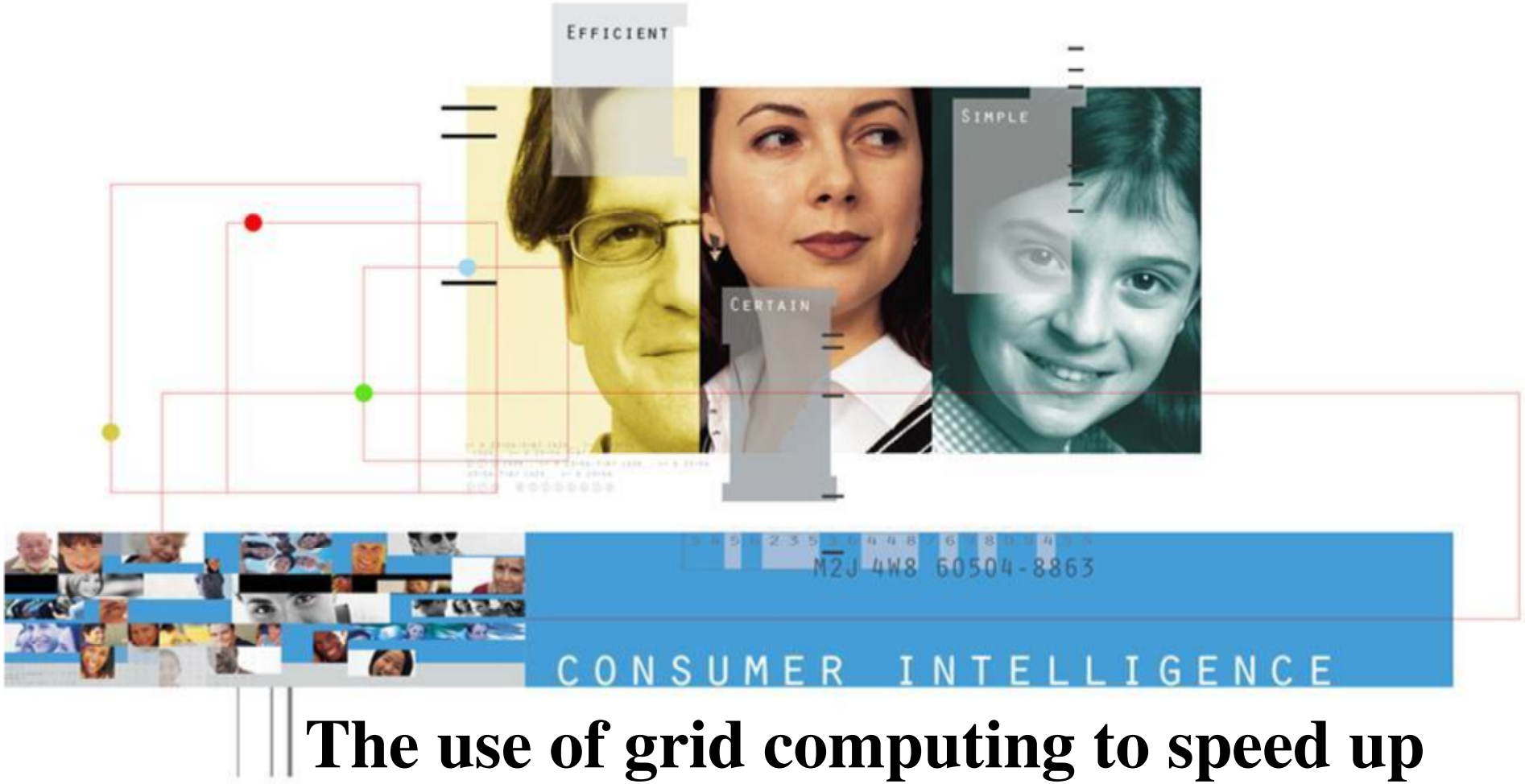
Generation5



The use of grid computing to speed up prediction

Alex Depoutovitch
29.10.2004

The use of grid computing to speed up prediction

Agenda

1. Problem definition
2. Available solutions
3. G5 MWM grid computing solution
4. Benchmarks of prediction with G5 MWM Grid
5. Upcoming developments

Problem

Size of data available for analysis growths exponentially with time!

