Theme 1 – Terabit Networking

A report at Australia/China SKA Big Data Workshop

Xiaoying Zheng (Shanghai Advanced Research Institute, Chinese Academy of Sciences, China) Paul Brooks (Trident Subsea Cable, Australia) Shaoguang Guo (Shanghai Astronomical Observatory, Chinese Academy of Sciences, China) Liang Li (Computer Network Information Center, Chinese Academy of Sciences, China) Yueqi Han (Computer Network Information Center, Chinese Academy of Sciences, China) Yuan Luo (Shanghai Jiaotong University, China)

2017/4/13

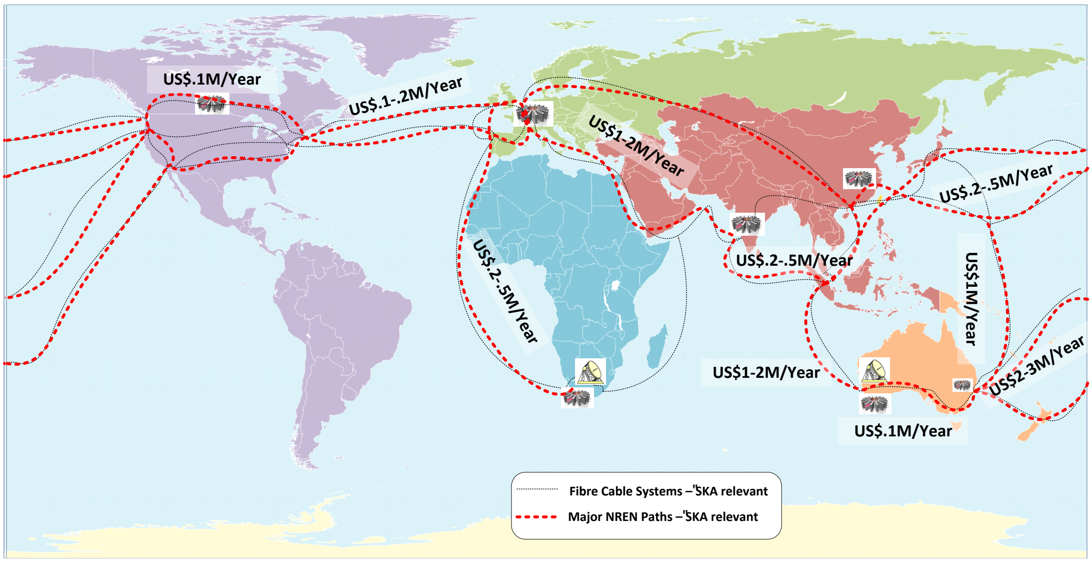
Index

Abstract

Theme-1 of Networking introduces the main challenges faced by the data distribution between SKA centers, including the infrastructure options between Perth and China, how to distribute data from Perth across the globe to regional centers, and how to distribute data from a regional center to end users, efficiently. Regarding the infrastructure options between Perth and China, Theme-1 discusses three infrastructure options, dedicated fiber, public hybrid, or hybrid. The conclusion drawn is that the current submarine infrastructure bandwidth is not a bottleneck for a 100Pbs scale transmission. The priority is to set up tests, evaluate the performance of three options, and suggest the candidate with the best cost performance Index. Regarding data distribution between Perth and regional centers, some candidate techniques are discussed. The objective is that each SKA data copy is transmitted only once to save the expensive Australian outband bandwidth. The regional centers can exchange their data copies to help the data distribution. Regarding data distribution from a regional center to end users, Theme-1 proposes several candidate solutions. The plan is to evaluate the candidates by literature survey, simulation and set up testbeds.

1. Background of bandwidth budget and subsea budget

1.1 Budget Bandwidth of SKA



v

Fig. 1 Budget Bandwidth of SKA [1]

Rough estimate prices $/yr for 100Gbps (0.1 Tbps)

100Gbps moves ~1 PB/day. 300PB/year = 82% utilisation 24x7x365! – 100Gbps not enough??

