HiBD
High-Performance
Big Data

Accelerating OpenStack Swift with RDMA for Building Efficient HPC Clouds

A Talk at OpenStack Summit 2017

by

Xiaoyi Lu
The Ohio State University
E-mail: luxi@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~luxi

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

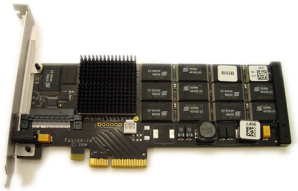
Outline

- Introduction
- Problem Statement
- Proposed Design
- Performance Evaluation
- Conclusion and Future Work

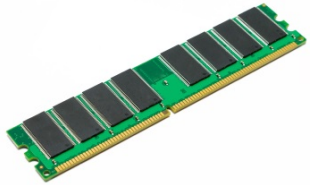
Drivers of Modern HPC Cloud Architectures

Multi-core Processors

High Performance Interconnects – InfiniBand (with SR-IOV)
<1usec latency, 200Gbps Bandwidth>



SSDs, Object Storage Clusters



Large memory nodes
(Upto 2 TB)

- Multi-core/many-core technologies
- Large memory nodes
- High-performance networking (InfiniBand and RoCE)
- Single Root I/O Virtualization (SR-IOV)
- Solid State Drives (SSDs), Object Storage Clusters

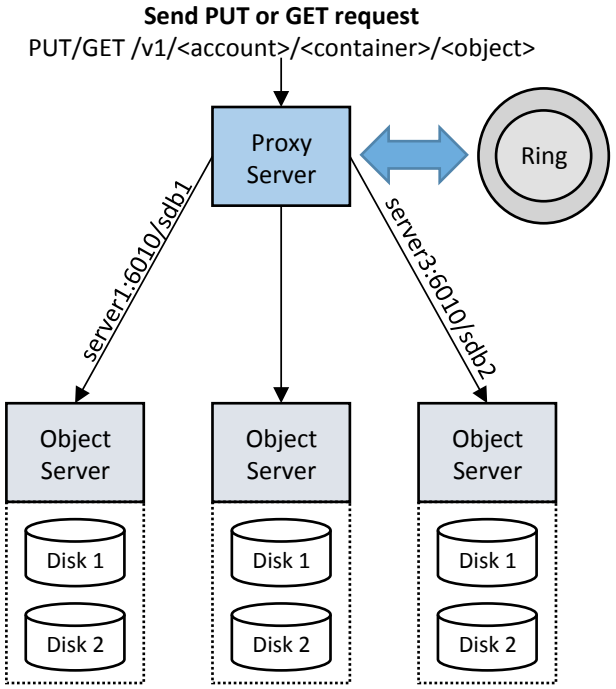
Microsoft Azure | amazon web services EC2 | Grid'5000 | ORACLE Cloud | Chameleon Cloud

Summary of HPC Cloud Resources

- High-Performance Cloud systems have adopted advanced interconnects and protocols
  - InfiniBand, 40 Gigabit Ethernet/iWARP, RDMA over Converged Enhanced Ethernet (RoCE)
  - Low latency (few micro seconds), High Bandwidth (200 Gb/s with HDR InfiniBand)
  - Remote Direct Memory Access (RDMA)
- Vast installations of Object Storage systems (e.g. Swift, Ceph)
  - Total capacity is in the PB range
  - Offer high availability and fault-tolerance
  - Performance and scalability is still a problem

OpenStack Swift Overview

- Distributed Cloud-based Object Storage Service
- Deployed as part of OpenStack installation
- Can be deployed as standalone storage solution as well
- Worldwide data access via Internet
  - HTTP-based
- Architecture
  - Multiple Object Servers: To store data
  - Few Proxy Servers: Act as a proxy for all requests
  - Ring: Handles metadata
- Usage
  - Input/output source for Big Data applications
  - Software/Data backup
  - Storage of VM/Docker images
- Based on traditional TCP sockets communication



Swift Architecture

Outline

- Introduction
- Problem Statement
- Proposed Design
- Performance Evaluation
- Conclusion and Future Work

Problem Statement

- **Challenges**
  - Proxy server is bottleneck for large scale deployments
  - Object upload/download operations are network intensive
  - Can an RDMA-based approach benefit?

