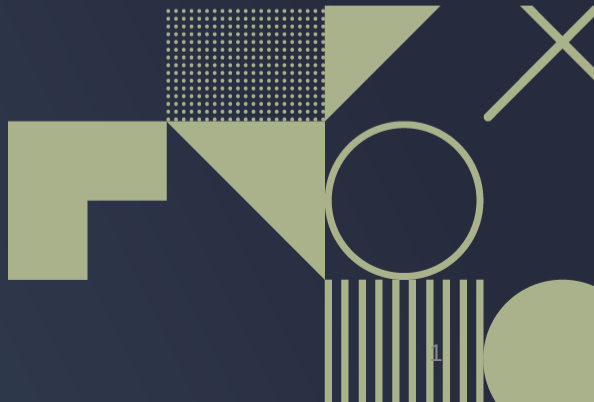
Bridging Network and Parallel I/O Research for Improving Data-Intensive Distributed Applications

Debasmita Biswas*, Sarah Neuwirth‡, Arnab K. Paul†, Ali R. Butt*

**Virginia Tech, ‡Goethe-University Frankfurt, †Oak Ridge National Laboratory*

{debasmita17, butta}@vt.edu, s.neuwirth@em.uni-frankfurt.de, akpaul@vt.edu

Presented by:
Debasmita Biswas,
Virginia Tech



Overview

- *Could Storage and Network Research be Related?*
- *Is there enough work that address this gap?*
- *Survey Technique*
- *Network And I/O Research*
- *Network Types*
- *Network Components*
- *Network Architecture*
- *Network Services*
- *Network Properties*
- *Network Performance Evaluation*
- *Key Insights And Research Challenges*
- *Conclusion*

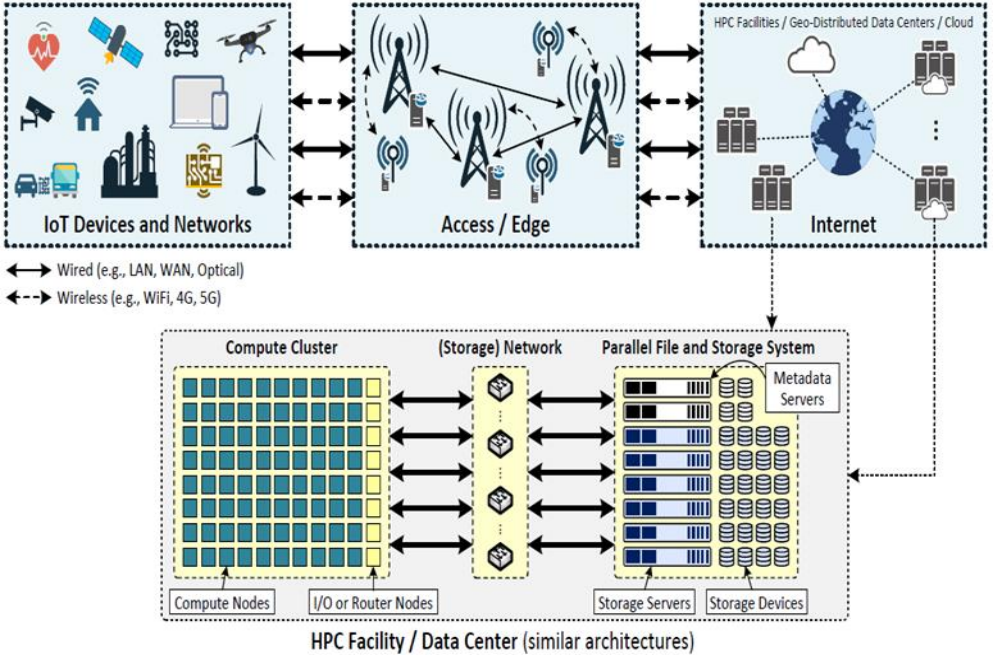
Could Storage and Network Research be Related?

- I/O capability is a major factor deciding an HPC storage system's merit.
- Networking- one of the main components in a distributed storage system in HPC– ***transmissions, internode communications, client to server communications***
- ***Modern Workloads are data intensive***

- Traditional methods of boosting I/O in HPC storage systems by scaling up resources may fall short.
- Is there is a direct relationship between network and HPC storage optimization research?
- Let's investigate!

