October 17th 2013

DataDirect™ NETWORKS
INFORMATION IN MOTION™

LUG2013
China and Japan

Enhancing Lustre Performance and Usability

Shuichi Ihara

Li Xi

DataDirect Networks, Japan

Agenda

DataDirect NETWORKS
Information in Motion

- Today's Lustre trends
- Recent DDN Japan activities for adapting to Lustre trends and new Lustre requirements/challenges

Today's Lustre trends

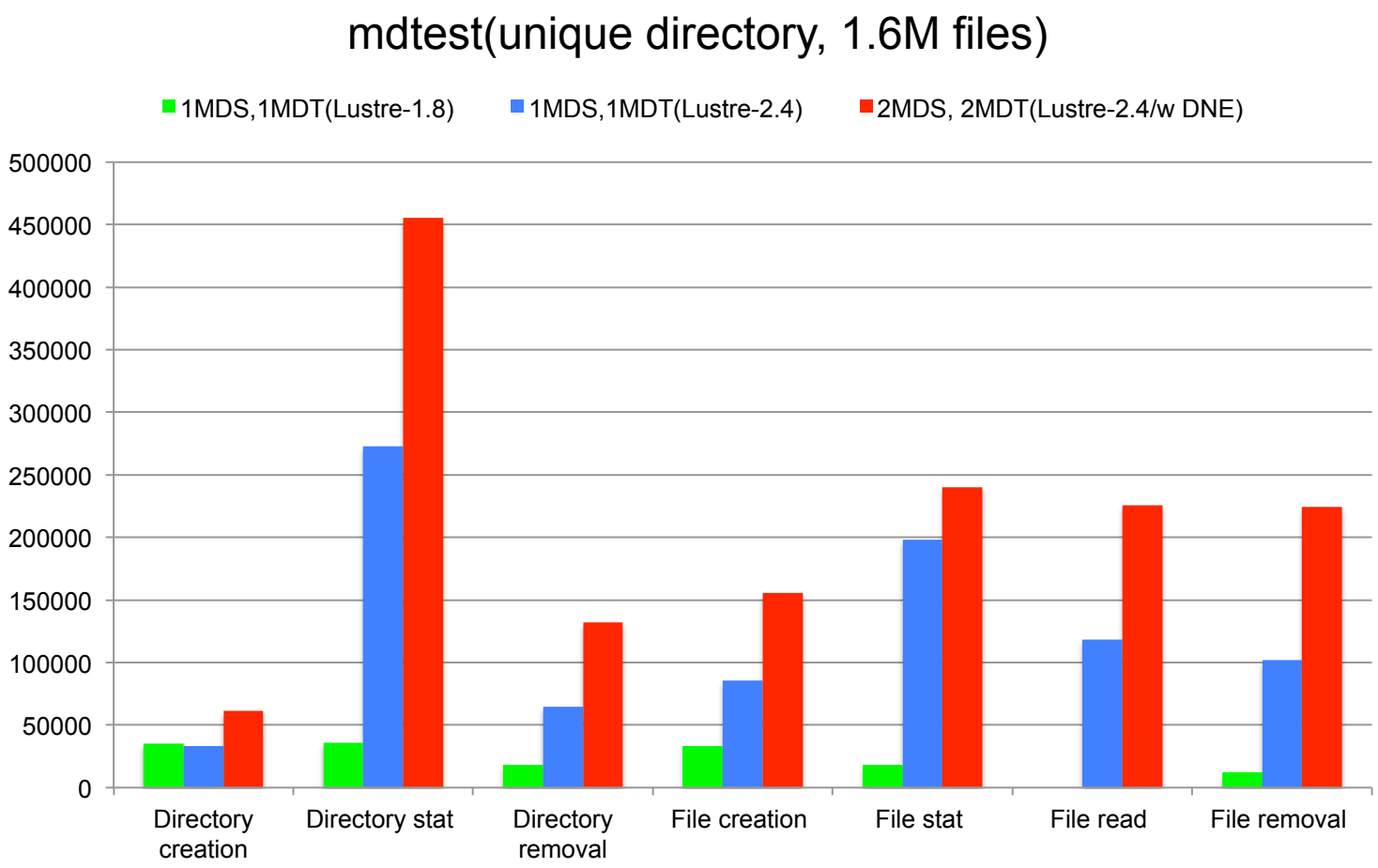
DataDirect NETWORKS
Information in Motion