1.2 Current international subsea fiber across the globe

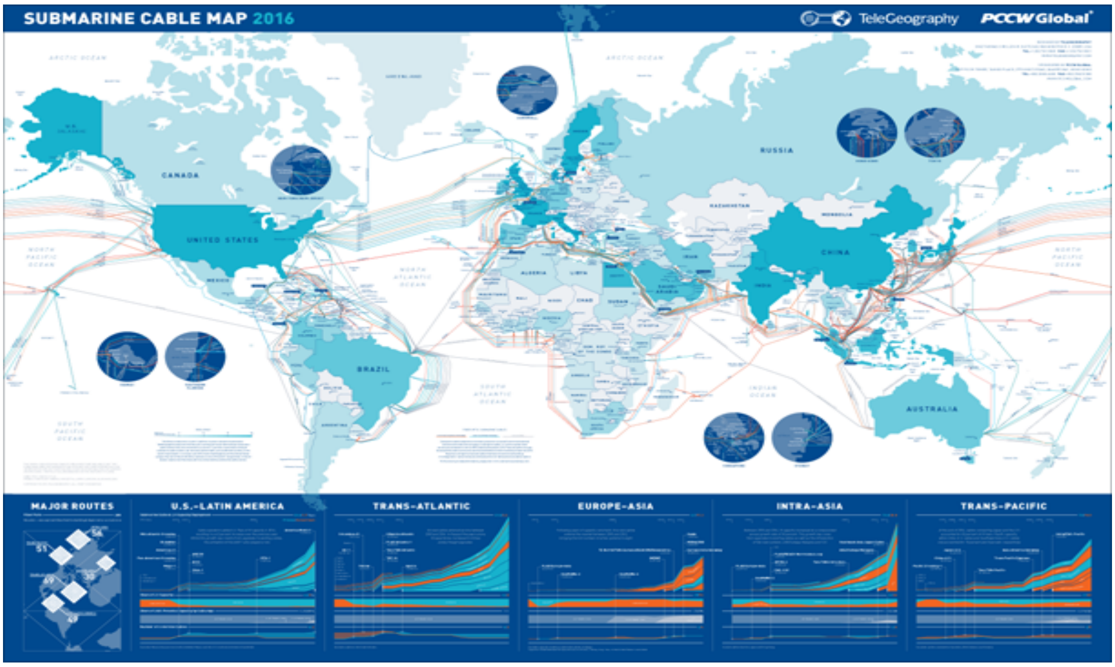


Fig. 2 Telegeography Cable Map across the globe [2][3]

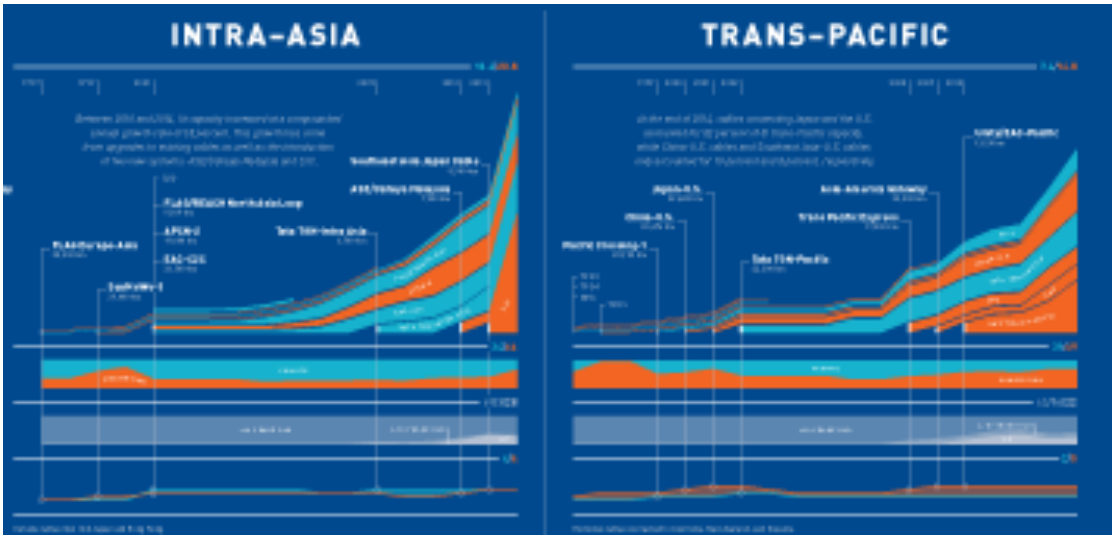


Fig. 3 The evolution of Telegeography Cables in Asia-Pacific area [2][3]