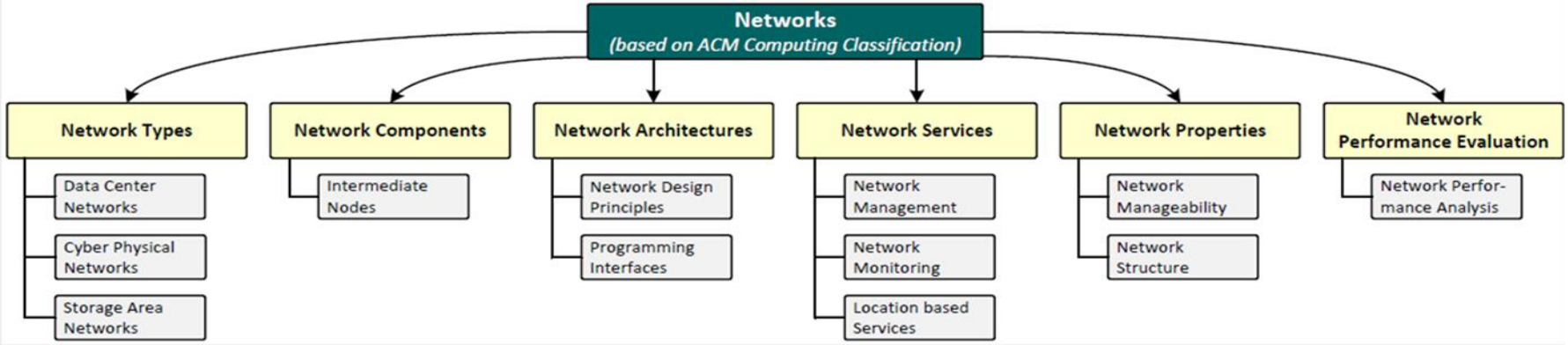
Is there enough work that address this gap?

- Emerging workloads exhibit different *I/O patterns*.
- Large scale data intensive *stresses the underlying network component.*
- Coffee File System- *network parameters for I/O optimization*
- *Fine grained routing* for to pair *Lustre clients* to their closest routers
- Research on accelerating network communication & I/O-
  does not address any direct relationship between Networking optimization and Storage Optimization



Survey Technique

- **Focus:** Research on Network Optimization that can also contribute towards HPC Storage optimization
- **Years:** 2015 to 2021
- **Classification Tree:** ACM Computing Classification System
- **Sources:** ACM Digital Library, Google Scholar
- **Keywords:** ***Network Optimization, HPC Storage Systems, Datacenters, Storage Area Network, IoT network, Edge, I/O optimization and, Data-intensive applications/workloads***



Network And I/O Research

- A subset of the ACM network classification is used to group publications on network optimization.
- We describe the optimization techniques borrowing from the research pertaining to each network classification
- Argue how they can possibly be applied to I/O optimization research.
- Categories being:
  - ***Network Types***
  - ***Network Components***
  - ***Network Architecture***
  - ***Network Services***
  - ***Network Properties***
  - ***Network Performance Evaluation***