How can high-performance and scalable RDMA-based communication schemes be designed to improve overall Swift performance?

Outline

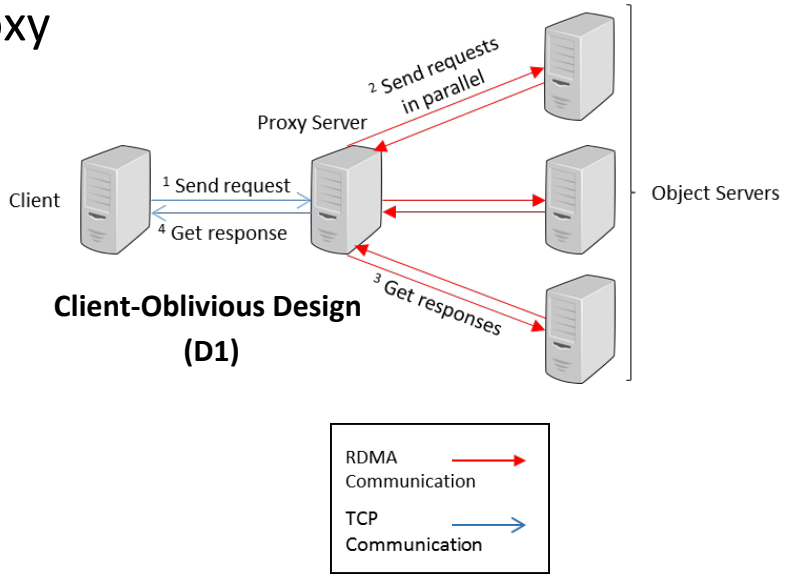
- Introduction
- Problem Statement
- Proposed Design
- Performance Evaluation
- Conclusion and Future Work

Proposed Design

- Re-designed Swift architecture for improved scalability and performance; Two proposed designs:
  - **Client-Oblivious Design**: No changes required on the client side
  - **Metadata Server-based Design**: Direct communication between client and object servers; bypass proxy server
- RDMA-based communication framework for accelerating networking performance
- High-performance I/O framework to provide maximum overlap between communication and I/O

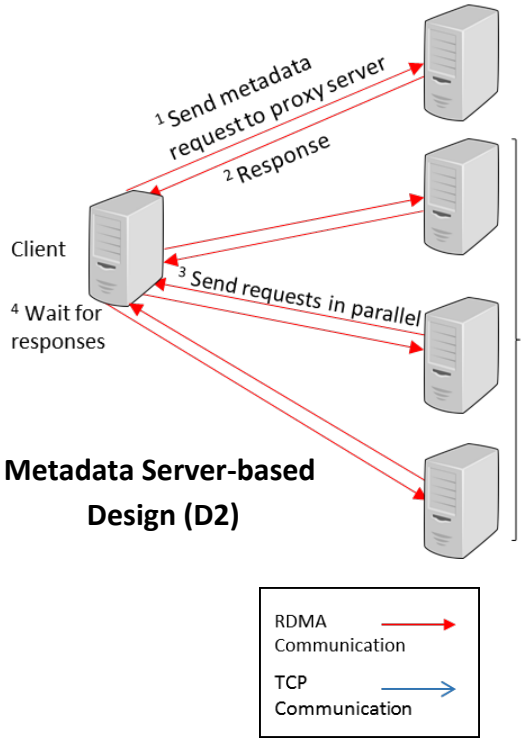
Client-Oblivious Design

- No change required on the client side
- Communication between client and proxy server using conventional TCP sockets networking
- Communication between proxy server using high-performance RDMA-based networking
- Proxy Server is still the bottleneck!



