
whamcloud

whamcloud

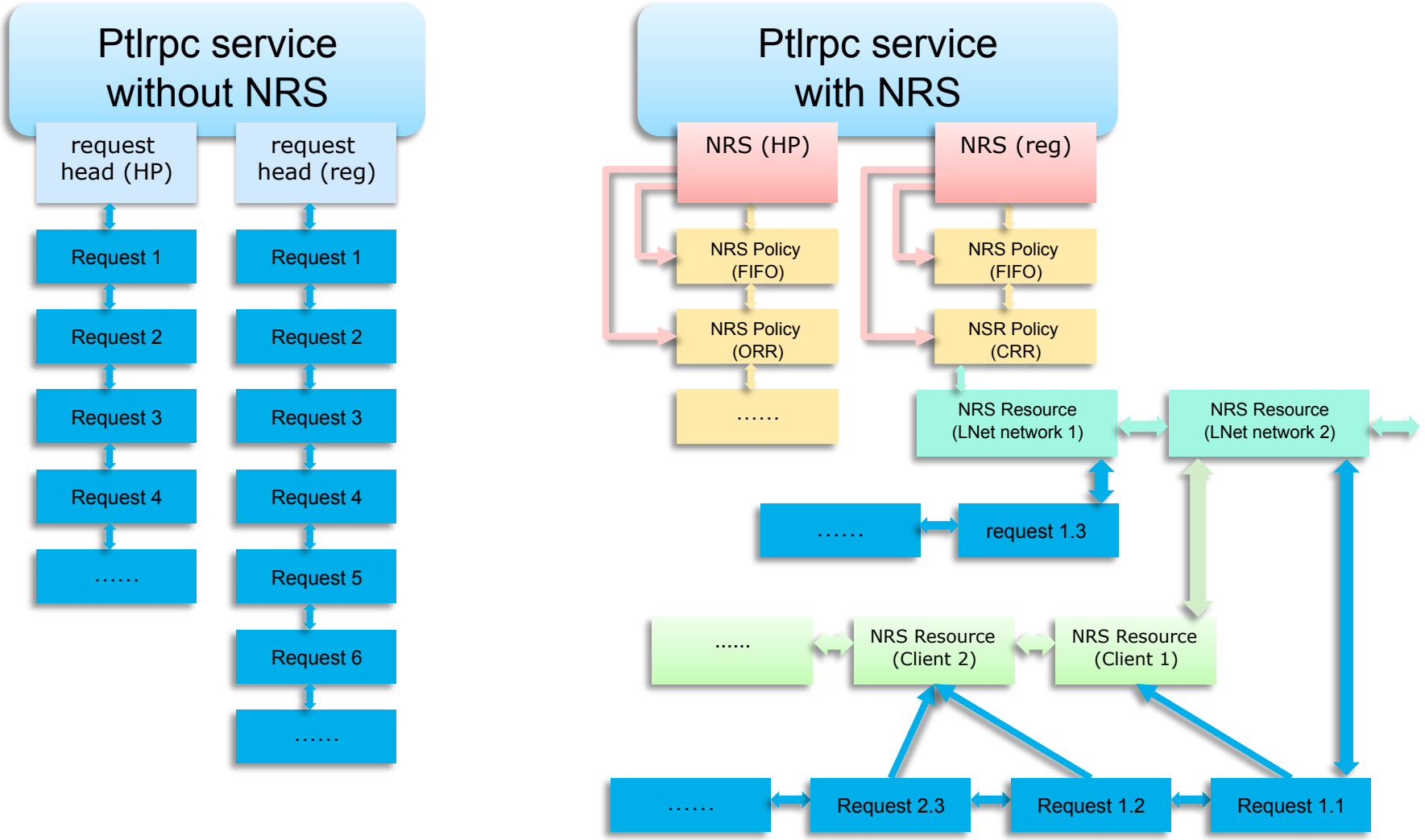
Lustre NRS (Network Request Scheduler)

- Liang Zhen
- liang@whamcloud.com

Why NRS

- Provide consistent performance by ordering request execution to avoid client starvation
- Balance workloads to backend storage
- Present a workload to the backend filesystem that can be optimized easily
- Job/network resource control

Ptlrpc service w/o and with NRS



Network Request Scheduler (NRS)

- Policy-driven request ordering
- Active policy
  - Attempt to schedule requests using this policy first
  - May allocate memory etc.
  - May fail
- Fallback policy
  - Used if active policy fails
  - Simple FIFO
  - May not fail
- Other policies
  - Exist transiently when active policy changed
  - Request dequeue is round-robin over all policies
  - Inactive policy removed when its last request dequeued
- Active policy selectable at runtime

