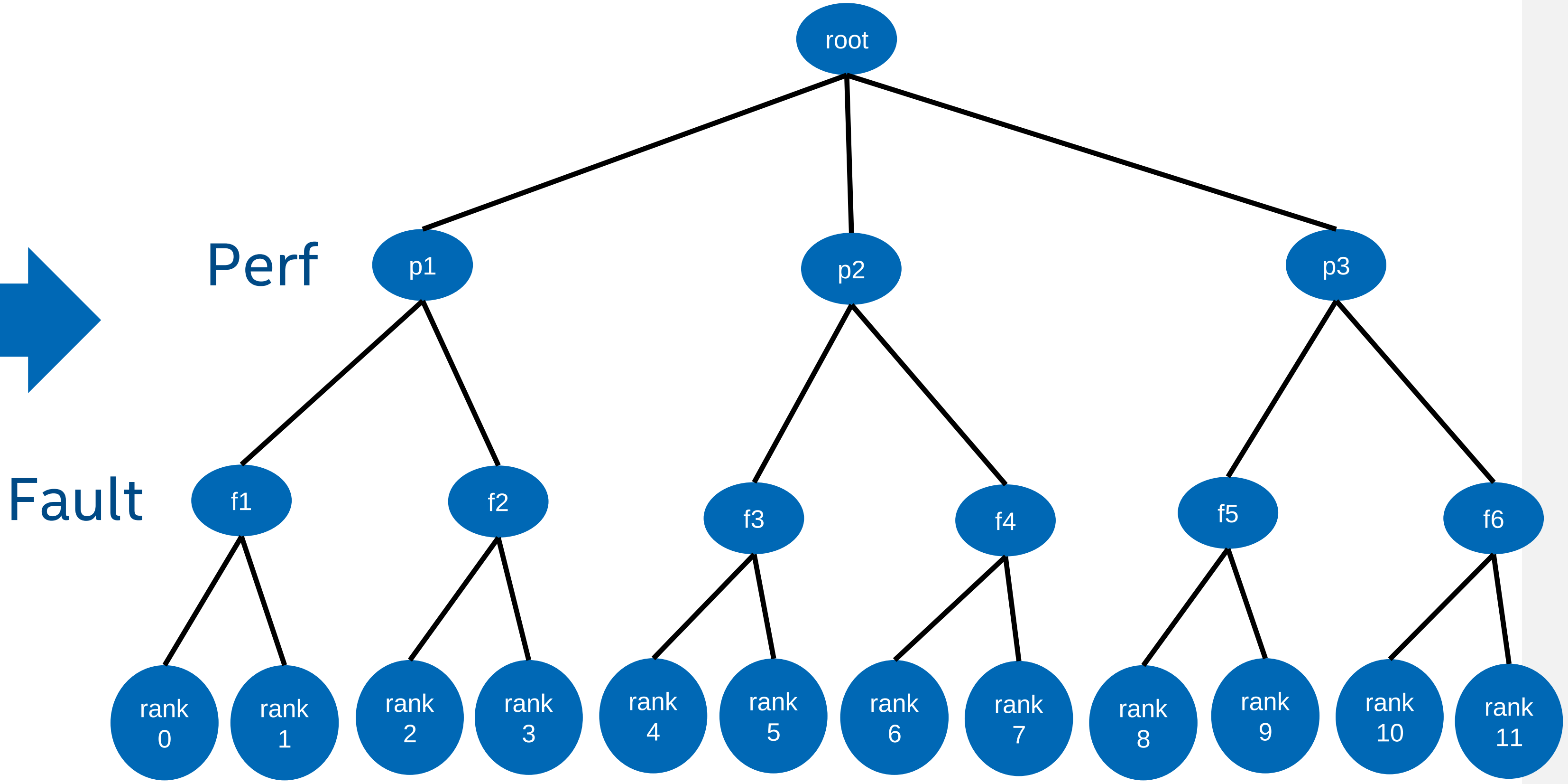
Building the domain tree

From each server:

fault_path: /p1/f1
fault_path: /p1/f2
fault_path: /p2/f3
fault_path: /p2/f4
…

OR

fault_cb: /etc/daos/fault.sh



Using fault domains

- Select redundancy level during container create:
  - rd_lvl=node (or user-defined fault domain)
  - rd_lvl=rank
- Default: node

Using performance domains

- Select performance domain level during pool and/or container create
  - perf_domain=root (e.g. whole system)
  - perf_domain=group (user-defined layer above fault domain)
  - Pool default: root
  - Container default: Inherit from pool
- Set performance domain affinity
  - rp_pda (replicated objects)
  - ec_pda (erasure coded objects)
  - Must be > 0
  - Lower value => more scattered

Performance domain affinity

- Higher values => keep shards in same domain
  - Best for objects with small number of replicas
  - Prioritize rebuild and server-to-server comms
  - Avoid extra network hops
- Lower values => spread shards over multiple domains
  - Best for large EC objects with many shards
  - Prioritize client bandwidth
  - Avoid bandwidth bottlenecks

Future improvements

- Arbitrary number of domain layers
- Configure performance domain to any layer
- System-defined node layer
- Intelligent rank selection in pool create
- User-defined layer naming in fault_path
  - Example:
    - fault_path: /cluster=“wolf”/psu=“psu100”/rack=“r130”/node=“wolf-123”

intel®