Efficient and Reliable Network Tomography in Heterogeneous Networks Using BitTorrent Broadcasts and Clustering Algorithms

Kiril Dichev[1] Fergal Reid[2] Alexey Lastovetsky[1]

[1]School of Computer Science
and Informatics
University College Dublin

[2]Clique Research Cluster

November 13th, SC12, Salt Lake City

Outline

Introduction

- Network properties significantly impact communication
- Communication libraries can use knowledge of network for more efficiency
- Various examples in HPC and distributed computing:
  - Early work includes MPI implementations for heterogeneous networks (MagPIe, PACX-MPI)
  - More recently – work on topology-aware collectives for multi-core clusters

Introduction

- Large body of work on efficiently using networks we know *a priori*
- But what if no *a priori* knowledge is available for complex networks (e.g. clouds or grids)?
- Discovery of network properties in heterogeneous networks:
  - Simple communication model (latency, bandwidth)
  - Isolated experiments for parameters at each link
  - Useful for communication on heterogeneous networks
- Isolated benchmarks do not reflect network properties during intense collective communication

Network Discovery and Network Tomography

- What about distributed computing?
- Network discovery has a long history and can involve all available components of a network
- More recently (late 90s), a sub-area called “network tomography” has emerged
- In network tomography, network properties are discovered only using end-to-end measurements

Two Phases in Network Tomography

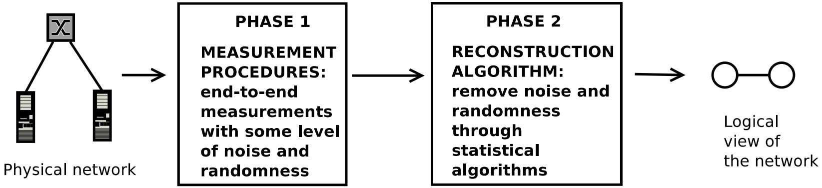


Figure: Network tomography can be considered as a two-phase approach

Network Tomography with Bandwidth as a Metric

- Majority of existing work covers metrics like delay and accessibility
- Work on discovering available bandwidth is limited
- Our contribution addresses the problem of “Multiple source, multiple destination network tomography”
  - Many peers simultaneously exchange large data volumes
  - What is the achievable bandwidth between each pair?

State-of-the-Art in Bandwidth Tomography

Recent work [1]:

- Objective: Establish logical links between nodes and capacity of each link
- Each benchmark is expensive in its nature
- Measurement procedure requires separate benchmarks for all triplets of nodes: $O(n^3)$
- Statistical analysis with acceptable runtime
- Real life experiments not feasible
- Approach only tested with simulation

[1] Bobelin, L., Muntean, T.: Algorithms for network topology discovery using end-to-end measurements

State-of-the-Art in Bandwidth Tomography

Previous work [2]:

- Objective: Find a simplified graph of a physical network reflecting the interferences of streams
- Measurement procedure for best-case scenarios (no interference): $O(n^2)$
- Real life experiments not feasible even for moderate node numbers
  - Very limited set of experiments – "about one hour for 20 nodes"
- Approach tested with simulation

[2] A. Legrand, F. Mazoit, M. Quinson: An application-level network mapper

Problem Statement

- The measurement procedures are inefficient, and yet focus is often on the reconstruction algorithm
- Due to their high complexity, existing MSMD network tomography methods are not practical

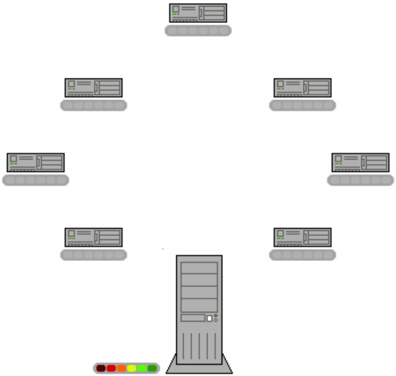
Proposed Solution

We propose a different MSMD tomography solution:

- For the measurement procedures, we use a highly efficient BitTorrent protocol
- As reconstruction algorithm, we employ reliable clustering techniques

BitTorrent Overview

Example

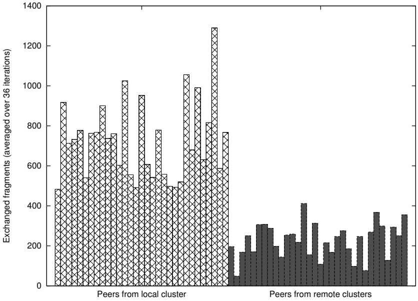


Source: Wikipedia

BitTorrent Overview

Why BitTorrent?

