On reliable data collection in large networks

Chadi BARAKAT

INRIA Sophia Antipolis, France
Planète research group

Email: Chadi.Barakat@sophia.inria.fr
WEB: http://planete.inria.fr/chadi

Problem statement

- Many situations in which one is interested in collecting information from a large number of sources spread over the Internet.
  - Measurements collected by hosts, routers, sensors or traffic capture devices.
  - Data generated by the different sites of an enterprise.
  - etc.
- Challenge: This collection, if done simultaneously, would congest the network and cause implosion at the collector.
- A transport solution is needed …

Framework for the study

- We look for end-to-end solution
  - No intermediate nodes are deployed to aggregate the information (as in ConCast for example).
- A priori, information to be entirely collected.
  - Need for reliability (sampling-based solutions don't work).
  - But one can lessen this requirement if needed.
- Congestion control
  - Avoids congestion of the network and the collector access link.
  - Do it as much friendly with TCP as possible.
- Mainly focus on small volume of data per source
  - Using TCP is no longer optimal due to three-way handshake overhead and slow start.

TICP: TCP-friendly Information Collection Protocol

- Initiated within a collaboration with Hitachi Sophia Antipolis, then continued with the help of:
  Karim Sbai (ENSI)
  Amaury Decreme (EPU)
  Mohammad Malli (PhD Planète)
- More information on http://planete.inria.fr/chadi/ticp
- Basic idea:
  - A central collector knows about all sources.
  - It probes them, they answer directly with their report packets containing their information.
  - The collector controls the probing rate. Retransmits probes in case of losses. Verifies reliability.

Protocol in brief: Congestion control

- A window-based flow control:
  - `cwnd`: maximum number of sources the collector can probe before receiving any information.
- The collector increases `cwnd` and monitors at the same time the loss ratio of reports (during a time window in the past).
  - The protocol has two modes: slow start and congestion avoidance.
- Congestion of the network is inferred when the loss ratio of reports exceeds some threshold.
- Upon congestion, divide `cwnd` by 2, and restart its increase.

Protocol in brief: Error Control

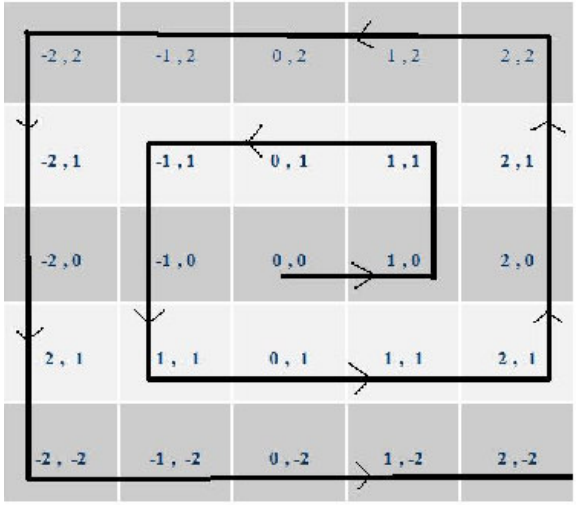
- The protocol is reliable in the sense that it ensures that all sources have sent their reports.
- To reduce the duration of the session:
  - In the first round, the protocol probes all sources
    - Order to be defined later.
  - In the second round, the protocol probes sources whose reports were lost in the first round.
  - In the third round, the protocol probes sources whose reports were lost in the first two rounds.
  - Continues in rounds until all reports are received.

Measuring the loss ratio

- The source disposes of a timer, called TO:
  - The timer is set to SRTT + 4 RTTVAR, where SRTT is the average round-trip time, and RTTVAR its mean deviation.
  - RTT is calculated over all sources (time and space dimensions).
  - The timer is rescheduled every time it expires.
  - The value of the timer can be seen as an upper bound on RTT.
- The timer serves to measure the loss rate.
  - All reports sent during one cycle of the timer have to arrive during the next cycle at the latest, otherwise they are supposed lost.

