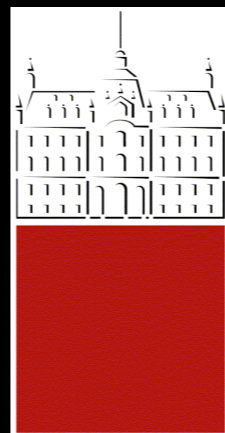


Challenges for data-intensive applications in heterogeneous environments

Uroš Čibej,
University of Ljubljana, Slovenia

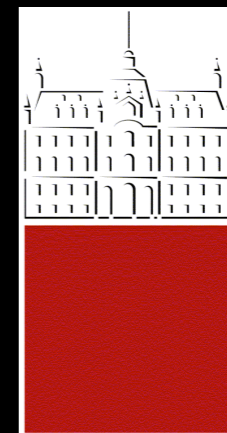
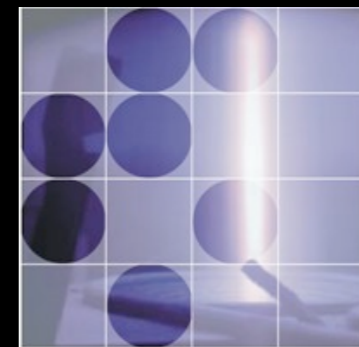


Overview

- HPC in Slovenia
- Data-intensive applications
- Infrastructural problems
- Optimization problems

HPC in Slovenia

- Institute Josef Stefan
- Turboinstitut
- University of Ljubljana



Institute JS

- Molecular dynamics
- Medical applications
- Distributed protocols and communication topologies
- Grid computing

Turboinstitut

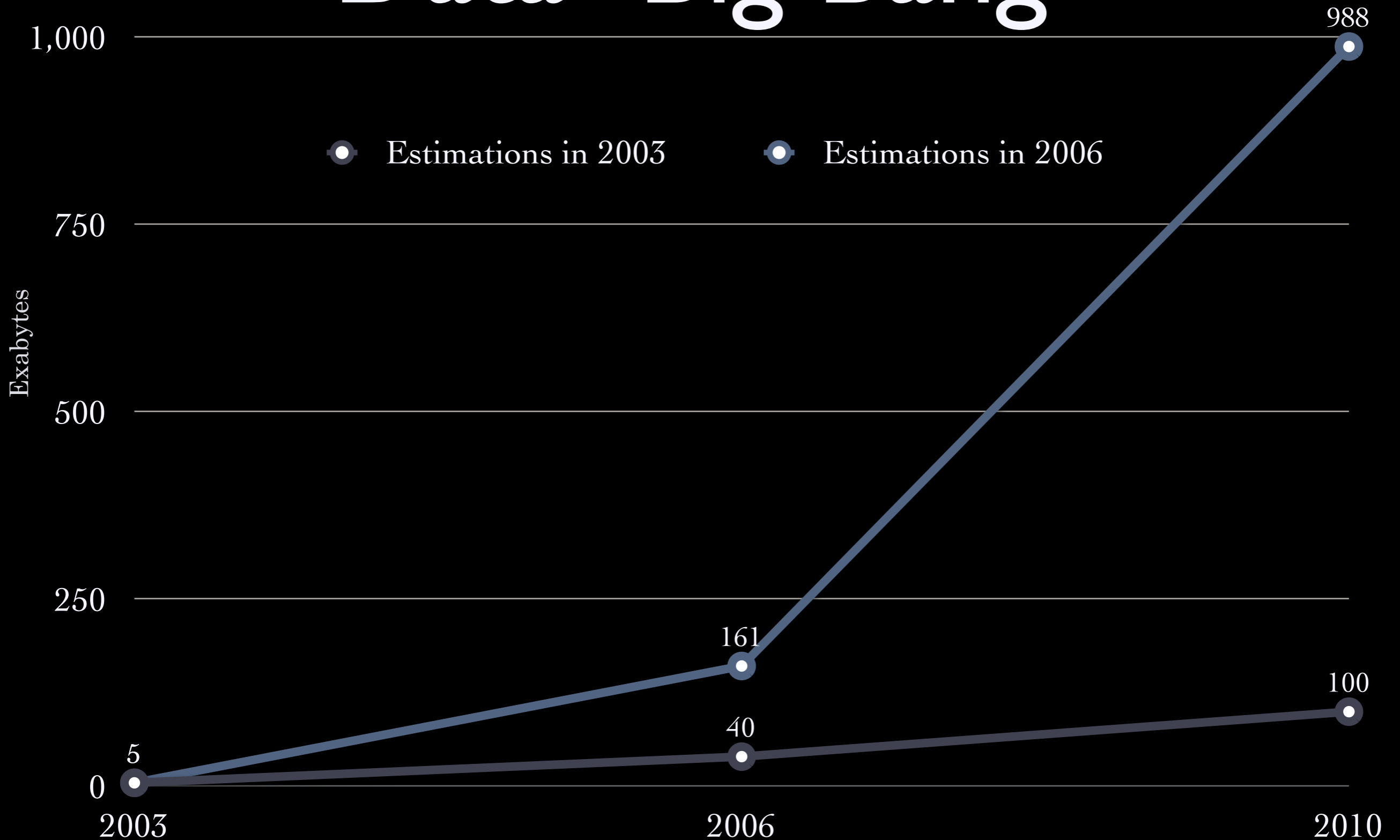
- Established Ljubljana Supercomputing Center (LSC ADRIA) - 4000+ processors
- Fluid flow simulations
- Other types of applications (e.g. pharmaceutical, medical, aerospace)

Data-intensive computation

- Efforts in HPC were/are focused mostly on computationally intensive tasks
- Time to change the focus to data-intensive tasks
- Motivation : amounts of data are soaring

Data “Big Bang”

Data “Big Bang”



Why the “Big Bang”?

- Increased possibilities of capturing data - sensors, cameras, instruments...
- Increased data generation - monitoring, archiving, backups, metadata, communications

The reality

- Vast majority of this data is hardly ever used
 - Hard to access/process
 - Useless data
- A need for an infrastructure to deal with these issues

The challenges

- Infrastructural challenges
- Optimisation challenges

Infrastructure

- Security
- Reliability
- Adding structure to data
- Data discovery

Infrastructure

- Security
- Reliability
- Adding structure to data
- Data discovery

Example of a data discovery mechanism

- Based on Ant-Colony Optimisation
- Each user/interest group provides examples of data they are interested in
- Ants crawl around, gathering interesting metadata in piles

Optimisation

- Data placement (replication) as the most important optimisation problem
- Equivalent to scheduling in computationally intensive tasks
- Scheduling and replication are complementary

Types of applications

- Clients
- Independent tasks
- Workflows

Types of applications

- Clients
- Independent tasks
- Workflows

Our approach

- Theoretical modelling of systems
- Design of static, centralized algorithms
- Transform static algorithms to dynamic and distributed

Theoretical models

- Use the knowledge from location theory
- Center and median location problems are related to the data placement problems in distributed systems
- Centralized algorithms need to be adapted for the distributed environment

Data-intensive workflows

- Workflows are gaining importance
- Workflows bring another level of complexity to optimisation
- Need to develop entirely new models

Summary

- Let's make data a first class citizen in our systems
- Still a lot of work in optimisation of data access
- New types of applications (e.g. workflows) need new approaches

Thank you