- Interesting Lustre performance metrics
  - Still throughput performance is scale
  - Huge metadata performance improvement by SMP Scaling and DNE (Lustre-2.3/Lustre-2.4)
  - Layered Flash device helps for random IO performance
- Moving forward to individual feature to framework or even more flexible
  - Lustre checksum (Lustre-2.3)
  - NRS (Network Request Scheduler (Lustre-2.4)
  - Lustre Quota (Lustre-2.3)
  - JobStat (Lustre-2.3)
  - Lustre HSM (Lustre-2.5)
  - .....

Metadata performance comparison

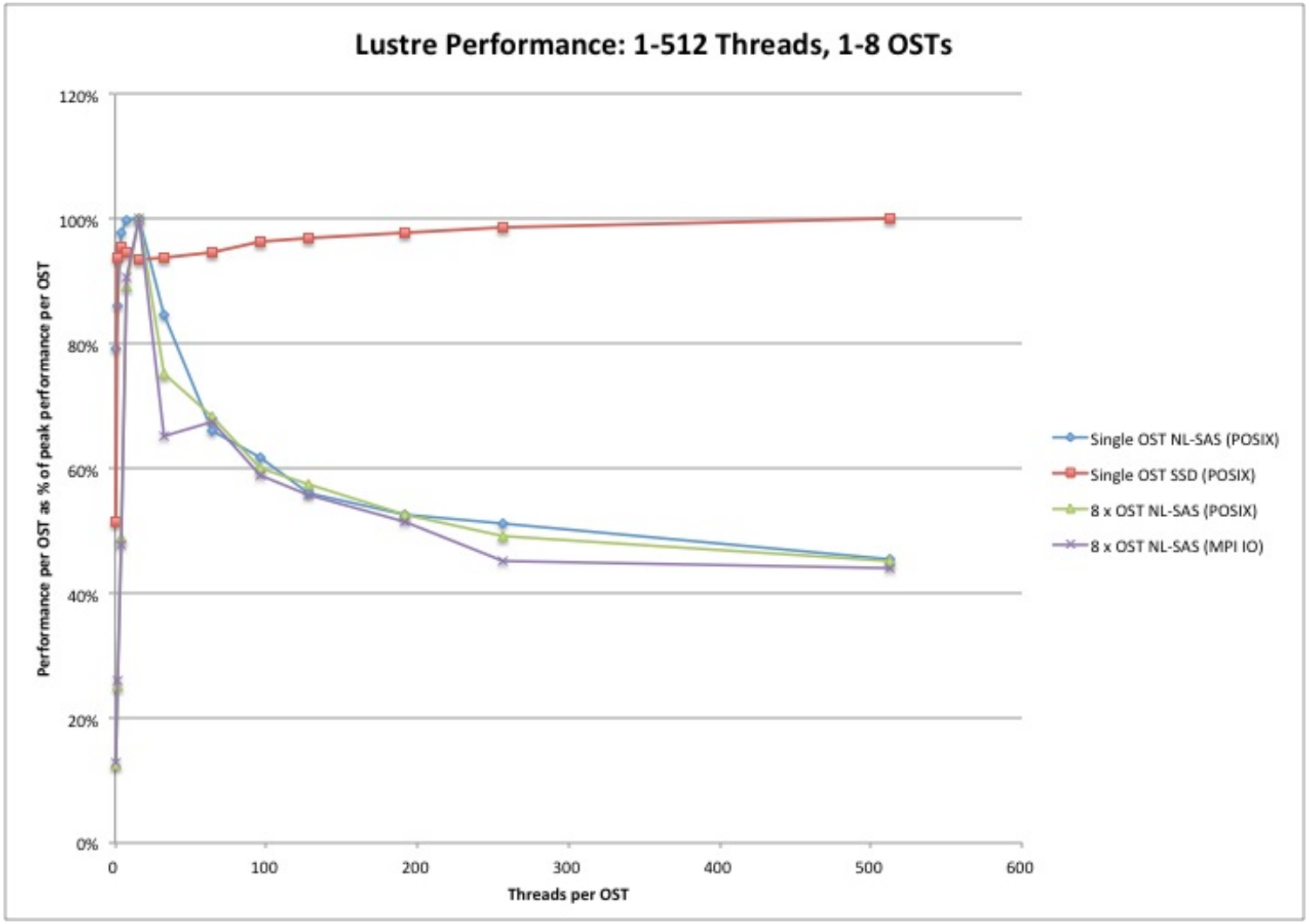
DataDirect NETWORKS
INFORMATION IN MOTION