~320 cables – the real World Wide Web!
Typical recent cable: 4 fibre-pairs, 100 x 100Gbps = 10Tbps/fibre-pair = **40,000 Gbps per cable**

1.3 Conclusion

- ~80 Tbps over 2 new cables between Australia – Singapore, then SNG-China in 2 years time
- Plenty of bandwidth across the globe. Infrastructure not the problem
- just need money to buy enough of it.

2. Q1: Infrastructure options – buy, Internet, hybrid?

2.1. the "Moore's Law" of Fiber

- Fiber price drops ~20% per year
- scale economies - 10x bandwidth only 4x price

Q: (When) Is it feasible to use cheap public Internet rather than dedicated circuits?

2.2 A1: Do pathfinder experiment

Request 1Gbps – 10 Gbps reserved link from existing NREN nets (AARNet, CSTNET)
Buy same as Internet access links in Australia & China
Test traffic & path performance over several months

Table 1. Evaluate and rate criteria:

Performance	•Throughput, Latency/delay,
Quality	•Consistency, Packet-loss, reliability
Ease of Control	
Budget & Cost	
Effect of Compression	
Scalability	•Ease of upgrade, or add more circuits
Complexity of Management	•...and cost

3. Q2: How to distribute data across the SKA network to regional centers?

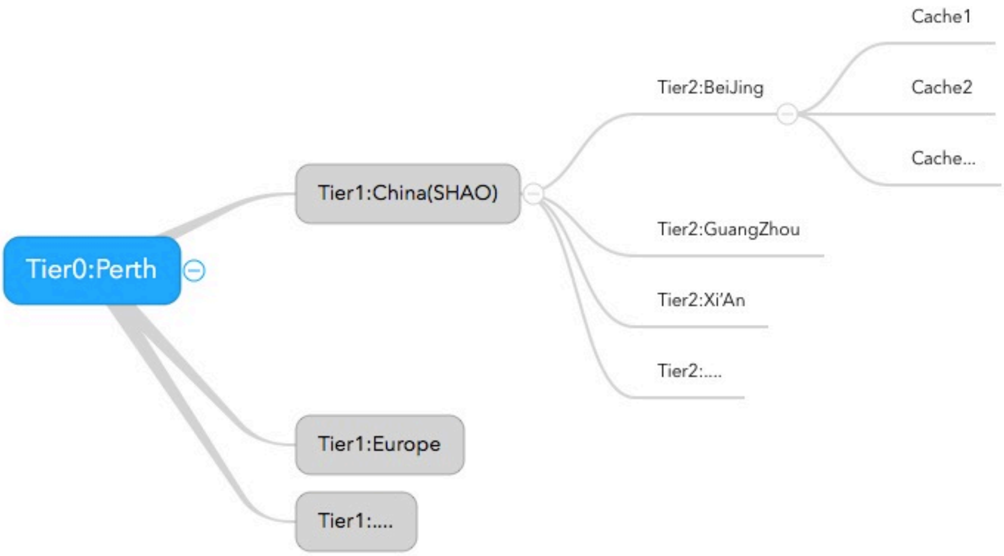


Fig. 4 The Tier Structure of SKA Network

Objective: Each data from Perth will be sent only once, then get replicated between the SRCs, to avoid overloading Perth outbound link

a) Tier0: Perth hold the complete copy of data.

b) Tier1: SRC, SKA REGIONAL CENTER, including: China(SHAO), Europe, etc.

c) Tier2: Data storage at Supercomputing Centers in the region charged by each SRC.

d) Cache: Data chache at each research institute.The purpose of the cache network is to push data to the edge of network, in a CDN (conten delivery network) fashion.