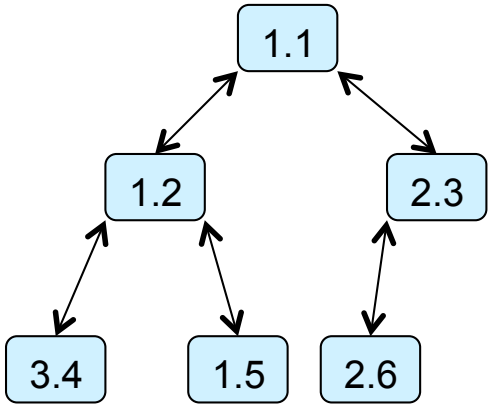
NRS Policy

- Implements abstract request operations
  - Initialize
    - Prepare request for enqueue
    - Identify resources to arbitrate between
    - Implementation may block and/or allocate memory
  - Enqueue
    - Adds a request to the policy's set of requests
    - Implementation may not block or allocate memory
  - Dequeue
    - Find/remove the policy's highest priority request
    - Take resource credits for the request
    - Implementation may not block or allocate memory
  - Finalize
    - Release resources taken by the request
    - Implementation may block and/or allocate memory
- Control interface to set properties at runtime
- Not aware of ptlrpc-service locks
  - Serialisation handled by callers

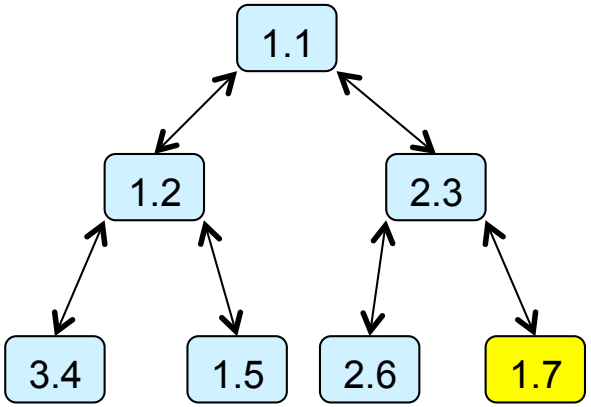
Other objects

- NRS request
  - Embedded in ptlrpc request
  - Properties
    - Resource credits taken by request
    - Priority values / resource references
- NRS resource
  - Placeholder for tracking/accounting resources tracked by NRS
  - Fast lookup (hash table)
  - Client Round Robin example: two-level resource tables
    - Round-robin between OSTs
    - Round-robin between clients on the same OST
- Binary Heap
  - Implements priority queue for NRS requests
  - Scales O(log n) with # queued requests
  - < 2uS insert + remove @ 10,000,000 requests

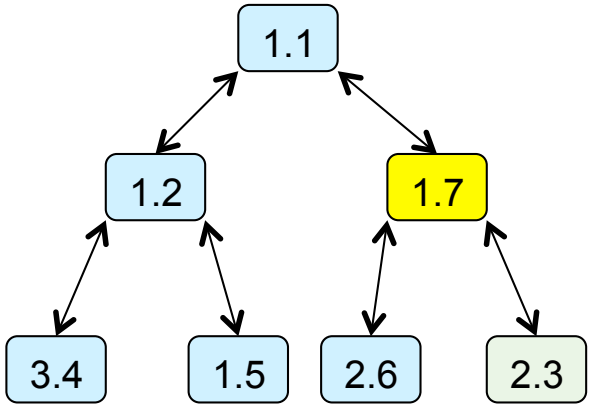
Sort prioritized requests by binheap



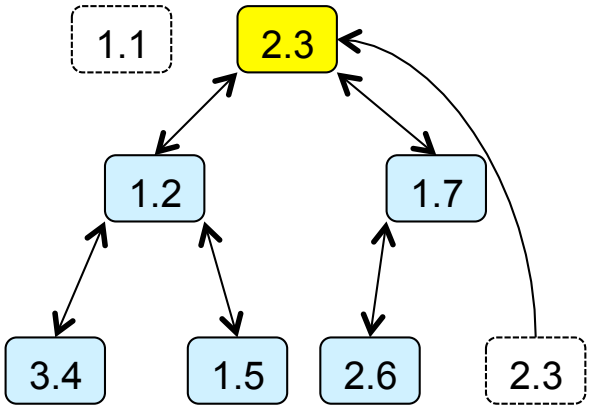
Add element step1



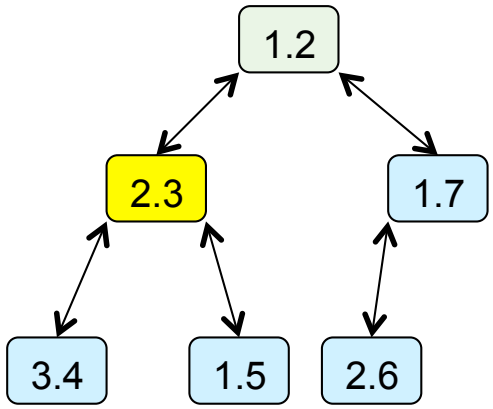
Add element step2



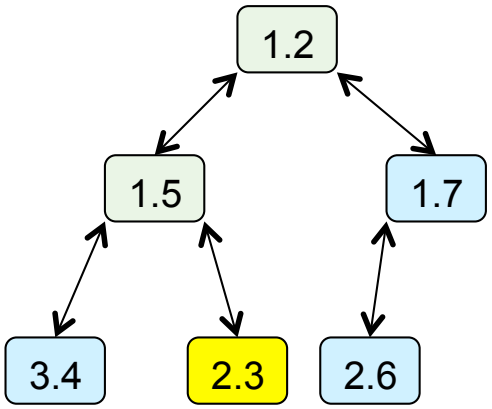
Add element step3



remove element step1



remove element step2



remove element step3

Thank You

- Liang Zhen
- liang@whamcloud.com