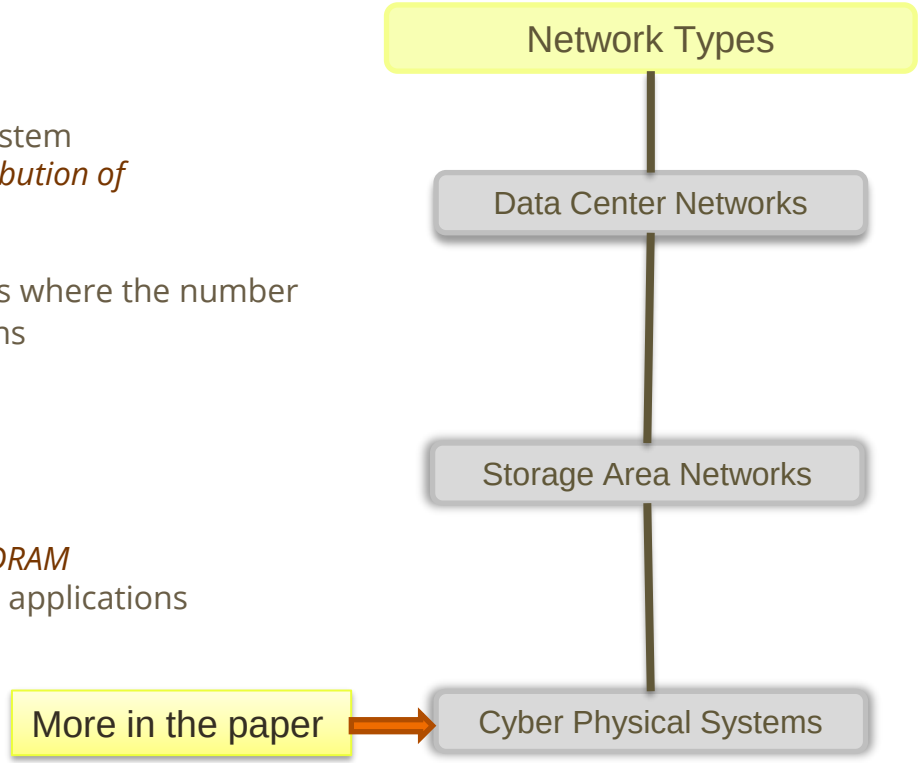
Network Types

- Data Center Networks:
  - **CliqueMap:** A hybrid RMA/RPC caching system
  - Highlights the *I/O benefits from careful distribution of work between RPC and RMA*

  ✓ Can be applied in data streaming applications where the number of read operations supersede write operations

- Storage Area Networks
  - **BlueDBM**: use distributed flash storage- *a low cost and energy efficient alternative to DRAM*
  - Can boost storage I/O for complex big data applications

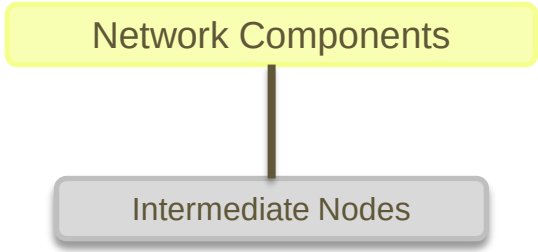
  ✓ May find application in local data centers



Network Components

- **Argo:** a user space distributed shared memory system
- Can *lower latency* produced due to communications between distant nodes
- Facilitates *faster synchronization between nodes*

- **NICE** (network- integrated cluster-efficient): *reduces network latency during request routing*
- Implements of *a ring of virtual storage nodes* in a Network Oblivious (NOOB) Storage system architecture.
- Leverages *SDN* (Software Defined Network)