32 clients, 64 processes, total 1.6M files



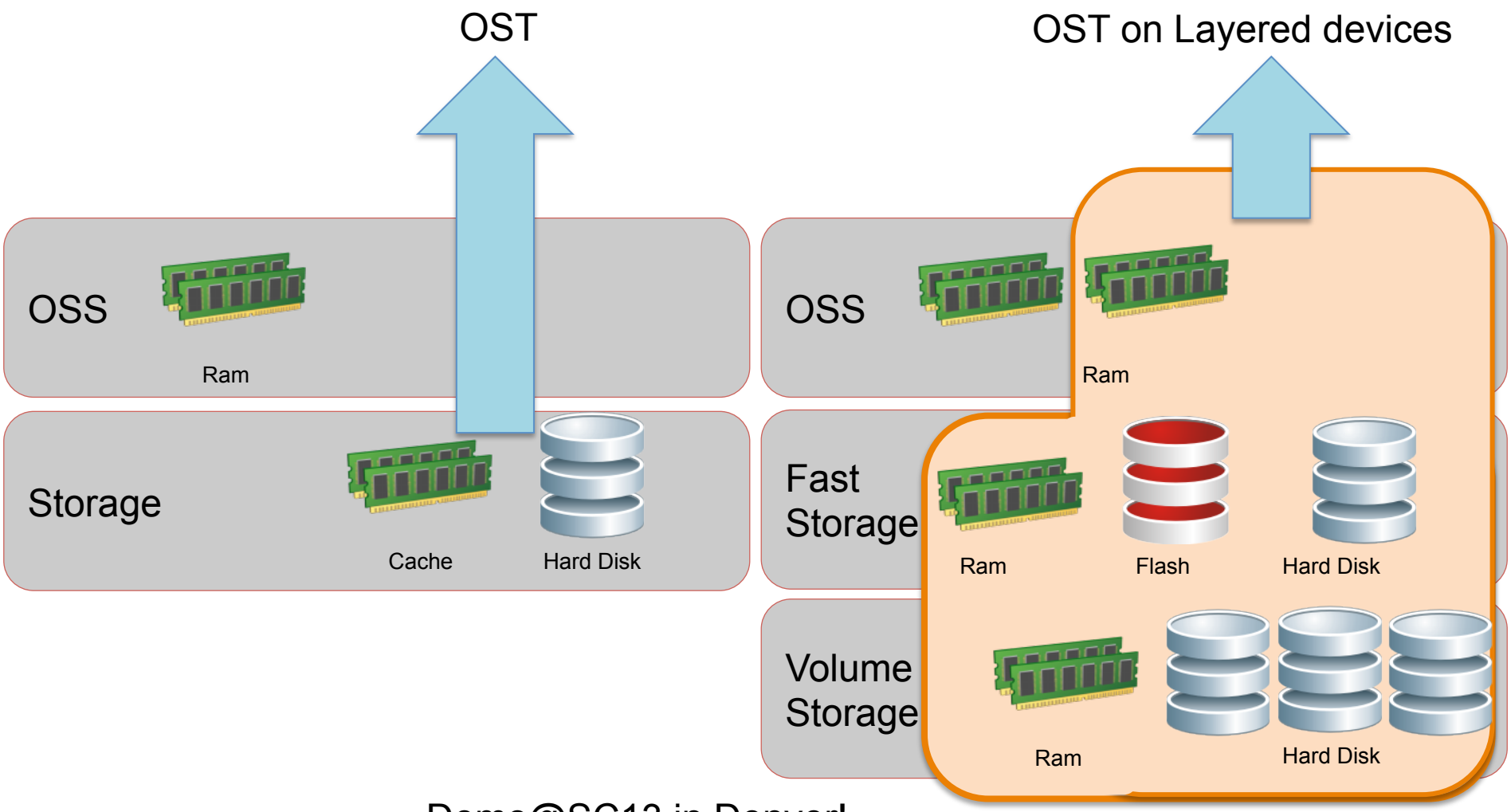
Performance ratio of peak Performance of OST