- Data size increase in one year
  - **At least 2 times for 29% of responders**
  - **At least 3 times for 13% of responders**
  - **At least 4 times for 8% of responders**

  (according to Winter Corporation survey)

- For example
  - **Hudson Bay: 10 times in 5 years**
    from 243GB in 1999 to 2TB in 2004
  - **Walmart: 200 times in 9 years**
    from 500GB in 1990 to 100TB in 1999
  - **Generation5: 15 times in 4 years**
    from 300GB in 2000 to 4,5TB in 2004

The solutions

- **Improving efficiency of algorithms**
  Most of widely used algorithms are known for tens of years and it is seems extremely hard to improve their performance dramatically
- **Sampling**
  This approach leads to results accuracy lost
- **Faster processors**
  Speed of processors doubles in every two 2 years. Hard drive size doubles every year and size of data growths even faster
- **Using symmetric multi-processor computers**
  - Cost ineffective: 8-CPU box 6 is times more expensive than 8 1 CPU boxes
  - Limited usually by 32 processors
  - Requires scaleable parallel algorithms

The solutions (continued)

Grid Computing

- Benefits:
  This approach gives the unique advantages of unlimited computational resource increase for a very low cost

- Challenges:
  One must use specially designed algorithms that allow independent execution without much data exchange and synchronization overhead.

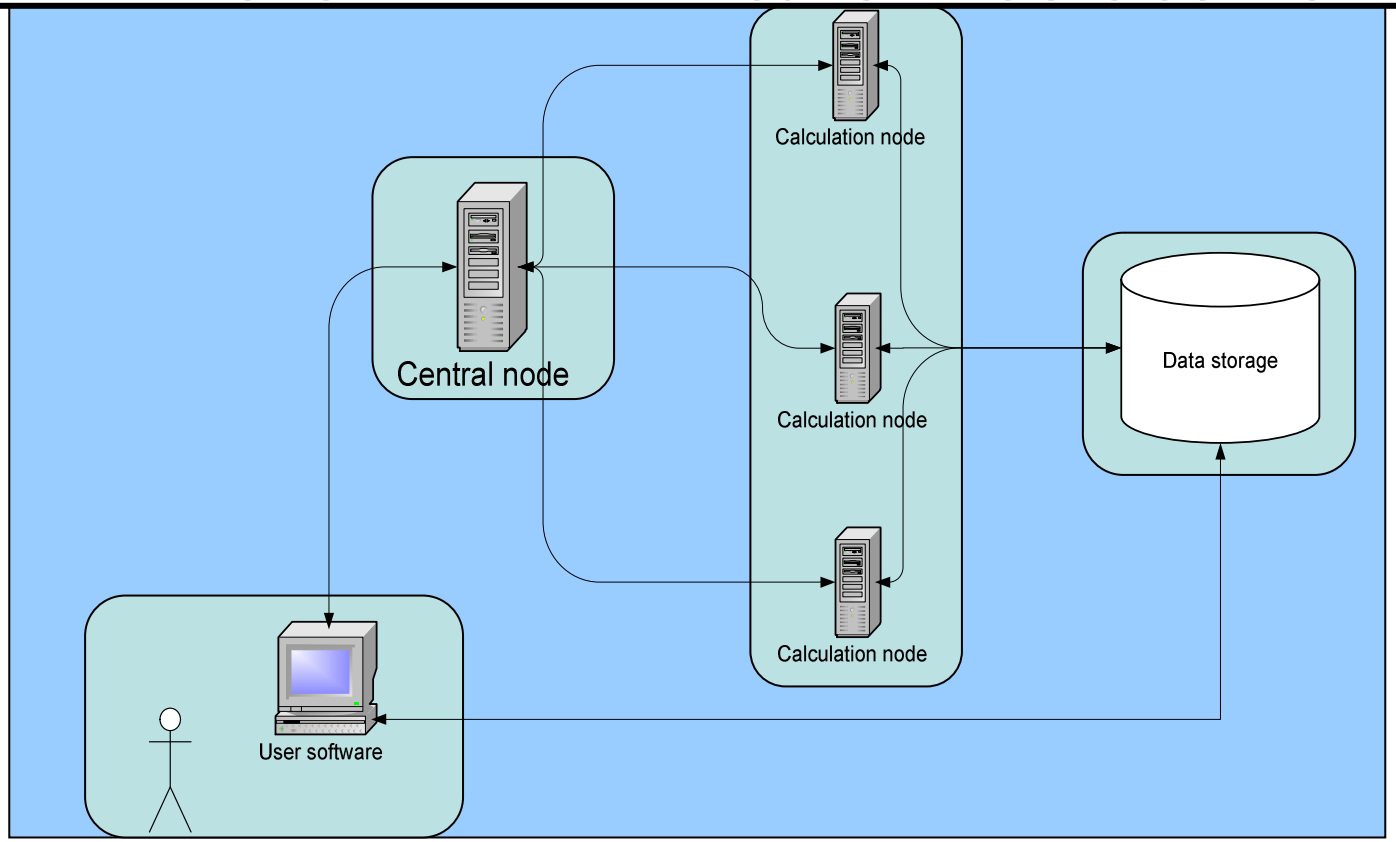
Support of grid computing in available statistical and data mining applications

