High Performance Computing and Data Mining

Performance Issues in Data Mining

Peter Christen

Peter.Christen@anu.edu.au

Data Mining Group

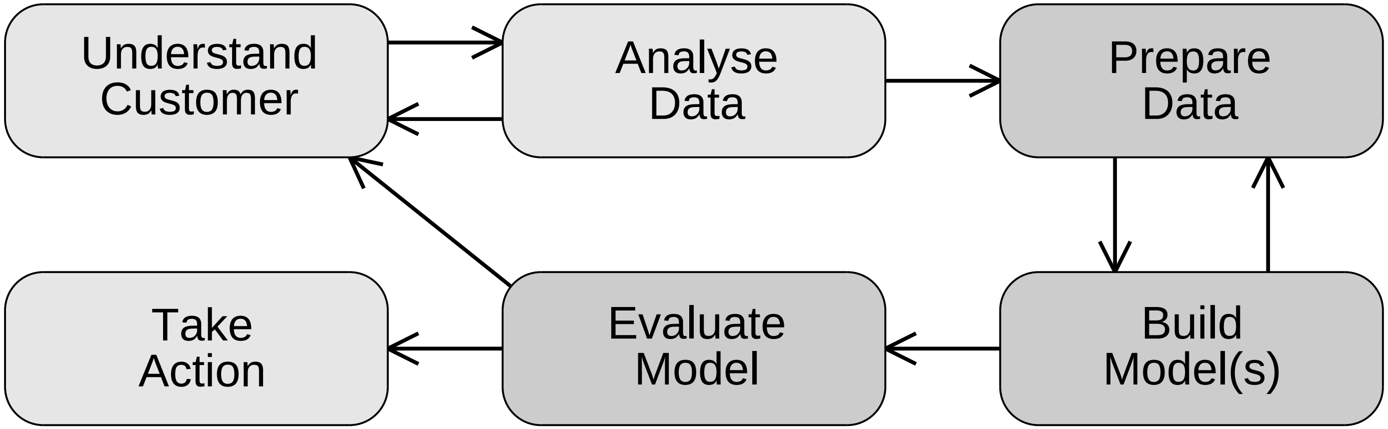
Department of Computer Science, FEIT

Australian National University, Canberra

http://csl.anu.edu.au/ml/dm/

THE AUSTRALIAN NATIONAL UNIVERSITY

The Data Mining Process



- Analysis: Fast data access, large memory, caching
- Preparation: Fast input and output, large memory, fast computing
- Modelling: Fast computing, large memory

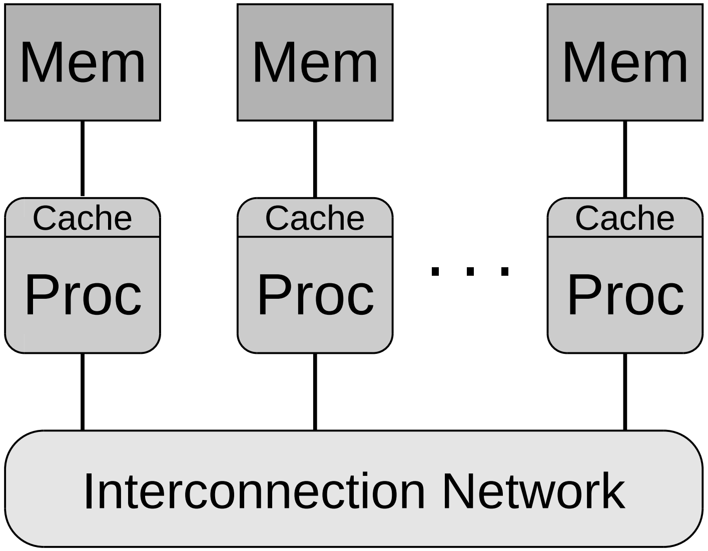
Why High Performance Computing

- Large data collections → Memory and disk space
- Long processing times → Processing speed
- Technical limitations
  - Processor speed
  - Input / output bandwidth
  - Memory size and bandwidth
- Many problems are inherently parallel
- Contemporary high performance computing always involves parallel computing

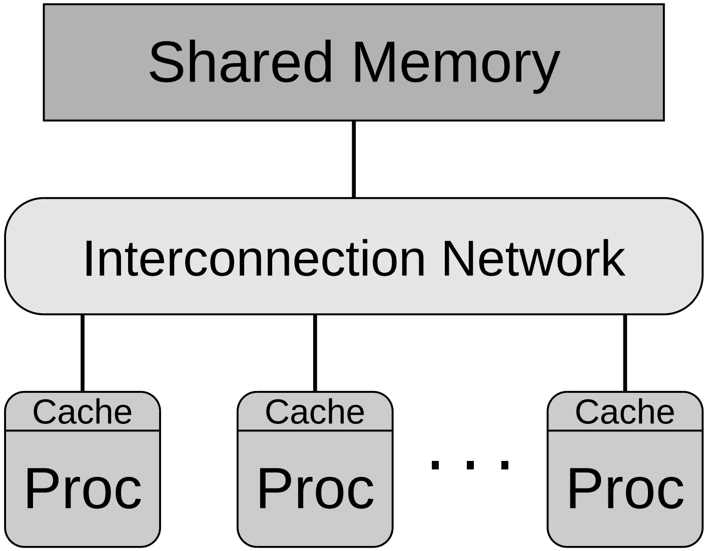
Different Kinds of Parallelism

- Functional parallelism
  - Each processor runs a *sub-job*, the result is passed on to the next processor in line
  - Pipeline principle (assembly-line)
- Data parallelism
  - All processors do the same job on different *sub-sets* of the data
  - Data decomposition