DataDirect NETWORKS
Information in Motion



OST on Layered devices

DataDirect NETWORKS
Information in Motion



Demo@SC13 in Denver!

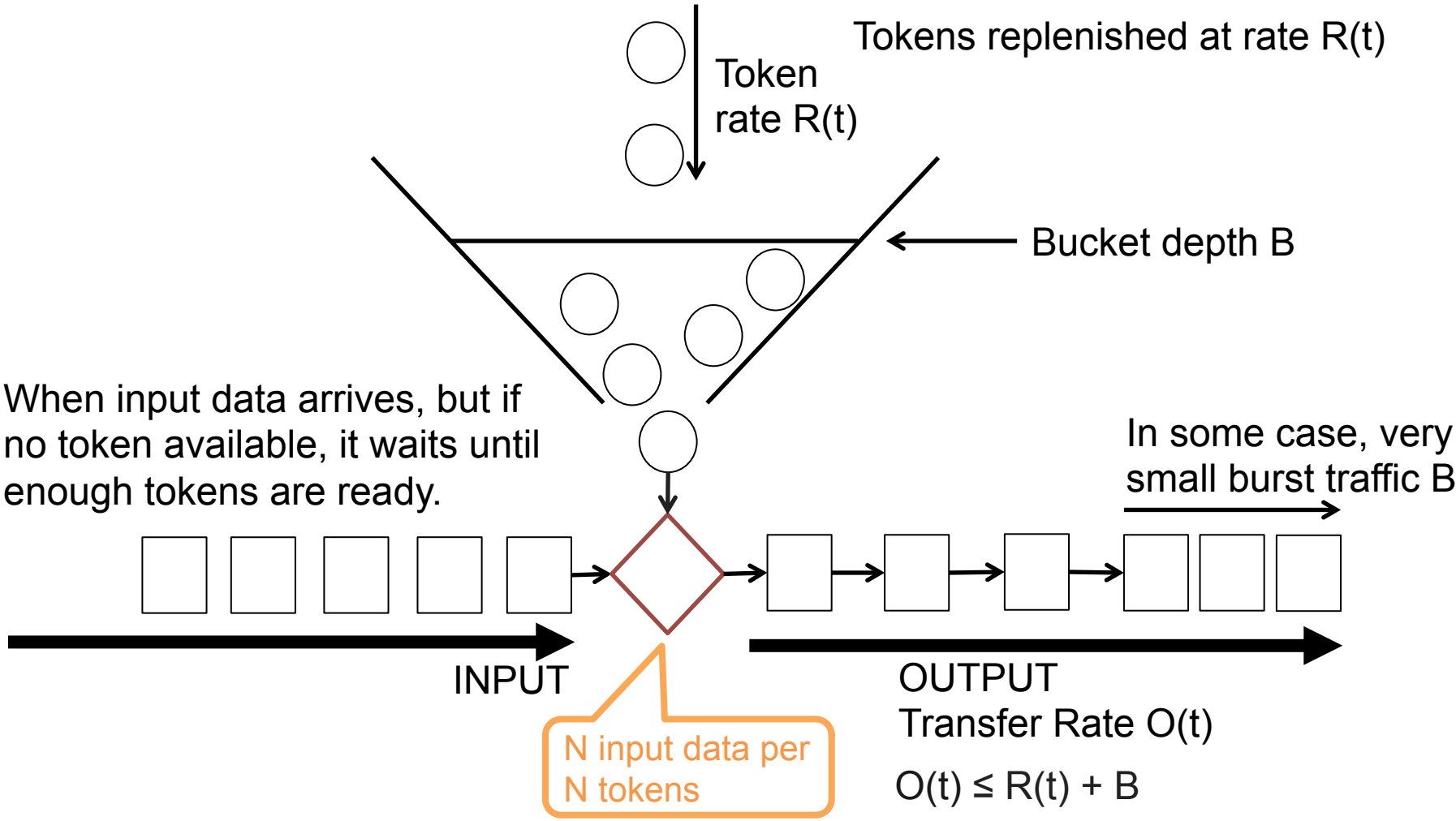
LQS: A QoS Policy for Lustre

DataDirect NETWORKS
Information in Motion

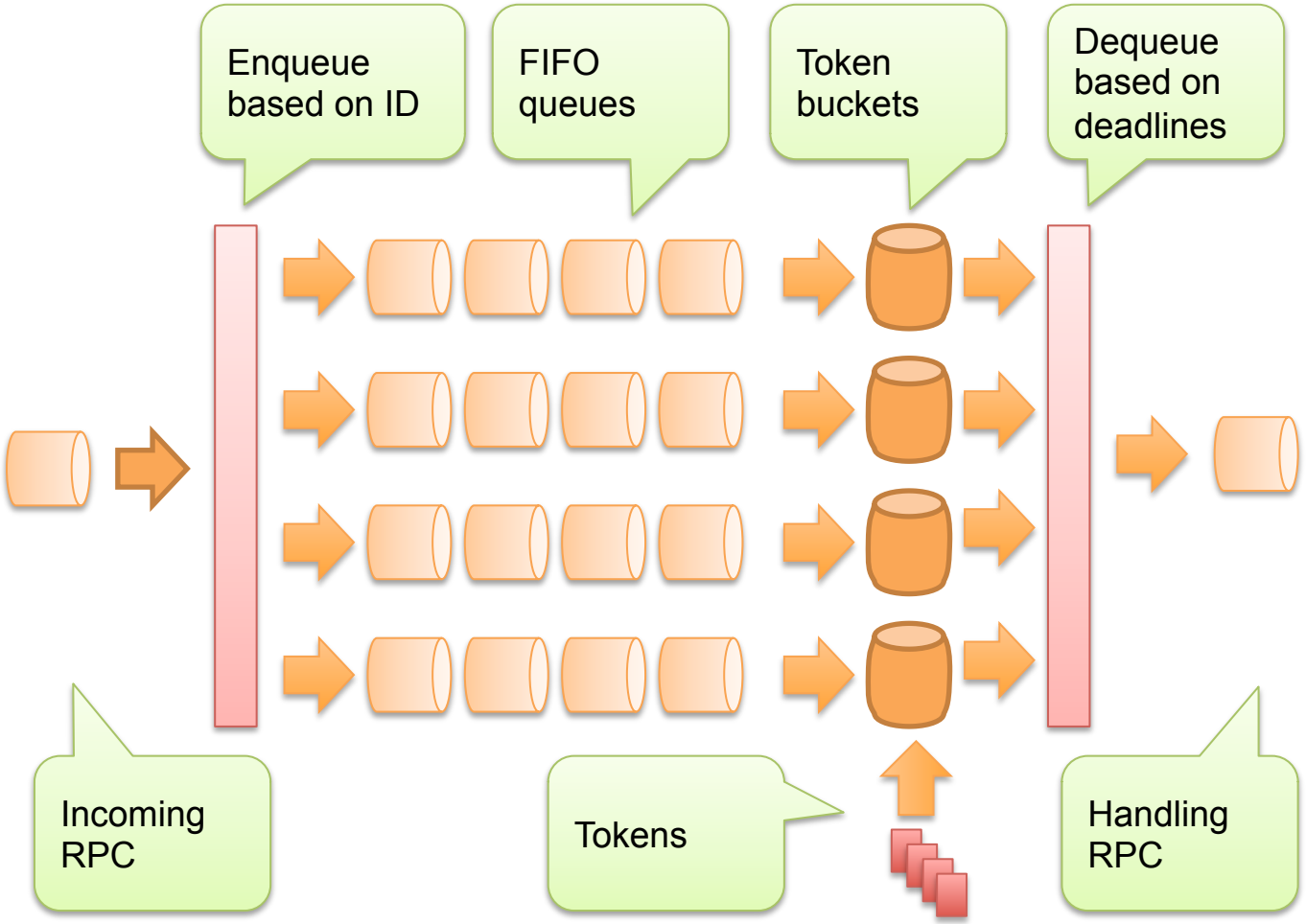
- We have developed a policy layer called “Lustre” QoS (LQS) that can provide QoS by controlling the number of RPCs handled on the Lustre servers
- LQS runs as a policy of the Network Request Scheduler (NRS)
- Eventually, LQS limits Lustre performance by limiting the number of bandwidth and/or metadata operations
- The Token Bucket Filter (TBF) is a major algorithm used in general network systems
  - It's simple and easy to implement
  - Many Ethernet switches and routers use TBF to enable QoS features