3.1 Candidate methods

a) Peer-toPeer [4]

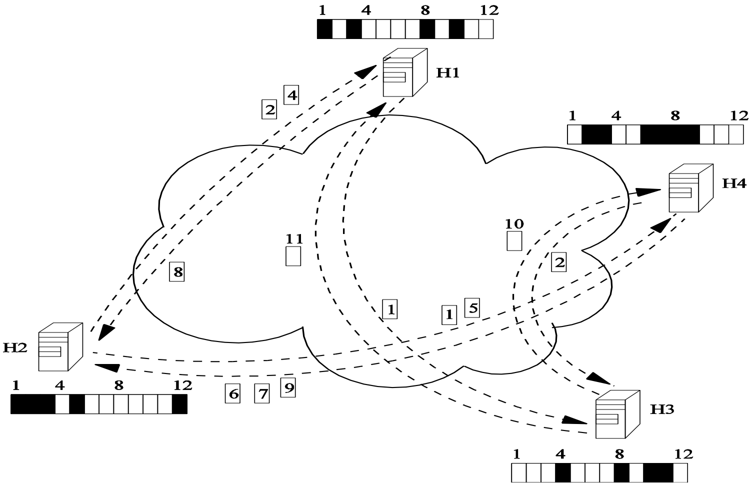
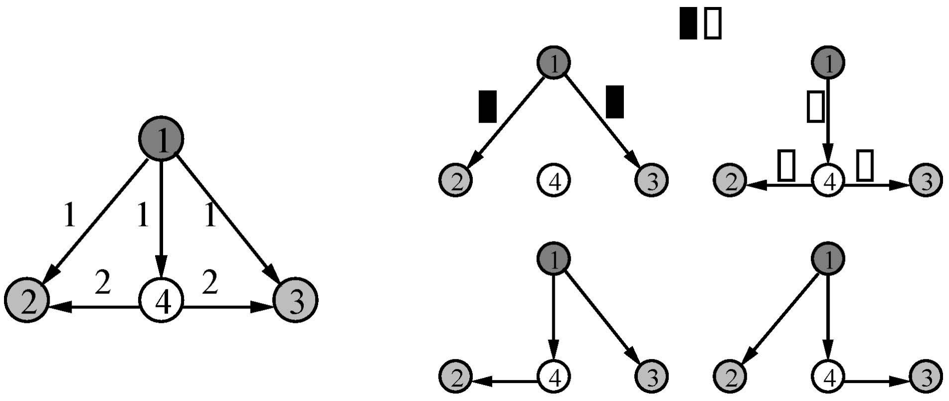


Fig. 5 A snapshot of P2P process

b) Swarming: distribute data over multiple multi-cast trees [5][6]



(a) (b)

Fig. 6 A Swarming example

c) Back-pressure based Packets Delivery [7]