- BitTorrent protocol exploits available bandwidth well
- A BitTorrent client opens a number of parallel connections with many peers
- More data flows along faster connections



Measurement Method

Idea: Measure peer-to-peer traffic in BitTorrent (Analogy: Flow of water in pipes)

- Let a BitTorrent broadcast be a synchronized distribution of a file from one peer to the rest
- All peers are instrumented to record incoming and outgoing fragments
- Define metric between two nodes v_1 and v_2 as the count of exchanged fragments within a BitTorrent broadcast:

$$w((v_1, v_2)) = v_1 \rightarrow v_2 + v_2 \rightarrow v_1 \tag{1}$$

Challenges with Chosen Metric

- BT "can be" very efficient: it is observed to scale as O(n)
- But it introduces a high degree of randomness and non-determinism
- There are two possible ways to address this issue:
  - Perform a number of iterations
  - Use a really good statistical algorithm
- Combining both is the best option

Clustering algorithm

- Our objective: Be useful to communication libraries (including MPI):
  - Cluster together nodes that can sustain high bandwidth even under heavy communication
  - Identify bottlenecks under heavy communication and separate nodes accordingly into different clusters
- Choice of clustering algorithm not obvious
- Based on experimental results, we chose the modularity-based clustering

Modularity-Based Clustering

- The modularity method is defined by following objective:

$$Q = \sum_i \left(e_{ii} - a_i^2\right) = Tr(e) - \left\|e^2\right\| \tag{2}$$

- e_{ii} – fraction of edges that would be intra-cluster in cluster i
- a_i – fraction of inter-cluster edges connecting to cluster i in a randomized model
- Larger Q indicates stronger community structure
- Maximal Q gives best clustering

Gluing the Pieces Together

- "Exchanged fragments" metric w used as input for clustering algorithm
- We choose the Louvain method as the algorithm to find a set of clusters that maximise the modularity criterion

Experimental Setup

- All experiments are performed on Grid'5000 infrastructure
- Runs involve 1,2,3 or 4 clusters at single site or different sites
- Single BT broadcast chosen (arbitrarily) for a fixed size 240 MB dummy file
- Graphviz visualization indicative of quality of clustering

Example: Bordeaux Topology

Three clusters, One Major Bottleneck

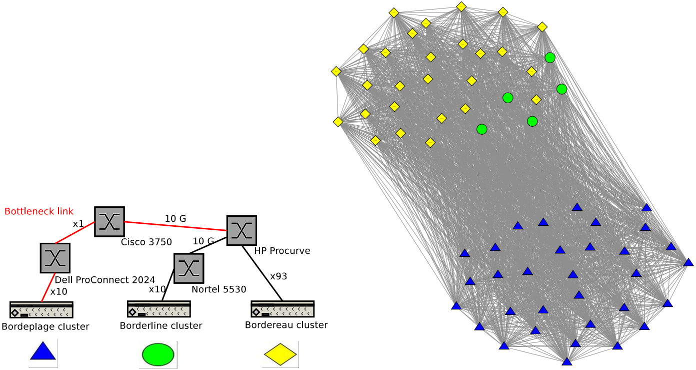


Figure: Three clusters within Bordeaux site, one bottleneck. Note: Used visualization uses each edge weight as a spring