- Commercial statistical and data mining software
  - SAS® — no support
  - Statistica ® — only validation
  - SPSS® — no support
  - KXEN® — no support
  - UNICA® — no support
- Emerging technologies
  - Linda (www.turboworx.com/) — General framework for grid computing
  - PaDDMAS (Rana et. al. 2000) — General framework and neural network algorithm only
  - D-DOAL (Pathasaraty & Subramonian 2001) — General framework and clustering algorithm only

G5 MWM GRID

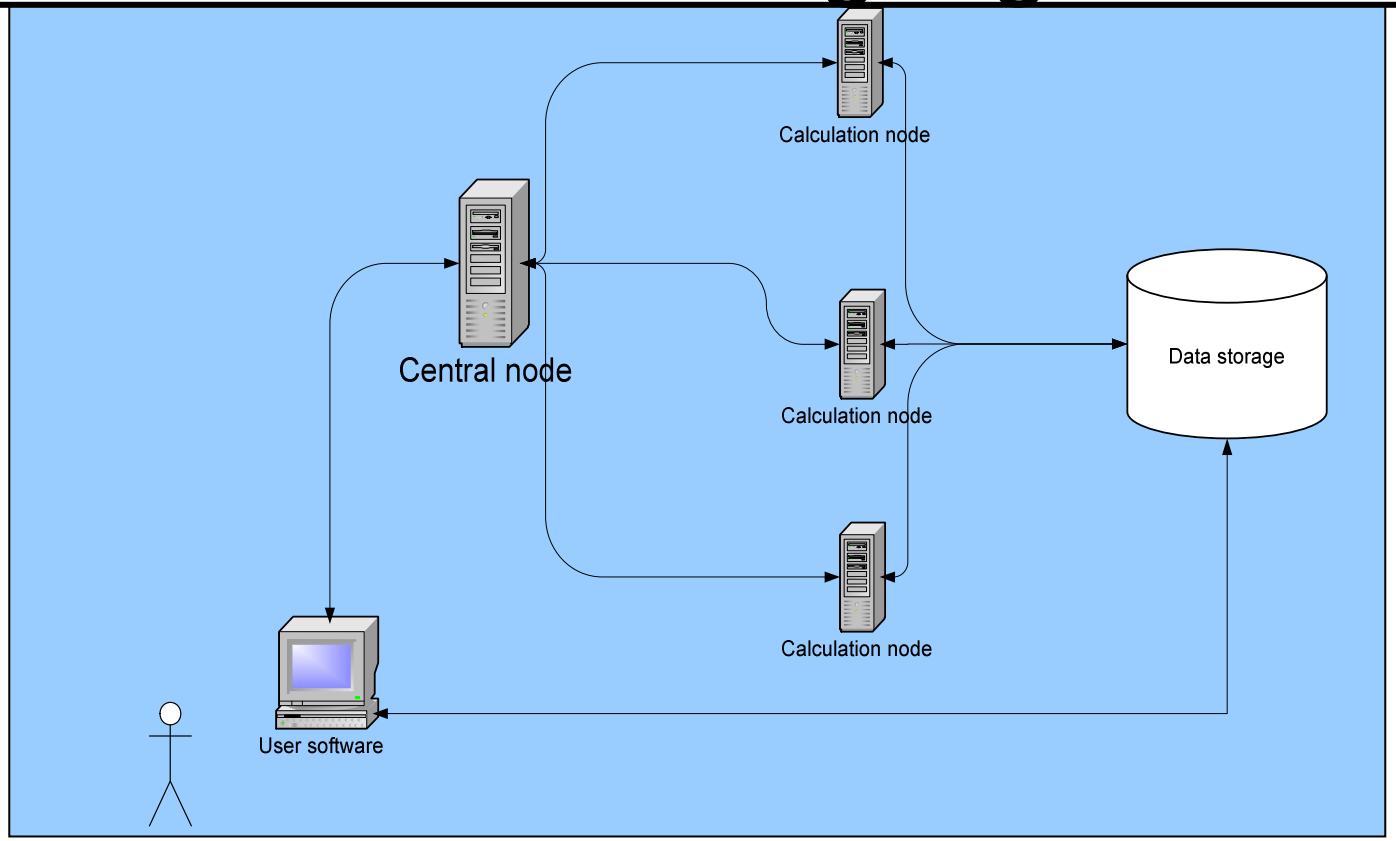
- Addresses the goal of speed increase with following benefits:
  - Achieves linear scalability of prediction algorithm with growth of number of computers in the grid
  - Combine set of heterogeneous computers including multiprocessor computers into one virtual supercomputer
  - It may be used with any statistical algorithm that allows parallelization of computations
  - Allows concurrent execution of several calculation tasks submitted by different users with dynamic load balancing between computers in the grid
  - Provides fault tolerance