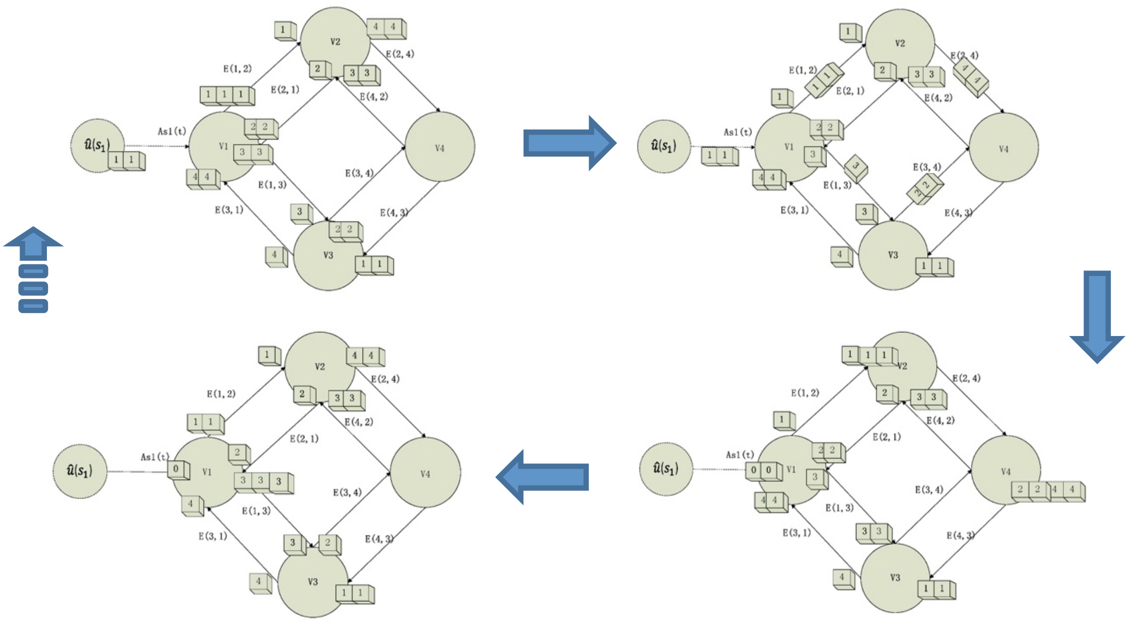


Fig. 7 An example of back-pressured data delivery

d) Other Candidate Solutions

Investigate data transmission principle from distributed file system, for instance, distributed Hash table and Google File System.

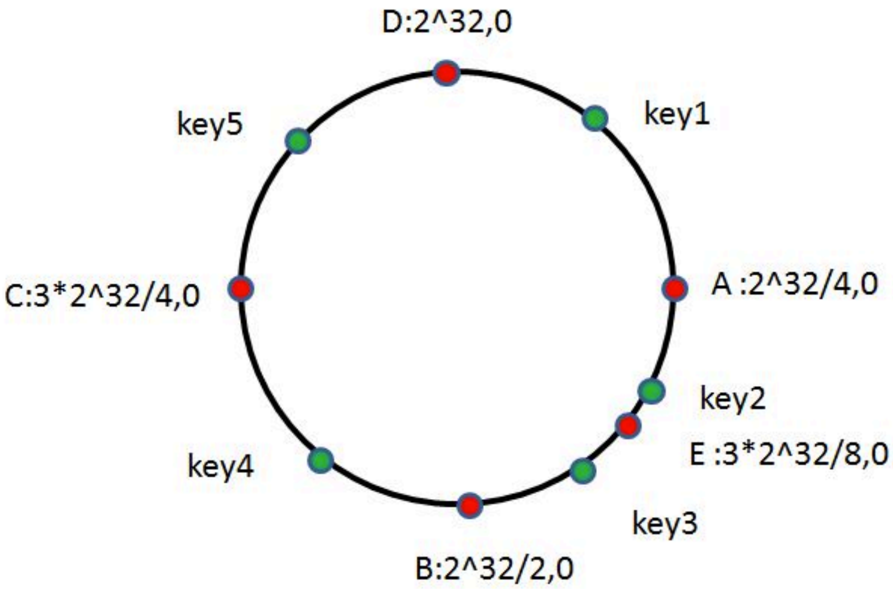


Fig. 8 Distributed Hash Table [8]

