中国科学院大学

Practical Parallel Computing on Supercomputers

Qinghai Miao
miaoqh@ucas.ac.cn

Introduction

- For practical engineering:
  - The scale of problems are usually too large to run a single computer/workstation.
  - The ability (Can we do?) to run the simulation is much more important than the performance (How well can we do?)

Introduction

- The performance of all numerical simulation method, especially in Three-Dimensional, is required to improve.

The Great Pyramid

- The Pharaoh Khufu pyramid, located in Giza, Egypt, is known as the biggest one.
- The height of the Pyramid is 147m and the width of the foundation is 230m.
- The layer of the masonry is 203 and the total number of blocks is estimated about 2.3 millions.

Content

- **Method and Technique**
- Profiling of Serial Program
- Parallelization
- Verification and Testing
- Run on Supercomputers
- Visualization

Method and Technique

- Discontinuous Deformation Analysis (DDA)
  - Method in rock mechanics and rock engineering.
  - Principles….

- In 2015, we proposed ***"Double 100"*** project that 3D-DDA :
  - 100 times speedup
  - 100 million blocks

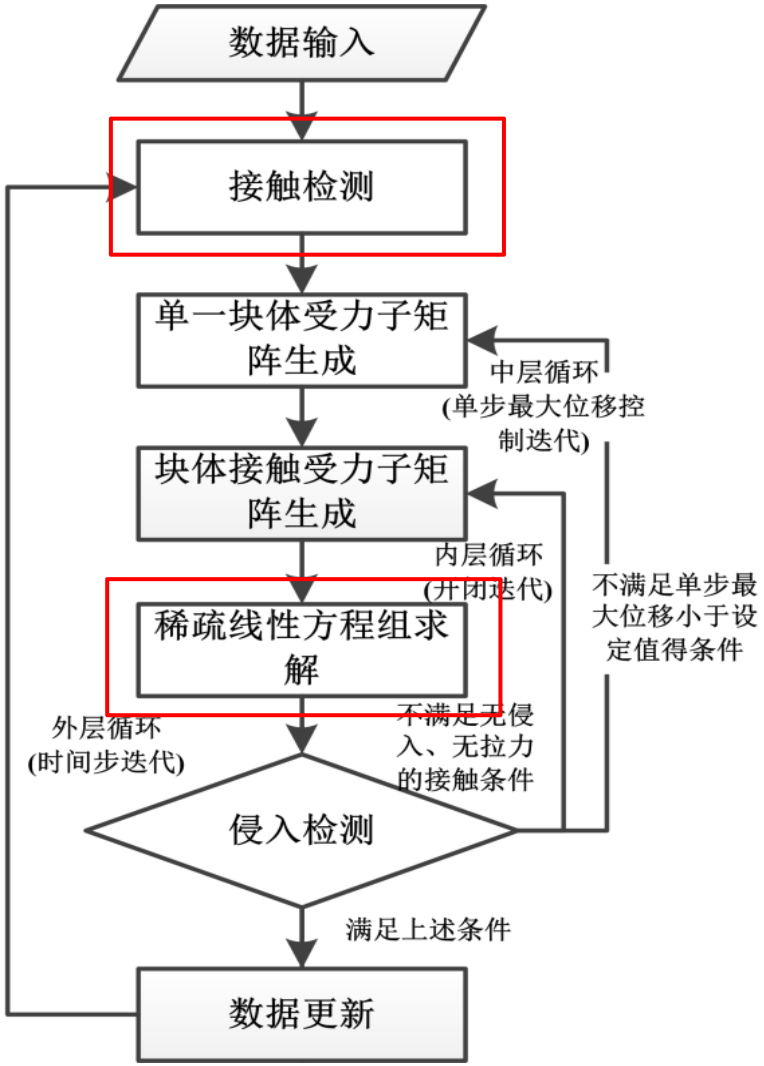
Related works

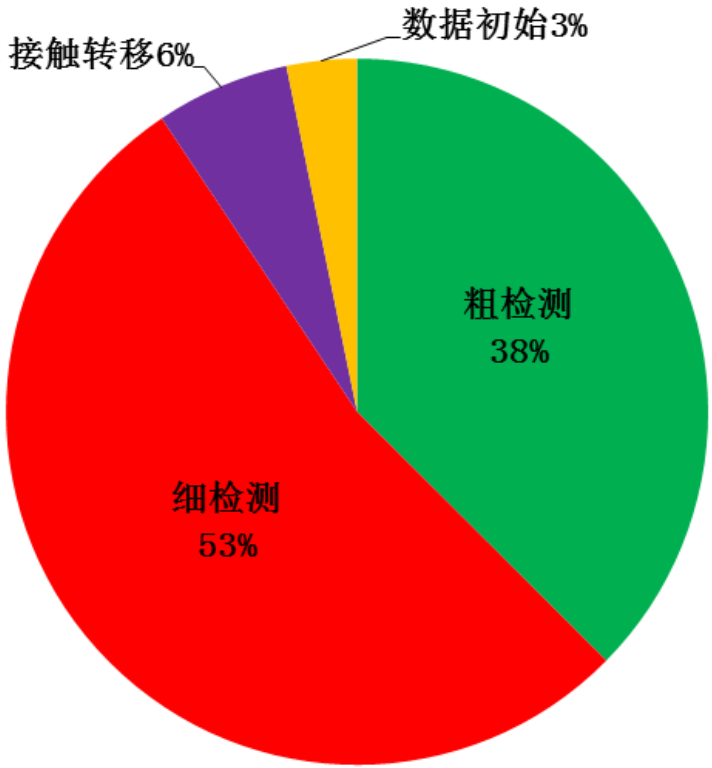
- GPU has become a popular parallel architecture and has been used to accelerate DDA computing.
  - Fu(2015), Song(2017), Xiao(2017)
  - 5~20 times of speedup;
  - Partial parallelization;
- But the most powerful platform is supercomputer like Tianhe-2:
  - Much larger scale;
  - Much better performance;
  - Complete parallelization;

Content

- Method and Technique
- **Profiling of Serial Program**
- Parallelization
- Verification and Testing
- Run on Supercomputers
- Visualization