G5 MWM architecture



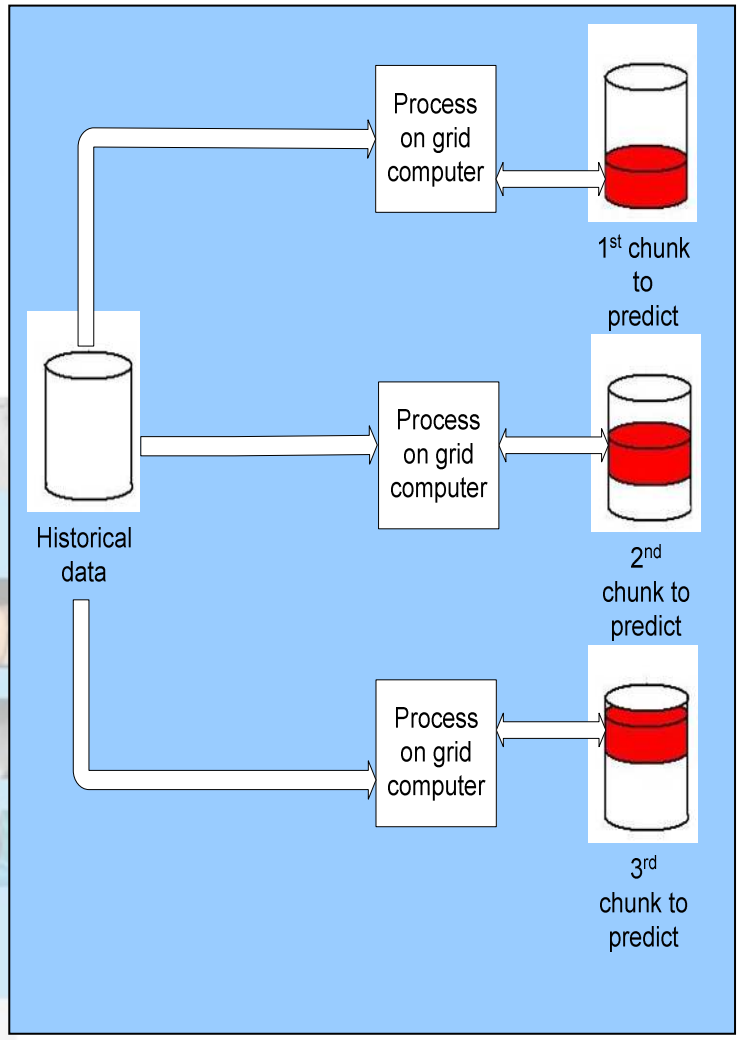
1. **Central node** - responsible for coordination of process running on computers in the grid.
2. **Calculation node** - single-processor computer or one processor in multi-processor computer and software running on it that executes tasks received from central node
3. **Data storage -** contains data that is available from central node and all calculation nodes
4. **Clients** - users of the system. Create jobs and submit them to central node

Load balancing algorithm



1. Central node starts itself, start calculation nodes and connects to them.
2. When prediction job is submitted by user, central node starts splitting it into tasks.
3. As soon as central node finds free calculation node it sends task for execution. When calculation node finishes task it will receive next one.

Prediction in G5 MWM grid



Prediction algorithm developed by G5 and based on Nearest neighbors approach with distance metrics that takes into account relevance of each input variables.

Benefits:

- Allows to predict each record independently from others.

Models for testing

Two prediction projects:

1. 40,000 records in training set
   40,000 records with missing values
   22 independent and 1 dependent variable