The Token Bucket Filter (TBF)

DataDirect NETWORKS
Information in Motion



TBF Implementation for Lustre



TBF patches for Lustre

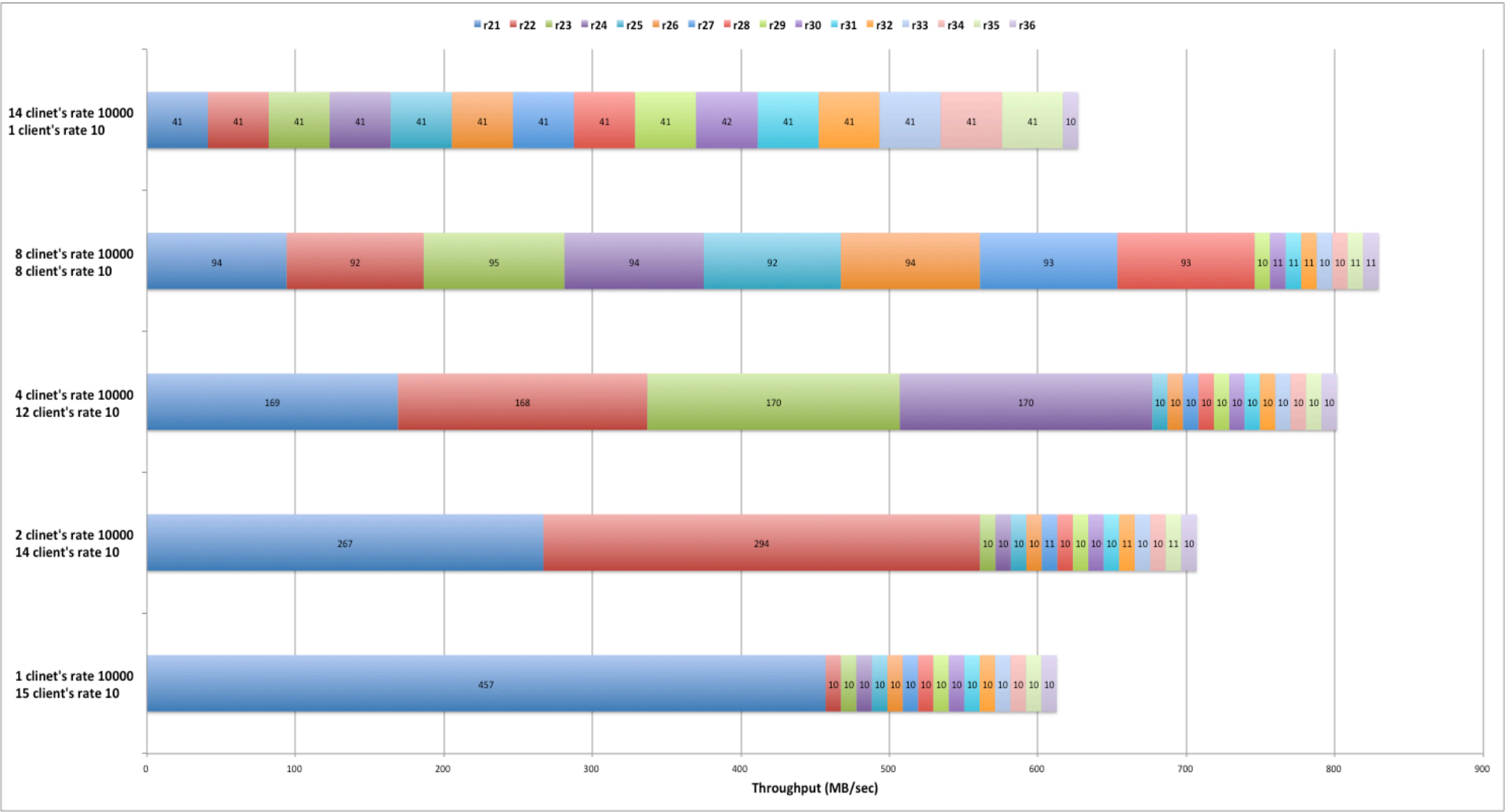
DataDirect NETWORKS
Information in Motion

