

US Patent

Pending



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

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

Copyright © Zettar Inc. 2013 - 2015

Zettar confidential: released to SLAC

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 ©

Scale-up or Scale-out?



- Too much data to store? Scale-out storage Too much data to analyze? Scale-out compute 2 3
 - Too much data to transfer? Why not scale-out transfer?

Bounded Throughput, SPOF, Low In-Transit Security



Copyright © Zettar Inc. 2013 - 2015

Scalable Throughput, HA, High In-Transit Security



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 ©

ZX – Boundlessly Scalable Data Transfer Throughput





SSD, eSATA, USB 3, Thunderbolt, Network Speeds (MiB/s)



Network

Fast Data Transfers Demand More Than Networks - I

All 7 pairs are 10GbE connected, why the throughput differences?



zcloud (< 1Yr in dev) vs bbcp (> 10Yrs in dev)

Data source: Dr. Reese, Stanford Research Computing and Dr, Chin Fang, Zettar Inc. 2013

Fast Data Transfers Demand More Than Networks - II



Evaluations/Demonstrations of 100 Gbps Disk-to-Disk WAN File Transfer Performance

Copyright © Zettar Inc. 2013 - 2015

Zettar confidential: released to SLAC

Verizon EVPL

Ethernet Access



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 ©

Data Transfer Patterns



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 ©

ZX – User-mode P2P-based Multipath Data Transfers



US Patent Pending



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 ©

Engineering Trade-offs To Meet SC14 Deadline



- 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

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 ©

Design And Implementation Challenges

Distributed, highly concurrent, *asynchronous* programming model

2 Resources for testing and verification. Debugging too!

3 Human factors, e.g. communication

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 ©

Supercomputing 2014 Live Demo Setup



The Same Setup, Now Hosted In SLAC Building 50



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 ©

Pilot Deployment At SLAC



Campus data transfer to start



LOSF cases?



Mix-sized and large file cases?

Lets use existing hardware as much as possible





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 ©

Upcoming SLAC Publications





- **SLAC-TN-15-001**: Using NVMe Gen3 PCIe SSD cards in high-1 density servers for high-performance big data transfers over multiple network channels. Published on Feb. 24, 2015
- A study of the Linux XFS and parallel file system overheads on 2 the data transfer performance using high-density servers equipped with NVMe PCIe SSD devices.



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

Picture credit: Caltech



US LHCNe



QuantaPlex T41SP-2U (4-Node)

- Intel Haswell CPUs
- 10Gbps NICs
- NVMe SSDs





100Gbps

100Gbps



US Patent

Pending



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

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

Copyright © Zettar Inc. 2013 - 2015

Zettar confidential: released to SLAC

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 ©



High-Speed TCP Data Transfers - Background

12 **Slow Start** Packet Loss (1.2s) 130-Congestion window size (segments) 10 **Congestion Control** - Multiplicative Decrease 100. 8 80. Rate (Gbps) 60-Exponential Growth 40 -4 Congestion Avoidance Cycle (34 minutes) **Congestion Avoidance** 20-2 5 6 7 8 9 10 11 **Round Trips** 0 0.0 0.5 1.0 1.5 2.0 Time (Hours) 12 12 Sum of Parallel Streams Sum of Distributed Parallel Streams 10 10 (WWWWWWWWW 8 8 Rate (Gbps) Rate (Gbps) Single Stream Sum of Synchronized Parallel Streams **Single Stream** 4 ndividual Streams 2 0 0 1.5 0.0 0.5 1.0 2.0 0.0 0.2 0.4 0.6 0.8 1.0 Time (Hours) Time (Hours)

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

Picture credit: Geoff Huston, APNIC & Ilya Grigorik, Google Inc.

Computer Memory Hierarchy Diagram



Illustration: Ryan J. Leng

Apply Storage Auto-Tiering To High-speed Data Transfers

AUTO TIERING - HOW IT WORKS



Picture credit: DataCore Software