   Total number of values to predict: 40,000

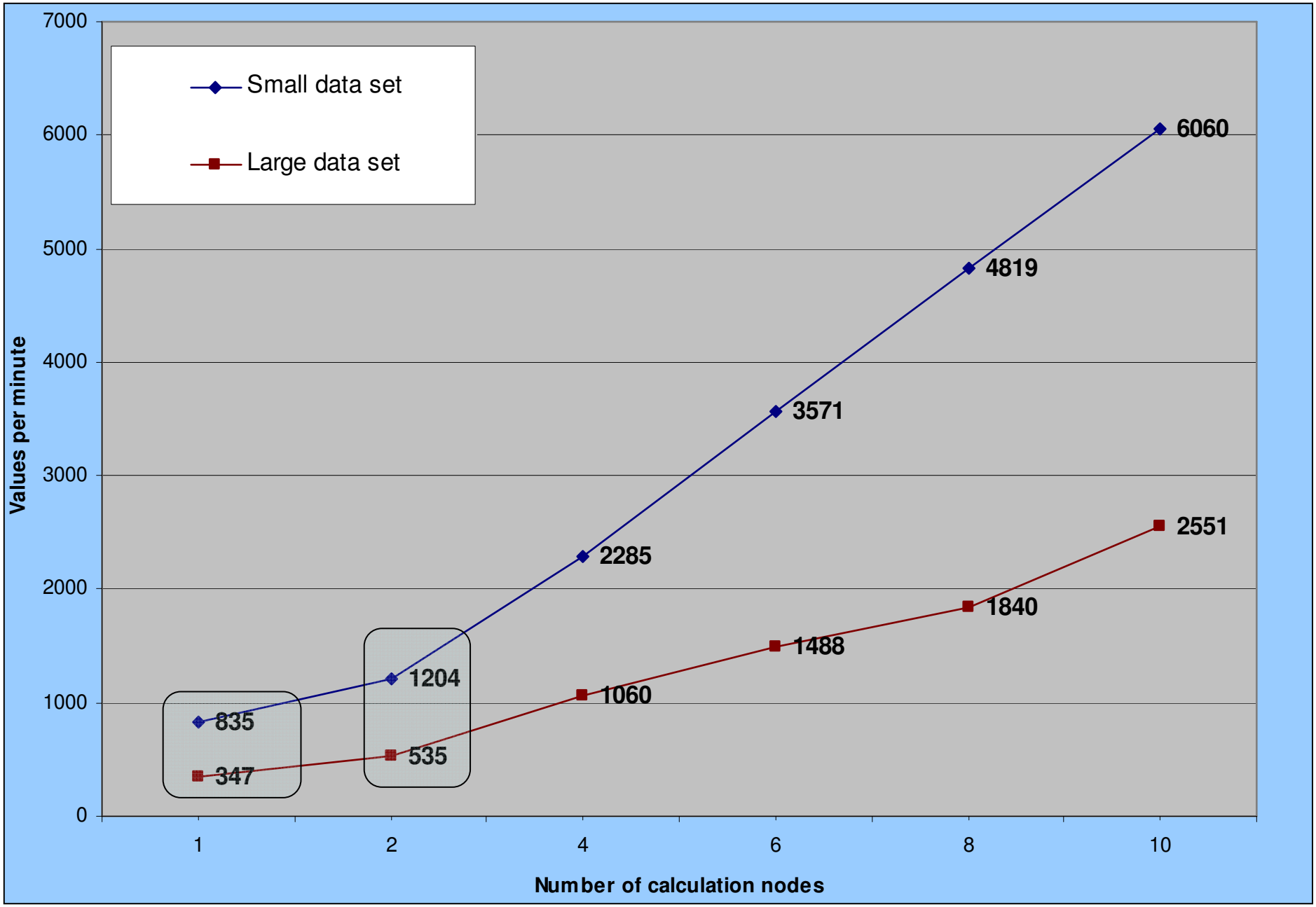
2. 250,000 records in training set
   250,000 records with missing values
   47 independent and 5 dependent variables

   Total number of values to predict: 1,25 millions

Test configuration

- **Central node**
  - Single processor 1.8GHz Pentium 4 running Windows XP
- **Calculation nodes**
  - From one to five
    Dual processor 2x 1GHz Pentium III running Windows 2000 Professional
- **Database server**
  - Was running on central node computer
- **Network**
  - Gigabit Ethernet

Execution results



Number of values predicted per minute for both data sets.

Resource utilization

10 calculation nodes:

- Data base server CPU time:
  ~3 min for small data set (30% of total execution time)
  ~25 min for large data set (4% of total execution time)

- Central node CPU time:
  less then 5 seconds

- Network utilization up to 4%

- Calculation node resource utilization: ~85%

We are far below bottleneck threshold!

Future developments

- Adding parallelization support to algorithms that solve other important data mining tasks such as:
  - Feature selection
  - Clustering
  - Association analysis
  - Validation
- To implement shared memory for easier parallelization of statistical methods

Conclusion

G5 MWM Grid is statistical package that support grid computations and provides following benefits:

- **Using ordinary PCs you can achieve performance of supercomputer for cost of several low budget computers.**
- **Calculation power is proportional to number of computers in grid.**
- **Simple configuration and user interface. User do not need to prepare data or write his algorithms to utilize distributed computations.**
- **Multi-user access**

Questions