- LU-3558 ptlrpc: Add the NRS TBF policy
  Main TBF code for NRS-based policy
- LU-3495 ptlrpc: Add rate counter for request handling
  New counters in /proc to show request handing
- LU-3494 libcfs: Add relocation function to libcfs heap
  Added a function to efficiently change the rank of queue

Lustre QoS for Clients
Test results

DataDirect NETWORKS
INFORMATION IN MOTION

"dd" command to single OST from multiple clients with various QoS rules
(Write, 1MB IO, max_rpc_in_flright=32)



OST pool and Quota on OST pools

DataDirect NETWORKS
Information in Motion

- OST number of Lustre clusters is growing rapidly
- OST pool feature enables users to group OSTs together for more flexible and controllable striping
- OST pools follow these rules:
  - An OST can be a member of multiple pools
  - No ordering of OSTs in a pool is defined or implied
  - Stripe allocation within a pool follows the same rules as the normal stripe allocator
  - OST membership in a pool is flexible and can change over time
- OST pool based quota is not supported today
  - But luckily current quota framework is powerful and flexible which makes it easy to add new extension.

OST pool based quota: Requirements

DataDirect NETWORKS
Information in Motion

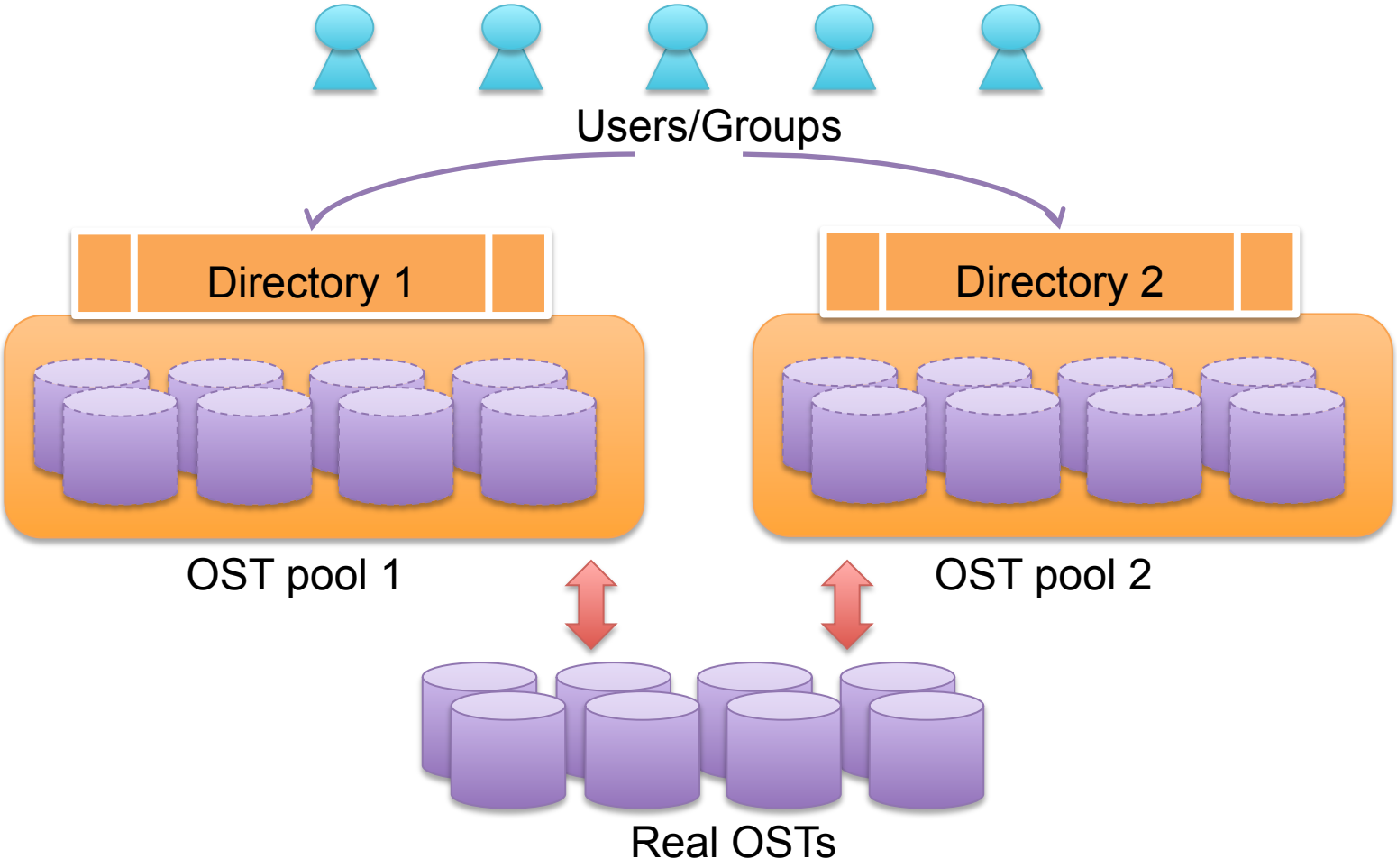
- Integrated in current quota framework
  - Ability to enforce both block and inode quotas
  - Support hard and soft limits
