Open source project study

On the complex network clustering using DryadLINQ

Stojan Trajanovski (st508)

MPhil in Advanced Computer Science
Motivation

Why going parallel in complex networks analysis?

• Online social networks, Internet graph
  o millions of users (Facebook, Twitter …)
  o increased computational complexity

• Why is prospective?
  o some actions are fully independent
  o increased hardware performance
    - multi-core
    - network clusters, global cloud clusters
Motivation

Why using PLINQ/DryadLINQ?

• Inherited LINQ behaviour
  o declarative and imperative programming
  o T-SQL syntax in your code
    - no more SQL server store-procedures
    - optimized performance
    - inherited SELECT, GROUP/ORDER BY

• + Dryad/Parallel processing
  o optimized job management
Why not (mainly pure technical reasons)?

- problems even with Microsoft concepts
  - requires .NET environment anyway
  - evaluated only on newest Microsoft OSs
  - head node:
    - Windows Server ’08 OS (problems with ’03)
    - more than 500G HD, 8 MB memory
  - computational nodes (at least Windows 7)
  - no Windows Azure support 😞
  - Someone mentioned Linux/MacOS? 😊
My application/solution?

Using PLINQ/DryadLINQ for network clustering?

• K-means clustering
  o parallel performs better
  o the approach:
    - parallelize the method
  o the results
    - significantly better time performance
  o TO DO
    - more clustering approaches, comparison …
Some plots

Parallel vs. non parallel LINQ ($dataset:$)

different values of $N=\{100,200,500,1000\}$
- Questions??

- Short Discussion
  - still work in progress ...