Metadata Server-based Design

- Re-designed architecture for improved scalability
- Client-based replication for reduced latency and high-performance
- All communication using high-performance RDMA-based networking
- Proxy Server no longer the bottleneck!

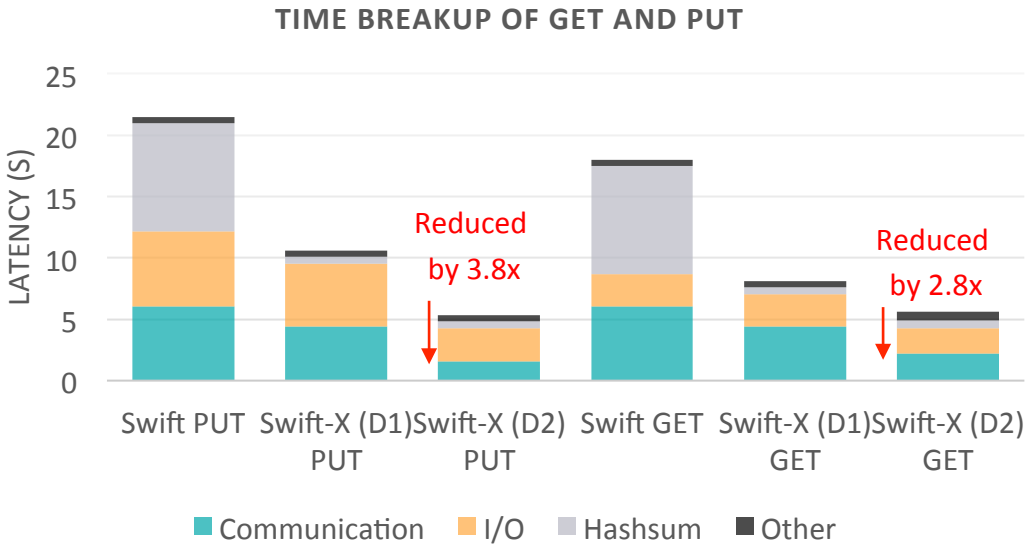


S. Gugnani, X. Lu, and D. K. Panda. Swift-X: Accelerating OpenStack Swift with RDMA for Building an Efficient HPC Cloud. The 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2017.

Outline

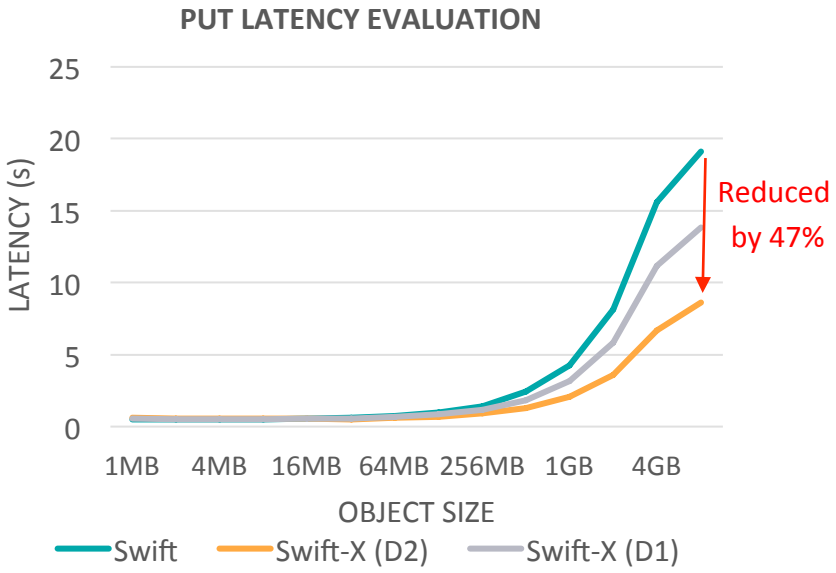
- Introduction
- Problem Statement
- Proposed Design
- Performance Evaluation
- Conclusion and Future Work