Ordering of sources

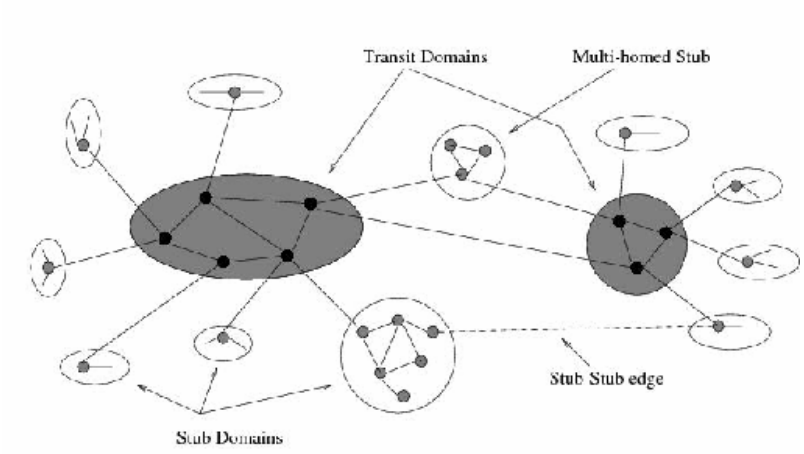
- Random, topology independent
  - Inefficient.
  - Hard to handle multiple bottlenecks at once.
  - RTT hard to predict (bad setting of the timer).
- Topology dependent
  - Cluster sources and rank clusters from closest to the collector to the farthest.
  - Use this ordering to probe sources.
  - Sources inside a cluster probed randomly.
  - We use Internet coordinates for clustering.



INRIA SOPHIA ANTIPOLIS

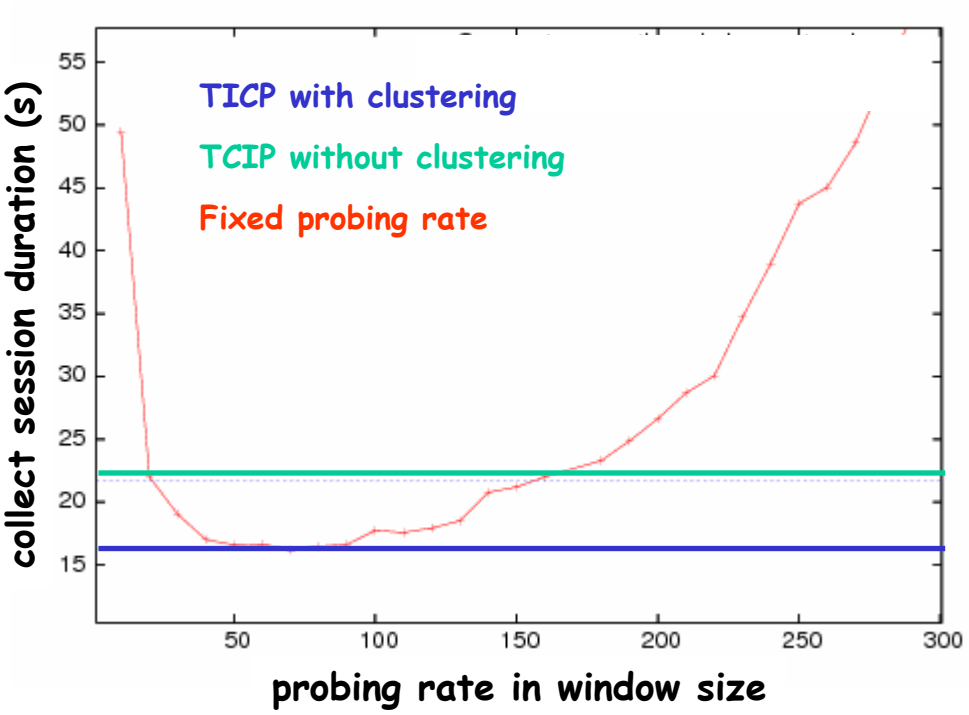
Performance of TICP: sample of results

- 500 sources of information generating a packet each.
- Cluster size equal to 50 ms, its optimal value over this topology.



This is an ns-2 simulation.

PlanetLab experiments show same behavior.
See technical report for more details.