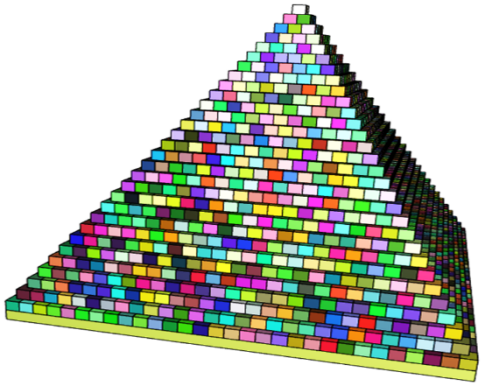
Performance of serial DDA

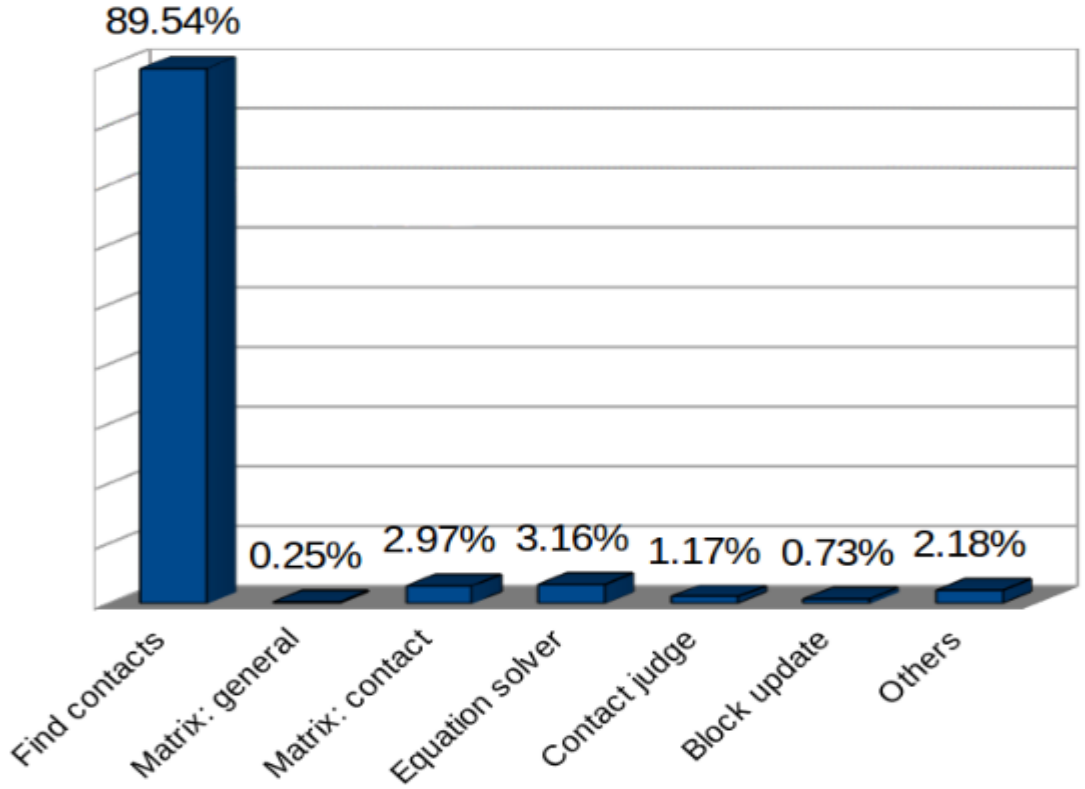




Performance of serial DDA

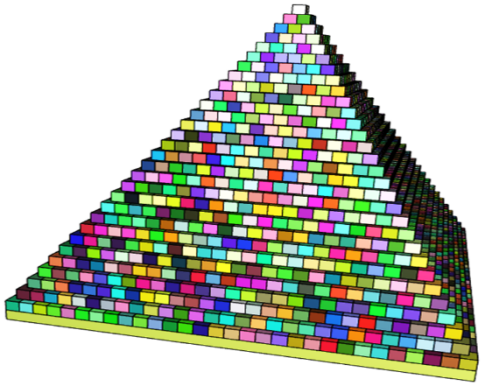
- Pyramid example, profiling with TAU.

