Streaming Algorithms for Cosmological Simulations and Beyond

Vladimir Braverman
Johns Hopkins University

supported in part by the National Science Foundation under Grant No. 1447639, by the Google Faculty Award and by DARPA grant N66001-1-2-4014.

Joint works with
Alexander S. Szalay, Zaoxing Liu, Greg Vorsanger, Vyas Sekar, Nikita Ivkin, Lin F. Yang, Mark Neyrinck, Gerard Lemson, Tamas Budavari, Randal Burns, Xin Wang, Stephen Chestnut, Harry Lang, Keith Levin...
What are Streaming Algorithms?

Goal: Compute $F(D)$

F(D) \approx F'(S)
How does it work?
Is X equal to Y?
Streaming Sketch

- If $h_1, \ldots, h_n$ are i.i.d., $h_i \sim U([-1,1])$

Compare the inner products:

$$\sum_{i=1}^{n} x_i h_i \quad \sum_{i=1}^{n} y_i h_i$$
New Theory

• The Johnson-Lindenstrauss Lemma and metric embedding
• Stable Distributions and Pseudorandom generators
• Dvoretzky Theorem (local theory of Banach spaces)
Algorithms for:

- Clustering
- Sliding Windows
- Correlations
- Trends
- Frequent Events
- ...

Streaming Algorithms for Halo Finders

Zaoxing Liu, Nikita Ivkin, Lin F. Yang, Mark Neyrinck, Gerard Lemson, Alexander S. Szalay, Vladimir Braverman, Tamas Budavari, Randal Burns, Xin Wang

Johns Hopkins University, Baltimore, MD, USA
Cosmological Simulations

Simulation:

• is a gravitational evolution of the system of particles
• provides distribution of particles in space and time
• helps to understand the processes of forming galaxies
Halo

In terms of Physics:
• Galaxies are thought to form in halos

Defining property:
• Macro structure with high mass concentration
Halo finding algorithms

1974 SO Press & Schechter
1985 FOF Davis et al.
1992 DENMAX Gelb & Bertschinger
1995 Adaptive FOF van Kampen et al.
1996 IsoDen Pfitzner & Salmon
1997 BDM Klypin & Holtzman
1998 HOP Eisenstein & Hut
1999 hierarchical FOF Gottloeber et al.
2001 SKID Stadel
2001 enhanced BDM Bullock et al.
2001 SUBFIND Springel
2004 MHF Gill, Knebe & Gibson
2004 AdaptaHOP Aubert, Pichon & Colombi
2005 improved DENMAX Weller et al.
2005 VOBOZ Neyrinck et al.
2006 PSB Kim & Park
2006 6DFOF Diemand et al.
2007 subhalo finder Shaw et al.
2007 Ntropy-fofsv Gardner, Connolly & McBride
2009 HSF Maciejewski et al.
2009 LANL finder Habib et al.
2009 AHF Knollmann & Knebe
2010 pHOP Skory et al.
2010 ASOHF Planelles & Quilis
2010 pSO Sutter & Ricker
2010 pFOF Rasera et al.
2010 ORIGAMI Falck et al.
2010 HOT Ascasibar
2010 Rockstar Behroozi

Cumulative number of halo finders as a function of time

The Halo-Finder Comparison Project
[Knebe et al, 2011]
Streaming Solution:

Our goal:
- Reduce halos finding problem to one of the existing problems in streaming setting
- Apply ready-to-use algorithms

haloes ≈ heavy hitters?
- To make a reduction to heavy hitters we need to discretize the space.
- Naïve solution is to use 3D mesh:
  - Each particle now replaced by cell id
  - Heavy cells represent mass concentration
  - Grid size is chosen according to typical halo size
Memory

• Dataset size: $\sim10^9$ particles
  • Any in-memory algorithm: 12 GB
  • Pick-and-Drop: 30 MB

• GPU acceleration
  • One instance of Pick-and-Drop algorithm can be fully implemented by separate thread of GPU
  • Count Sketch algorithm have two time-consuming procedures: evaluating the hash functions and updating the queue. The first one can be naively ported to GPU
Enabling a “RISC” Approach for Software-Defined Monitoring using Universal Streaming

Zaoxing Liu*, Greg Vorsanger*, Vladimir Braverman*, Vyas Sekar†

*Johns Hopkins University, † Carnegie Mellon University
A “RISC” method called Universal Monitoring (UNIVMON)

Calculate different metrics without processing the traffic again
Aggregative queries for OLAP

- HyperRoll: Start-up company, Israel
- Acquired by Oracle in 2009
- From *InformationWeek*, September 2009:

  “...HyperRoll has acquired customers in the retail, financial services, and consumer goods sectors... Its products can shorten data warehouse loading times and speed up query executions by a factor of 10...”
Applications

- Distributed computing
- Astronomy
- Web applications
- Security
- Compressed Sensing
- Machine Learning
- Databases
- Networking
- NLP

Streaming Algorithms
Thank you

• vova@cs.jhu.edu

• http://www.cs.jhu.edu/~vova