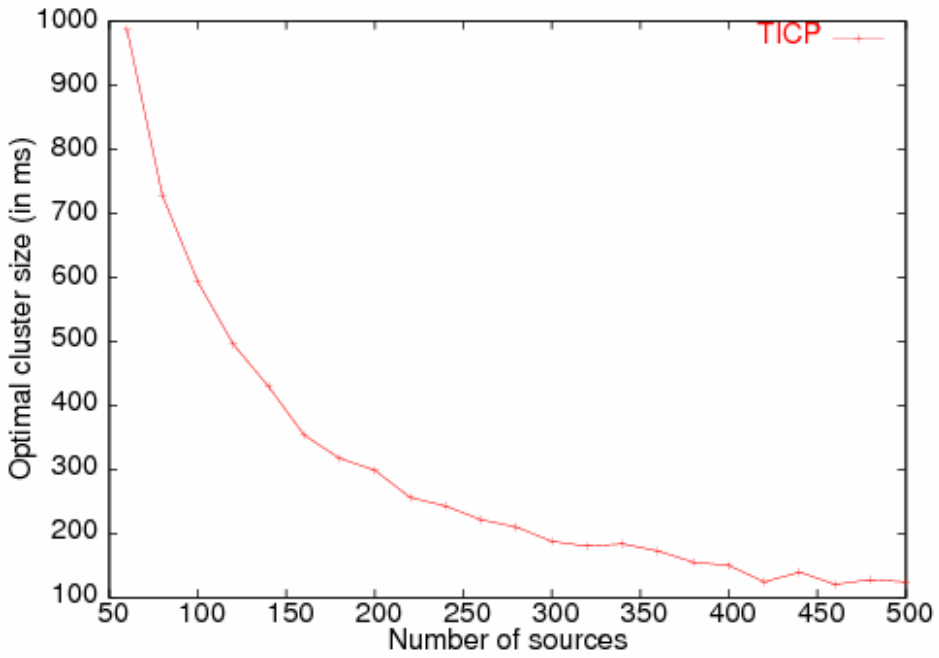
INRIA SOPHIA ANTIPOLIS

Cluster size

- Important parameter of the protocol to set.
- Open issue …
- Our observation: As the number of sources increases, it converges to some constant value function of the underlying topology.

For example, over PlanetLab,

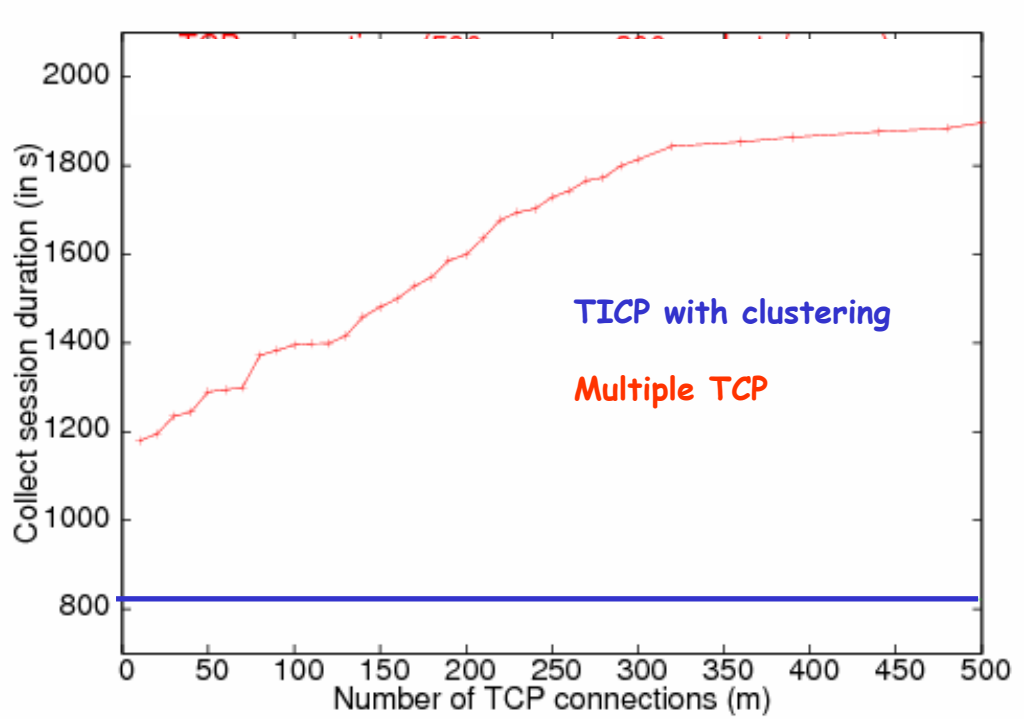
100 ms is a good choice …



INRIA SOPHIA ANTIPOLIS

Compared to parallel TCP

- What if parallel TCP connections are used to collect ?
- TICP behaves better due to its multiplexing capability.
  - See it as multiple TCP connections with one congestion window.
- Simulation results …



INRIA SOPHIA ANTIPOLIS

Ongoing work

- We also studied the delegation issue where the central collector can ask other sources to collect on its behalf.
  - A kind of two level collection.
  - Major observation: one level is enough until some threshold, beyond it delegate to as many proxy collectors as the threshold.
- We also started to use TICP for network probing architectures.
  - Probing can be seen as information collection !
  - We have nice results …
- ns code exists.
- C++ code exists. Only delegation still to be implemented.