Parallel Computer Architectures



Distributed Memory
Architecture



Shared Memory
Architecture

Parallel Performance

- Goal: Being *P* times faster with *P* processors
  - *Speedup* is usually less than *P*
  - Sequential parts in a program limit speedup
- Scalability
  - Measurement how well speedup scales with increasing number of processors
- Data distribution and load balancing are critical
- Parallel programs need to be tuned for new architectures

ANU Beowulf Linux Cluster Bunyip

- 96 Dual Pentium III nodes
- 36 Gigabytes main memory
- 1,305 Gigabytes disk space
- Fast-Ethernet network
- Gordon Bell prize winner 2000

Australian Partnership for Advanced Computing (APAC)

- ANU Data Mining is 1 of 13 Expertise Programs
  - Conduct research and development projects
  - Provide high-level user support services
- National Facility at ANU opened in May 2001
  - Peak performance close to 1 Tera-Flops
  - 480 Compaq Alpha processors
  - Each with 1 Gigabyte of main memory
  - Connected by a fast, low latency switch
  - Disk capacity around 10,000 Gigabytes

APAC National Facility

THE AUSTRALIAN NATIONAL UNIVERSITY

Research at ANU Data Mining Group

- *DMtools* facilitate analysis and preprocessing
  - Access to parallel database server
  - Caching for fast retrieval
  - Uniform interface for parallel data mining algorithms
- Parallel scalable data mining algorithms
  - Predictive modelling
  - Clustering and association rules
- Aim: Harness the power of high performance computing with a flexible toolbox

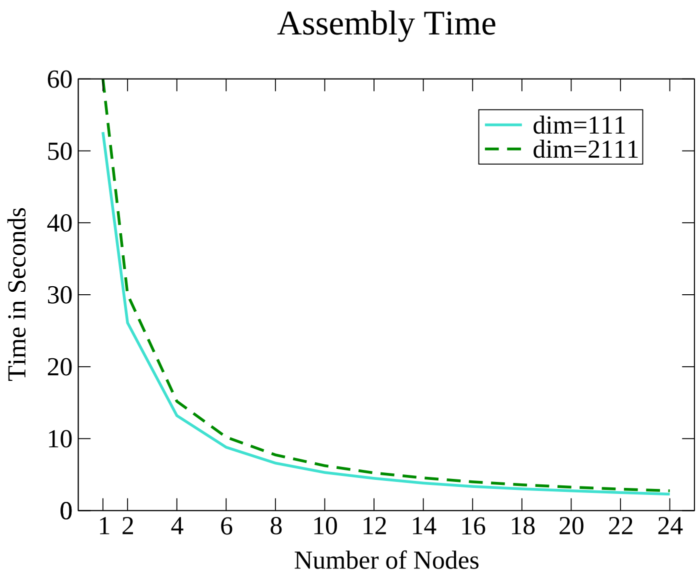
Parallelism in DMtools

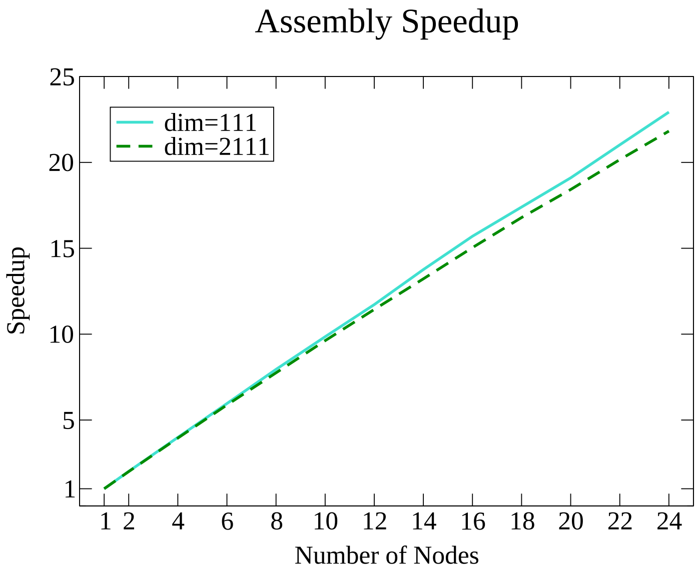
- Parallel database access
  - Many database servers are capable of running queries in parallel
  - *DMtools* start several queries over different tables in parallel, then collect results and process them
- Controlling and steering of parallel data mining applications

Parallel Predictive Modelling

- Our algorithms for predictive modelling are scalable with the size of data collections and number of processors
  - Read distributed data in parallel (only once) and build models locally on each processor
  - Combine (reduce) models to final model, then solve the (linear) system
  - Size of the model does not depend on the size of the data, only on the accuracy of the model

Example Timing Results





- Assembly of linear systems for additive models
- Platform: ANU Beowulf Linux cluster *Bunyip*

THE AUSTRALIAN NATIONAL UNIVERSITY

Outlook: Current and Future Work

- Integration of parallel data mining algorithms into *DMtools*
- Integration of statistical and graphical packages into *DMtools*
- Extension of predictive modelling
  - Sparse grids
  - Complex data types (sets, vectors, etc)

- Visit our web site at:

  http://csl.anu.edu.au/ml/dm/