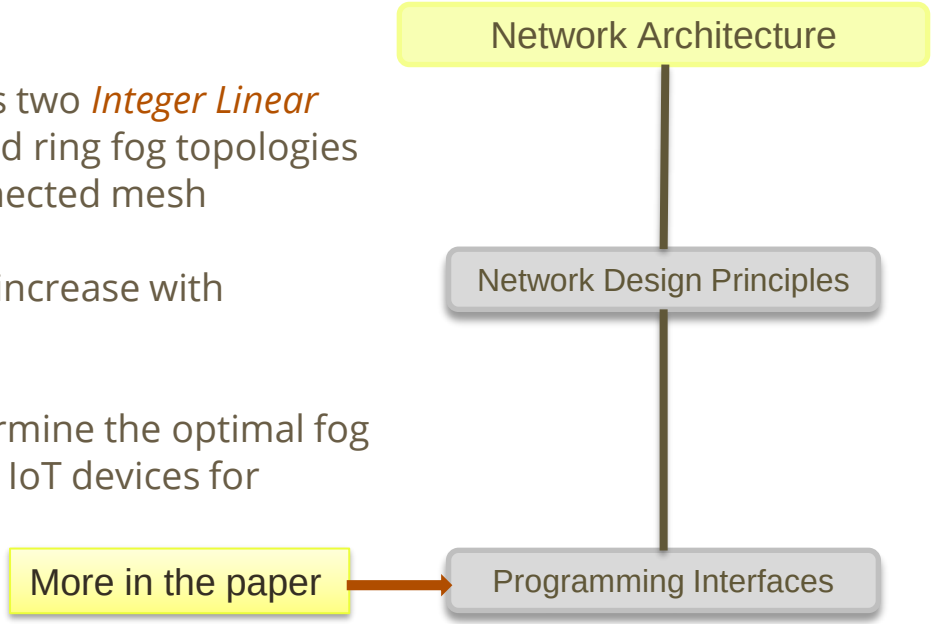


✓ May find application HPC centers embracing SDNs like in *data centers and IoT*

Network Architecture

- Network Design Principles:
  - Proposes, implements and evaluates two *Integer Linear Programming* (ILP) models on star and ring fog topologies
  - *Star topology* outperforms fully connected mesh topology
  - *Ring topology* costs can theoretically increase with increasing system complexity

  ✓ The ILP models can be used to determine the optimal fog topology between HPC systems and IoT devices for data intensive workloads



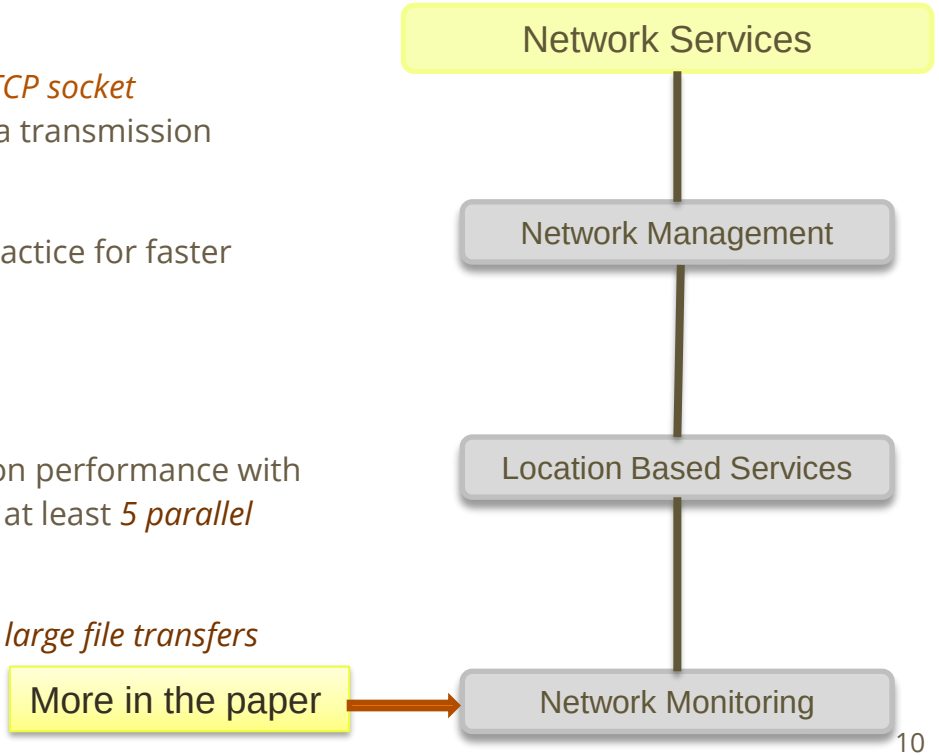
Network Services

- Network Management:
  - *Bandwidth-Delay-Product* to predict the optimal *TCP socket buffer size and the number of TCP streams* for data transmission
  - **BDP ≤ buffer × streams**

  ✓ Will be helpful once SDN becomes a common practice for faster data transmission

- Location Based Services:
  - BeeGFS: streaming *834GB*, best data transmission performance with *4-8 nodes*, connected by *InfiniBand* over *GridFTP*, at least *5 parallel TCP streams*,*16 MiB TCP socket buffer size*.