  - Support user/group (and maybe pool) accounting
- Full support of pool
  - Dynamic change of pool definition
  - Separate quotas of users/groups for each pool
- No significant performance impact

Status for quota on OST pool

DataDirect NETWORKS
Information in Motion

- Main framework has been completed
- LU-4017 quota: Add pool support to quota
  - Main codes for pool support of quota
  - The patch is a big one which involves quite a lot of components
  - According to early test, the patch works well
  - Will be split into multiple parts for review
- User space command update
  - Use '-p pool_name' argument to specify which pool to configure
- Test suits for pool based quota
  - Verify the correctness and efficiency of pool based quota
- LDISKFS support is ready, but ZFS support is not yet finished

UseCase :
Quota of users/groups for directories



Conclusion

- Lustre throughput performance is very well and balanced for HPC. Optimized and layered devices help for new performance characteristics and I/O challenges.
- New features in Lustre-2.x are flexible and easy to add new functional capability
- We adapted a standard Token Bucket Filter (TBF) algorithm to Lustre and implemented LQS, a QoS policy based on the Lustre Network Request Scheduler (NRS) framework
- We are working on Quota enhancement in conjunction with OST pool to support OST based quota.

DataDirect™
NETWORKS
INFORMATION IN MOTION™

LUG2013
China and Japan