Ultra-high Performance Parallel Big Data Transfer Software for Data-intensive Science

Chin Fang, Ph.D., Founder
fangchin@zettar.com
650-644-9722
Agenda

Problem solved & motivation of the design

Unique capabilities & a key to high-speed data transfers

Use cases

Architecture overview

Engineering trade-offs to meet SC14 deadline

Design and implementation challenges

A quick review of the SC14 setup

SLAC pilot deployment discussions

Upcoming SC15 excitements 😊

Q&A
Too much data to store? Scale-out storage
Too much data to analyze? Scale-out compute
Too much data to transfer? Why not scale-out transfer?
Bounded Throughput, SPOF, Low In-Transit Security
Scalable Throughput, HA, High In-Transit Security
Agenda

Problem solved & motivation of the design

Unique capabilities & a key to high-speed data transfers

Use cases

Architecture overview

Engineering trade-offs to meet SC14 deadline

Design and implementation challenges

A quick review of the SC14 setup

SLAC pilot deployment discussions

Upcoming SC15 excitements 😊

Q&A
ZX – Boundlessly Scalable Data Transfer Throughput

Data Transfer Throughput

- 10Gbps
- 20Gbps
- 30Gbps

Storage service link
Data transfer links

Copyright © Zettar Inc. 2013 - 2015
Zettar confidential: released to SLAC
SSD, eSATA, USB 3, Thunderbolt, Network Speeds (MiB/s)

- **Intel P3700 (2670 MiB/s)**: NVMe SSD
- **Intel 730 (524 MiB/s)**: SATA SSD
- **eSATA (375 MiB/s)**: Interface
- **USB 3.0 (625 MiB/s)**: Interface
- **Thunderbolt (1,250 MiB/s)**: Interface
- **1Gbps (125 MiB/s)**: Network
- **10Gbps (1,250 MiB/s)**: Network
- **100Gbps (12,500 MiB/s)**: Network

Copyright © Zettar Inc. 2013 - 2015

Zettar confidential: released to SLAC
All 7 pairs are 10GbE connected, why the throughput differences?

Data source: Dr. Reese, Stanford Research Computing and Dr. Chin Fang, Zettar Inc. 2013
Fast Data Transfers Demand More Than Networks - II
Verizon EVPL

Ethernet Access
Agenda

Problem solved & motivation of the design
Unique capabilities & a key to high-speed data transfers

Use cases

Architecture overview

Engineering trade-offs to meet SC14 deadline

Design and implementation challenges

A quick review of the SC14 setup

SLAC pilot deployment discussions

Upcoming SC15 excitements 😊

Q&A
Problem solved & motivation of the design
Unique capabilities & a key to high-speed data transfers
Use cases

**Architecture overview**

Engineering trade-offs to meet SC14 deadline
Design and implementation challenges
A quick review of the SC14 setup
SLAC pilot deployment discussions
Upcoming SC15 excitements 😊
Q&A
ZX – User-mode P2P-based Multipath Data Transfers

Up to \((p \times q \times N)\) paths/xfer
Agenda

Problem solved & motivation of the design
Unique capabilities & a key to high-speed data transfers
Use cases
Architecture overview
Engineering trade-offs to meet SC14 deadline
Design and implementation challenges
A quick review of the SC14 setup
SLAC pilot deployment discussions
Upcoming SC15 excitements 😊
Q&A
Engineering Trade-offs To Meet SC14 Deadline

1. KISS to trade for short-term scalability & performance
2. Routing complexity to trade for scalability & performance
3. Demanding programming model to trade for product longevity

Never compromised: ease-of-use + good OOTB performance
Agenda

Problem solved & motivation of the design
Unique capabilities & a key to high-speed data transfers
Use cases
Architecture overview
Engineering trade-offs to meet SC14 deadline

**Design and implementation challenges**

A quick review of the SC14 setup
SLAC pilot deployment discussions
Upcoming SC15 excitements 😊
Q&A
Design And Implementation Challenges

1. Distributed, highly concurrent, \textit{asynchronous} programming model

2. Resources for testing and verification. Debugging too!

3. Human factors, e.g. communication
Agenda

Problem solved & motivation of the design
Unique capabilities & a key to high-speed data transfers
Use cases
Architecture overview
Engineering trade-offs to meet SC14 deadline
Design and implementation challenges
A quick review of the SC14 setup
SLAC pilot deployment discussions
Upcoming SC15 excitements 😊
Q&A
Supercomputing 2014 Live Demo Setup
The Same Setup, Now Hosted In SLAC Building 50
Agenda