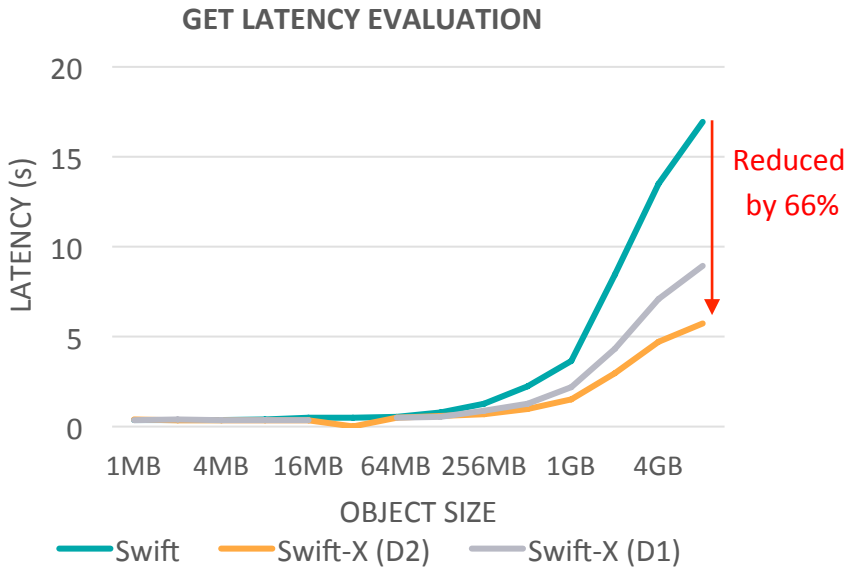
Breakdown of Put and Get Operations



- Communication time reduced by up to **3.8x** for PUT and up to **2.8x** for GET

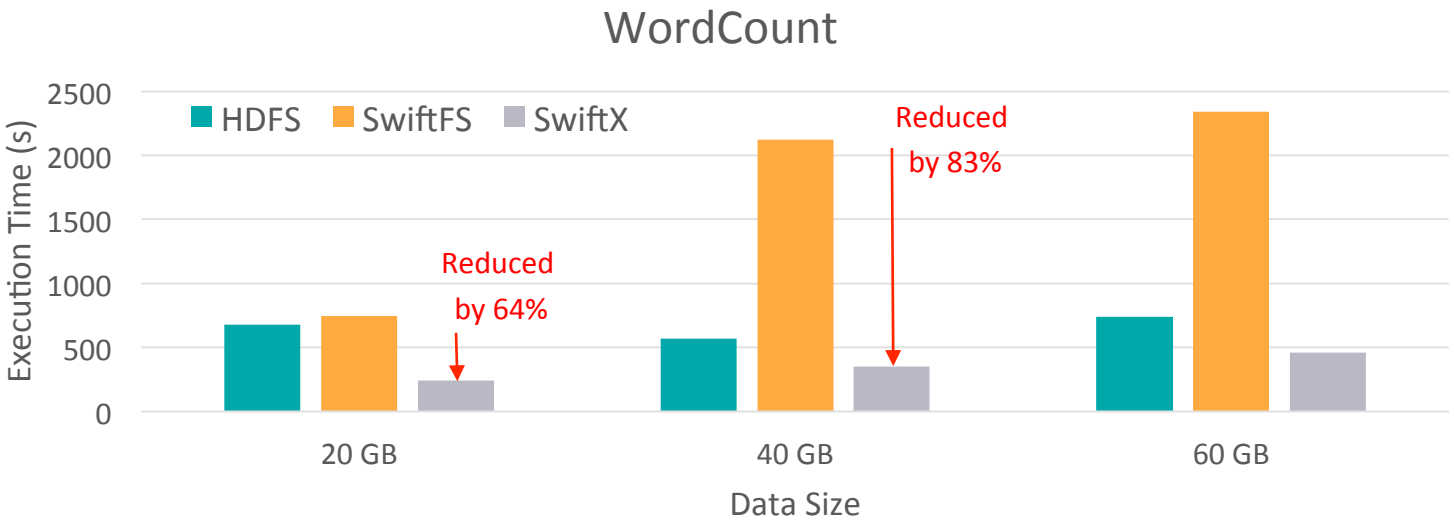
Experimental Evaluation for Put and Get Operations