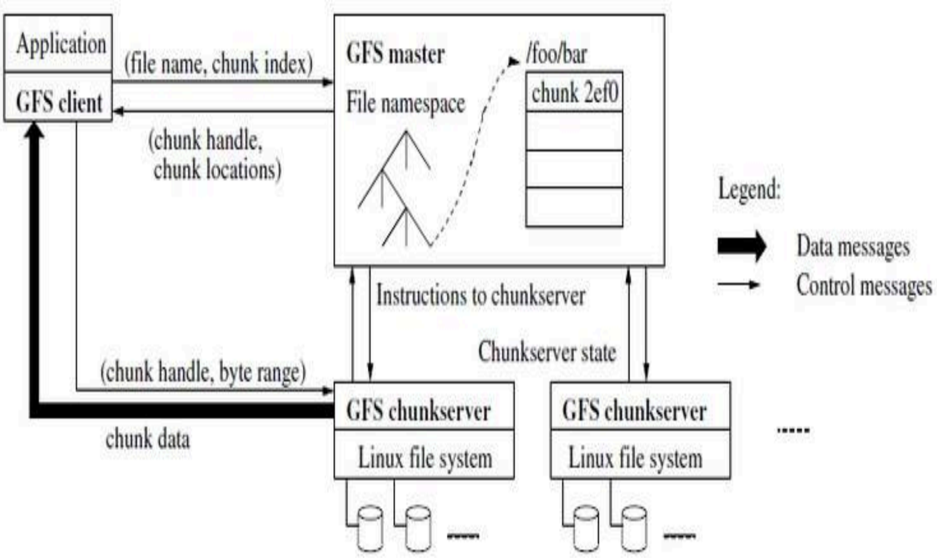


Fig. 9 Google File System [9]

3.2 A2: Evaluate Candidate solutions

The proposed approaches to evaluate candidate solutions are as the following.

- Literature search, check prior work
- Simulation of each candidate method
- Set up Candidate testbed and test all parameters

4. Q3: How to distribute data inside the east Asia region efficiently?

4.1 Borrow ideas from Section 3?

a) the difference between Section 2 and Section 3

- Each T-1 regional center owns up to the complete copy of SKA data; a T-2 node and cache node only store a portion of SKA data
- T-0 to T-1 bandwidth is more expensive than T-1 to T-2 to cache
- More cache nodes in T-2 network

We leave it as open question?

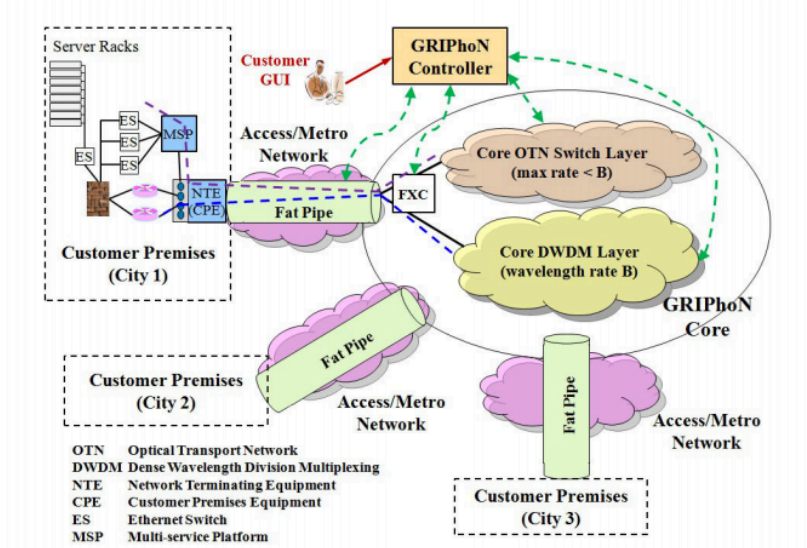
b) Push or pull or hybrid?

- Push data from T-1 node to T-2 or cache nodes
- Cache nodes pull data from T-2 node

c) Move data or move codes?

- Build a few caches near super computing centers and move codes to caches?
- Build more storage caches at the network edge, and move data to users

d) We can also the example of GRIPhoN in [17].

- 

Server Racks
ES
ES
ES
ES
MSP
NTE
(CPE)
Customer GUI
GRIPhoN Controller
Access/Metro Network
Fat Pipe
FXC
Core OTN Switch Layer (max rate < B)
Core DWDM Layer (wavelength rate B)
Customer Premises (City 1)
GRIPhoN Core
Fat Pipe
Customer Premises (City 2)
Access/Metro Network
Fat Pipe
Access/Metro Network
Customer Premises (City 3)
OTN Optical Transport Network
DWDM Dense Wavelength Division Multiplexing
NTE Network Terminating Equipment
CPE Customer Premises Equipment
ES Ethernet Switch
MSP Multi-service Platform

Fig. 10 Example of GRIPhoN

The proposed approaches to evaluate candidate solutions are as the following.

