CS 598: Communication cost analysis of algorithms

May 13, 2016

Description: Efficiency and parallel scalability of data-intensive applications are most often constrained by data movement in the memory hierarchy and the network. This course will focus on analysis of algorithms through the lens of communication and synchronization models. Communication lower bounds and algorithms that attain them will be surveyed for fundamental combinatorial and numerical problems. The course will emphasize general analytical techniques, but will also connect to full-scale applications.

Prerequisites: algorithms, linear algebra, and basic parallel programming (e.g. CS 420, 450, 473)

Outline of lectures:

1. quantifying communication and synchronization costs
   - memory hierarchy models: red-blue pebble game, disk I/O models
   - message-passing models: $\alpha - \beta$, BSP, LogP, LogGP

2. cache-efficient algorithms
   - communication cost in divide and conquer algorithms, cache-obliviousness
   - communication bounds for FFT, sorting, and permutation in I/O model

3. collective communication algorithms
   - trees for single-item broadcast and reduction
   - bandwidth-efficient multi-item collective communication, scans

4. avoiding communication in parallel dense linear algebra (I)
   - 2D algorithms: Cannon, SUMMA, LU, QR
   - 3D algorithms for matrix multiplication and dense matrix factorizations

5. avoiding communication in parallel dense linear algebra (II)
   - advanced techniques for QR and the symmetric eigenvalue problem
   - nonsymmetric eigenvalue problem

6. avoiding communication in parallel sparse linear algebra
   - in-time blocking for stencil computations
   - asynchronous algorithms for iterative methods
7. lower bounds on communication in parallel algorithms
   - techniques for analyzing dependencies: graph expansion, rank, volumetric inequalities
   - applications to dense matrix computations and stencils

8. lower bounds on synchronization
   - tradeoffs: work vs synchronization, communication vs synchronization
   - dependency interval growth in matrix factorizations and iterative methods

9. communication-avoiding shortest-path algorithms
   - single-source shortest-paths: tradeoffs from Dijkstra’s algorithm to Bellman-Ford
   - all-pairs shortest-paths: Floyd-Warshall and path-doubling, betweenness centrality

10. graph partitioning
    - dependency graph partitioning and communication cost
    - parallel algorithms for graph partitioning

11. reducing communication in scientific applications: (I) molecular dynamics
    - decompositions: particle, force, spatial
    - lower and upper bounds for communication in spatial decompositions

12. reducing communication in scientific applications: (II) electronic structure calculations
    - dense and sparse matrix computations in Density Functional Theory
    - communication cost of Hartree-Fock (HF) and post-HF methods

13. communication complexity of fast bilinear algorithms
    - lower and upper communication bounds for Strassen’s algorithm
    - communication cost of symmetric tensor contraction algorithms

14. network topologies
    - communication in real networks: wormhole routing, congestion, bisection bandwidth
    - topology-aware collective communication protocols

15. topology-aware communication-avoiding algorithms
    - examples: dense linear algebra on torus networks, divide and conquer on hierarchical networks
    - network oblivious algorithms, topology-dependent lower bounds