- Up to **47%** reduction in PUT latency
- Up to **66%** reduction in GET latency

Evaluation with Hadoop WordCount



- Up to **83%** improvement over SwiftFS
- Up to **64%** improvement over HDFS
- With HDFS, data needs to be copied to/from Swift

Outline

- Introduction
- Problem Statement
- Proposed Design
- Performance Evaluation
- Conclusion and Future Work

Conclusion

- Analyzed Swift architecture and identified major bottlenecks
- Proposed two designs to accelerate Swift performance and improve scalability
  - Client-Oblivious Design
  - Metadata Server-based Design
- Designed high-performance RDMA-based communication framework
- Experimental Evaluation shows promising results
- Future Work
  - Evaluation with additional application scenarios
  - Support of S3 and POSIX APIs with our designs
  - RDMA-design will be publicly available

One More Presentations

- Thursday, 11 May, 3:10pm, Hynes Convention Center - Level Three - MR 312

  Building Efficient HPC Clouds with MVAPICH2 and OpenStack over SR-IOV enabled InfiniBand Clusters

Funding Acknowledgments

Funding Support by

Equipment Support by

Personnel Acknowledgments

Current Students

- A. Awan (Ph.D.)
- R. Biswas (M.S.)
- M. Bayatpour (Ph.D.)
- S. Chakraborthy (Ph.D.)
- C.-H. Chu (Ph.D.)
- S. Guganani (Ph.D.)
- J. Hashmi (Ph.D.)
- H. Javed (Ph.D.)
- M. Li (Ph.D.)
- D. Shankar (Ph.D.)
- H. Shi (Ph.D.)
- J. Zhang (Ph.D.)

Current Research Scientists

- X. Lu
- H. Subramoni

Current Research Specialist

- J. Smith

Past Students

- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- W. Huang (Ph.D.)
- N. Islam (Ph.D.)
- W. Jiang (M.S.)
- J. Jose (Ph.D.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- K. Kulkarni (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- K. Kandalla (Ph.D.)
- P. Lai (M.S.)
- J. Liu (Ph.D.)
- M. Luo (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- V. Meshram (M.S.)
- A. Moody (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- X. Ouyang (Ph.D.)
- S. Pai (M.S.)
- S. Potluri (Ph.D.)
- M.-W. Rahman (Ph.D.)
- R. Rajachandrasekar (Ph.D.)
- G. Santhanaraman (Ph.D.)
- A. Singh (Ph.D.)
- J. Sridhar (M.S.)
- S. Sur (Ph.D.)
- H. Subramoni (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Venkatesh (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)

Past Research Scientist

- K. Hamidouche
- S. Sur

Past Programmers

- D. Bureddy
- M. Arnold
- J. Perkins

Past Post-Docs

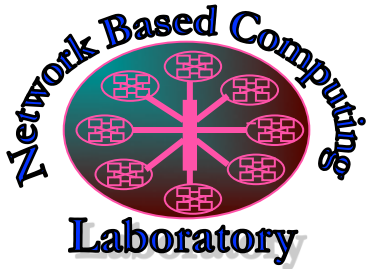
- D. Banerjee
- X. Besseron
- H.-W. Jin
- J. Lin
- M. Luo
- E. Mancini
- S. Marcarelli
- J. Vienne
- H. Wang

Thank You!

{luxi, panda}@cse.ohio-state.edu

http://www.cse.ohio-state.edu/~luxi

http://www.cse.ohio-state.edu/~panda



Network-Based Computing Laboratory

http://nowlab.cse.ohio-state.edu/

The High-Performance Big Data Project

http://hibd.cse.ohio-state.edu/