- Literature search, check prior work
- Simulation of each candidate method(pull, or push, or hybrid; move data or move codes)
- Set up candidate testbed and test all parameters(efficiency, bandwidth, latency, robustness)

5. Q4: Intra-datacenter network topologies?

5.1 Some mature candidate intra-DC network topologies

Tree/FatTree[10]， VL2[11], DCell[12]， Portland[13]， BCube[14]， Jellyfish[15]， Jupiter[16]

5.2 Do survey on these candidates and give some suggestion

We will evaluate the following performance metrics, including

- Aggregate throughput
- Load balance
- Routing complexity
- Cost
- Scalability
- Robustness

References

[1] private communication with Nick Rees

[2] http://submarine-cable-map-2016.telegeography.com/

[3] http://www.submarinecablemap.com/

[4]http://www.bittorrent.com/

[5] Xiaoying Zheng, Chunglae Cho, Ye Xia, Algorithms and Stability Analysis for Content Distribution over Multiple Multicast Trees，IEEE Transactions on Parallel and Distributed Systems，2015，26（5）：1217-1227.

[6] Xiaoying Zheng, Chunglae Cho, Ye Xia，Content Distribution by Multiple Multicast Trees and Intersession Cooperation: Optimal Algorithms and Approximations，Computer Networks，2015，83：100-117.

[7] Lin Tong, Xiaoying Zheng, Ye Xia, Mingqi Li，Delay Tolerant Bulk Transfers on Inter-datacenter Networks, IEEE GC16 Workshops CCSNA, 2016.

[8] Frank Dabek. A Distributed Hash Table. PhD thesis, Massachusetts Institute of Technology, 2005.

[9] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. SOSP'03, 2003.

[10]Mohammad Al-Fares, Alexander Loukissas, and Amin Vahdat. 2008. A scalable, commodity data center network architecture. SIGCOMM Comput. Commun. Rev. 38, 4 (August 2008), 63-74.

[11]Albert Greenberg, James R. Hamilton, Navendu Jain, Srikanth Kandula, Changhoon Kim, Parantap Lahiri, David A. Maltz, Parveen Patel, and Sudipta Sengupta. 2009. VL2: a scalable and flexible data center network. In Proceedings of the ACM SIGCOMM 2009 conference on Data communication (SIGCOMM '09).

[12l] Chuanxiong Guo, Haitao Wu, Kun Tan, Lei Shi, Yongguang Zhang, and Songwu Lu. 2008. Dcell: a scalable and fault-tolerant network structure for data centers. SIGCOMM Comput. Commun. Rev. 38, 4 (August 2008), 75-86.

[13] Radhika Niranjan Mysore, Andreas Pamboris, Nathan Farrington, Nelson Huang, Pardis Miri, Sivasankar Radhakrishnan, Vikram Subramanya, and Amin Vahdat. PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric. In Proceedings of the ACM SIGCOMM 2009 conference on Data communication (SIGCOMM '09).

[14] Chuanxiong Guo, Guohan Lu, Dan Li, Haitao Wu, Xuan Zhang, Yunfeng Shi, Chen Tian, Yongguang Zhang, Songwu Lu. BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers. In Proceedings of the ACM SIGCOMM 2009 conference on Data communication (SIGCOMM '09).

[15] Ankit Singla, Chi-Yao Hong, Lucian Popa, and P. Brighten Godfrey. 2012. Jellyfish: networking data centers randomly. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation* (NSDI'12). USENIX Association, Berkeley, CA, USA, 17-17.

[16] Arjun Singh, Joon Ong, Amit Agarwal, Glen Anderson, Ashby Armistead, Roy Bannon, Seb Boving, Gaurav Desai, Bob Felderman, Paulie Germano, Anand Kanagala, Hong Liu, Jeff Provost, Jason Simmons, Eiichi Tanda, Jim Wanderer, Urs Hölzle, Stephen Stuart, Amin Vahdat. Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network. Communications of the ACM, Vol. 59 No. 9, Pages 88-97.

[17] Mahimkar A, Chiu A, Doverspike R, et al. Bandwidth on demand for inter-data center communication[C]//Proceedings of the 10th ACM Workshop on Hot Topics in Networks. ACM, 2011: 24.