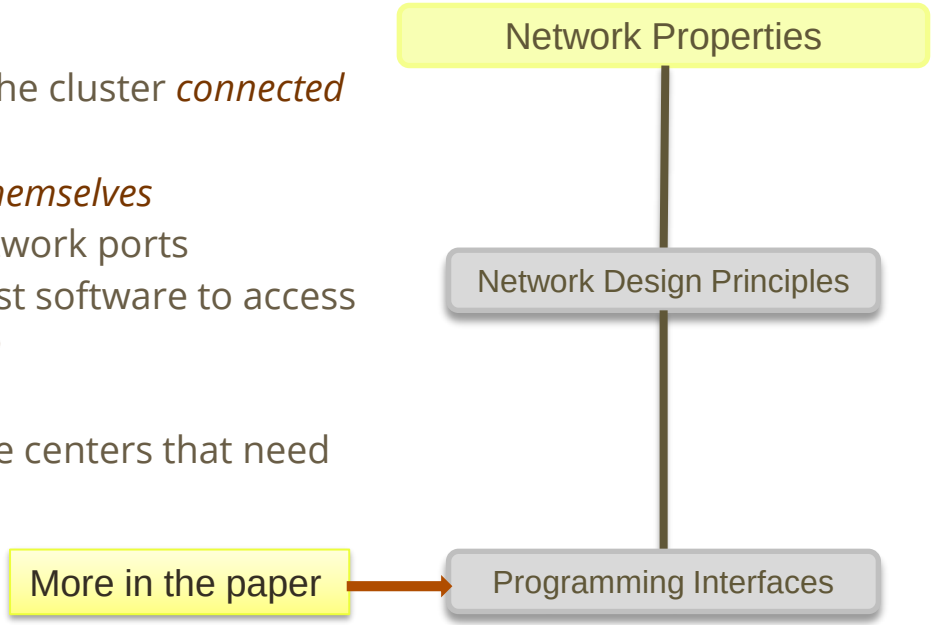
  ✓ Can help storage facilities using BeeGFS for *very large file transfers*



Network Properties

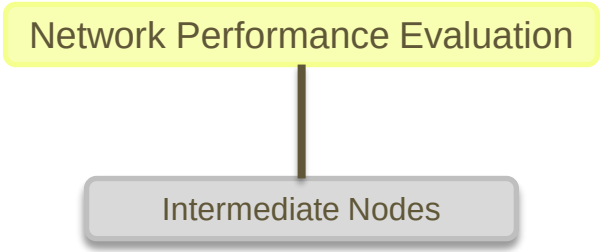
- Network Structure:
  - **BlueDBM:** each storage device in the cluster *connected with serial links*
  - Forming a *separate network among themselves*
  - Each storage device has multiple network ports
  - *Removes overhead* of going to the host software to access individual storage devices, *boosts I/O*

  ✓ Can be adapted in large scale stoarge centers that need *superfast data access*



Network Performance Evaluation

- Addresses the challenges imposed by the three popular in **NTC** (Network Traffic Control)
- *Deep learning* based method to classify the network traffic in *communication systems and networks*
- Outputs generate a final prediction- *Average accuracy of 98%* on the Cambridge Internet Traffic dataset



✓ May be applied to *data-intensive applications generating erratic network traffic patterns* like in IoT, shared high performance computing facilities for scientific research

✓ Evolving workloads that rely on *SDNs* to boost storage system performance *by efficiently analyzing the network utilization and dynamically adjusting the networking parameters* for maximum I/O.

Key Insights And Research Challenges

- Software Defined Networks
- Configuring the network based on the relationship between BDP, number of TCP streams and TCP socket buffer size to optimize throughput for large data transmission in geo-distributed data centers
- Network Load Balancing

- Challenges:
  - ***Complexity***
  - ***Monetary cost***
  - ***Temporal cost***
  - Determining the ***best design approach for non-homogeneous workloads*** hosted on a single HPC storage cluster.

Conclusion

- We present a brief snapshot of the recent *network research landscape targeting data-intensive science applications from a network perspective*.
- We have tried to identify *possible synergy effects between network and parallel file and storage system research*.
- A *realtionship between Network optimization and Storage optimization research* does exist.
- It is worth exploring how these two research areas can work together towards boosting I/O performance in an HPC facility.

Thank you!

debasmita17@vt.edu, s.neuwirth@em.uni-frankfurt.de,
akpaul@vt.edu, butta@vt.edu

Questions?