Problem solved & motivation of the design
Unique capabilities & a key to high-speed data transfers
Use cases
Architecture overview
Engineering trade-offs to meet SC14 deadline
Design and implementation challenges
A quick review of the SC14 setup
SLAC pilot deployment discussions
Upcoming SC15 excitements 😊
Q&A
Pilot Deployment At SLAC

1. Campus data transfer to start
2. LOSF cases?
3. Mix-sized and large file cases?

Lets use existing hardware as much as possible

Recommended

- SSDs
- 10GbE NICs

1U server
Switch
Network Cloud
Agenda

Problem solved & motivation of the design
Unique capabilities & a key to high-speed data transfers
Use cases
Architecture overview
Engineering trade-offs to meet SC14 deadline
Design and implementation challenges
A quick review of the SC14 setup
SLAC pilot deployment discussions

Upcoming SC15 excitements 😊

Q&A
Upcoming SLAC Publications


2. A study of the Linux XFS and parallel file system overheads on the data transfer performance using high-density servers equipped with NVMe PCIe SSD devices.

3. More to come…
SC15: Surpass The Best LOSF Throughput Over 100Gbps

Source: Optimizing Large Data Transfers over 100Gbps Wide Area Networks, 2013
SC15: Transferring Big Data ~100 Gbps With A 2U Server

40GE Server Design Kit

- SandyBridge E5 Based Servers:
  - (SuperMicro X9DRI-F or Dell R720)
  - Intel E5-2670 with C1 or C2 Stepping
  - 128GB of DDR3 1600MHz RAM
- Mellanox VPI CX-3 PCIe Gen3 NIC
- Dell / Mellanox QSFP Active Fiber Cables
- LSI 9265-8i, 8 port SATA 6G RAID Controller
- OCZ Vertex 3 SSD, 6Gb/s (preferably enterprise disks like Deneva 2)
- Dell – Force10; Z9000 40GE Switch

Server Cost = ~ $15k

QuantaPlex T41SP-2U (4-Node)
- Intel Haswell CPUs
- 10Gbps NICs
- NVMe SSDs

100Gbps 100Gbps

Copyright © Zettar Inc. 2013 - 2015

Zettar confidential: released to SLAC

Picture credit: Caltech
Ultra-high Performance Parallel Big Data Transfer
Software for Data-intensive Science

Chin Fang, Ph.D., Founder
fangchin@zettar.com
650-644-9722
Problem solved & motivation of the design
Unique capabilities & a key to high-speed data transfers
Use cases
Architecture overview
Engineering trade-offs to meet SC14 deadline
Design and implementation challenges
A quick review of the SC14 setup
SLAC pilot deployment discussions
Upcoming SC15 excitements 😊

Q&A
High-Speed TCP Data Transfers - Background

TCP Congestion Performance (RTT 70ms, 1500 MSS, 10Gbps, 256Mb queue)

- Slow Start (1.2s)
- Congestion Avoidance Cycle (34 minutes)

Congestion Control
- Exponential Growth
- Multiplicative Decrease

Packet Loss

Congestion Avoidance

Sum of Parallel Streams
- Single Stream
- Individual Streams

Sum of Distributed Parallel Streams
- Sum of Synchronized Parallel Streams
- Single Stream

Time (Hours)

Picture credit: Geoff Huston, APNIC & Ilya Grigorik, Google Inc.
Computer Memory Hierarchy Diagram

- Processor
- CPU
  - Processor Register
  - CPU Cache
    - Level 1 (L1) Cache
    - Level 2 (L2) Cache
    - Level 3 (L3) Cache
- Physical Memory
  - Random Access Memory (RAM)
- Solid State Memory
  - Non-Volatile Flash-based Memory
- Virtual Memory
  - File-based Memory
- EDO, SD-RAM, DDR-SDRAM, RD-RAM and More...
- SSD, Flash Drive
- Mechanical Hard Drives

Simplified Computer Memory Hierarchy Illustration: Ryan J. Leng
Apply Storage Auto-Tiering To High-speed Data Transfers

AUTO TIERING – HOW IT WORKS

Flash/SSD

Fast Disks

Capacity Disks

Cloud Storage

Tier 1

Tier 2

Tier 3

Tier N

ABOVE average moves UP a tier

BELOW average moves DOWN a tier

Copyright © Zettar Inc. 2013 - 2015

Zettar confidential: released to SLAC