Example: Two Clusters on Different Sites

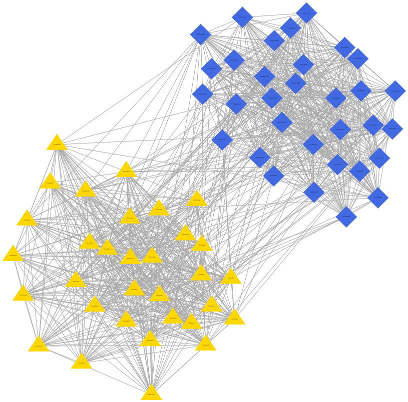


Figure: Two distributed clusters – Grenoble and Toulouse

Example: Three Clusters on Different Sites

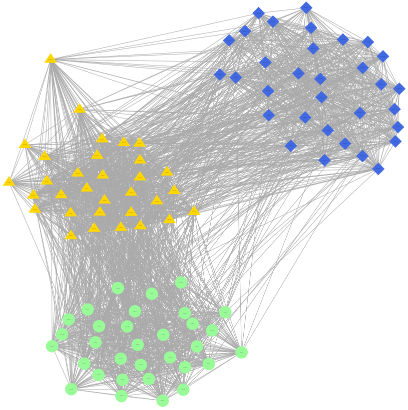
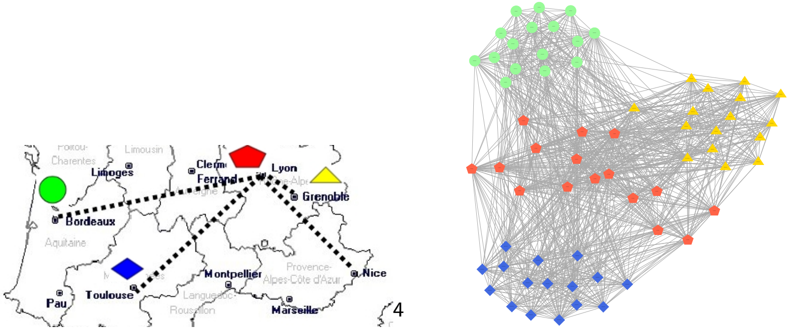


Figure: Three distributed clusters – Bordeaux, Grenoble and Toulouse

Example: Four Clusters Across France



[4]

Figure: Four distributed clusters – and an interesting recognition of star topology

[4] Source: Grid5000 Webpage

Measuring Accuracy of Proposed Solution

- A measure called NMI (normalized mutual information) is common in clustering algorithms
- This index compares the **ground truth** – the a-priori knowledge of the network – against the **clustering results**
- We “borrow” NMI to measure the accuracy of the proposed network tomography

Ground Truth

How do we produce our ground truth, i.e. our *a priori* knowledge on the network?

- Intra-site network:
  - Documentation (Wiki) – sometimes not reliable
  - Network administrator – reliable source of information
- Inter-site network (optic fiber):
  - Documentation – generally reliable
  - But we still perform NetPIPE benchmarks

How Efficient and Reliable is the Proposed Approach?

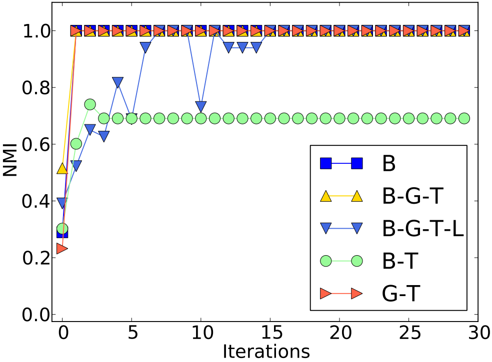


Figure: The NMI quickly converges as the number of measurement iterations increases

How Efficient and Reliable is the Proposed Approach?

- Results demonstrate that nearly all runs converge to perfect accuracy after at most 15 BT broadcasts
- Efficiency:
  - Each of the BT broadcasts requires around 20 seconds for 64 nodes (even when geographically distributed)
  - At most 5 minutes in total for full accuracy with 64 nodes
  - Related work would need more than 10 hours for measurement on similar setup

Conclusion

- We presented a new method of network tomography
- Both the BitTorrent-based measurement and coupling with a clustering algorithm are unconventional
- Randomness and non-determinism of BitTorrent easy to overcome through iteration
- Clustering algorithm provides reliable results
- Due to efficiency of measurement procedures, proposed solution is the only one that can be used for real platforms

Thank you!