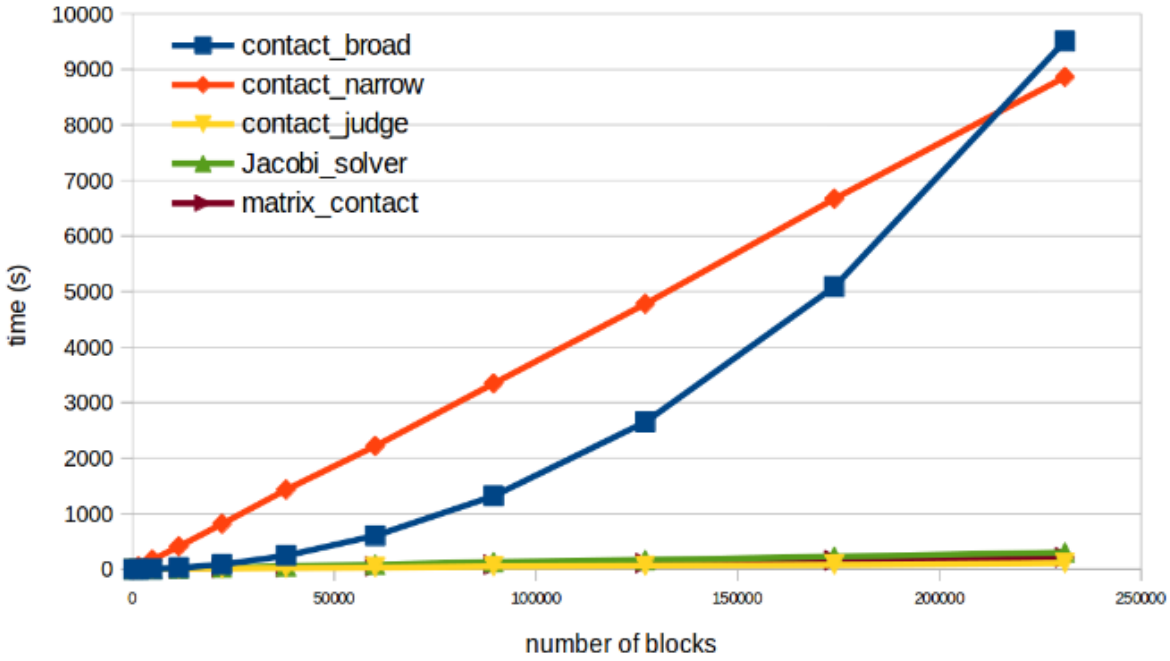




Performance of serial DDA

- Pyramid example, profiling with TAU.





Conclusion

- Contact accounts for nearly 90% of total computing time;
- Equation solver is the second time consuming part.

- Solution:
  - Parallelize the DDA process, especially contact and solver.

Content

- Method and Technique
- Profiling of Serial Program
- **Parallelization**
- Verification and Testing
- Run on Supercomputers
- Visualization

Methods

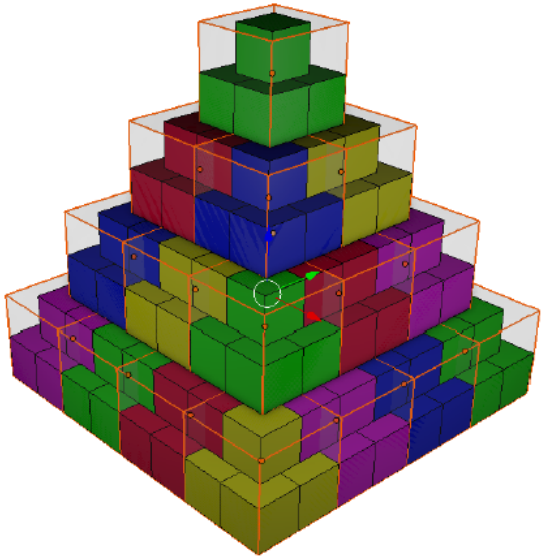
- Domain-Decomposition:
  - To take advantages of supercomputers, the parallel DDA takes the strategy of domain decomposition based on MPI.
  - By decomposing the original domain with N blocks into m subdomains, all the subdomains can be computed on m separated process simultaneously.

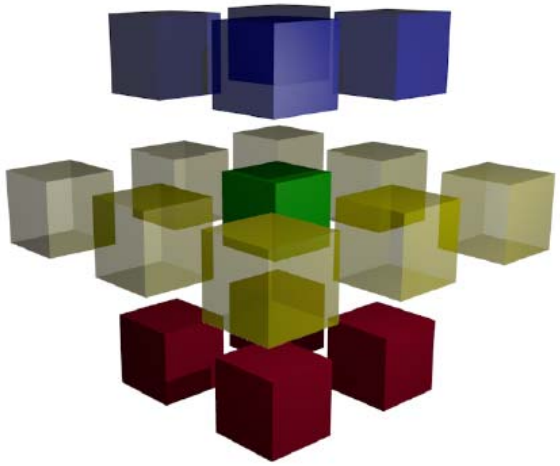
Methods

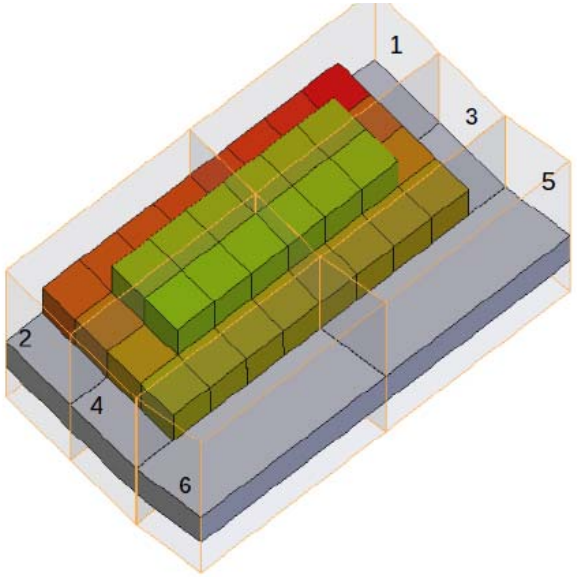
- Domain-Decomposition:
  - For algorithms with linear time complexity O(N), the computing time can be theoretically reduced to 1/m plus the time for communications between subdomains.
  - For algorithms with exponential time complexity $O(N^2)$, the complexity becomes $O((N/m)^2)$, which can dramatically reduce the computing time.

Methods

- Domain-Decomposition:





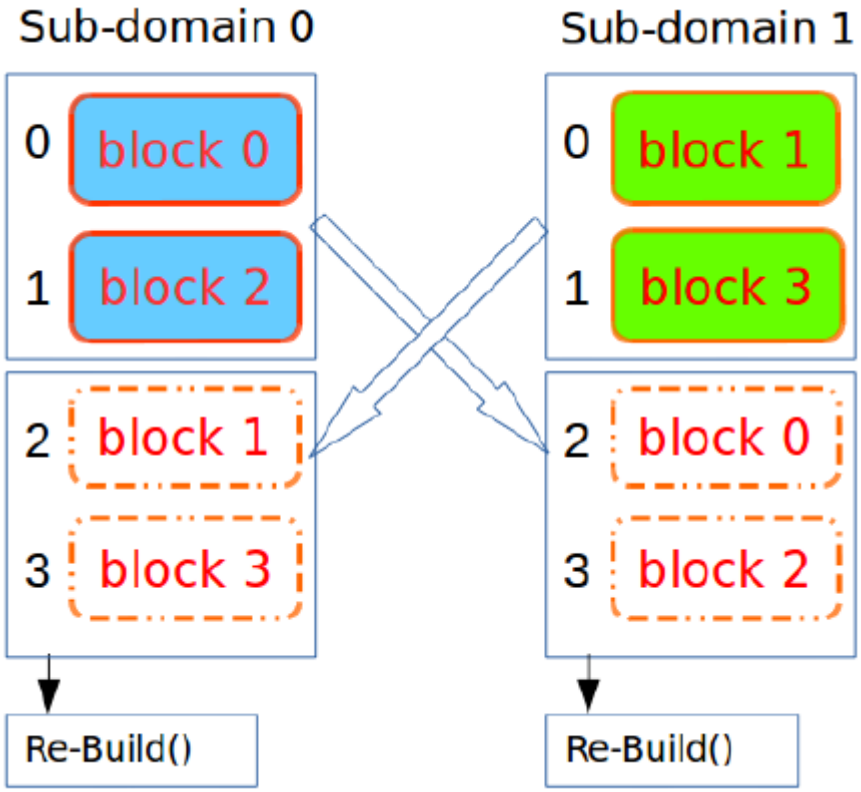


Methods

- Transfer blocks between subdomains
  - As DDA allows large displacement, blocks may move from one subdomain to its adjacent subdomains.
  - When transferred to a new subdomain, a block need to be coherently reconstructed as a complete local block to keep the simulation stable.
  - A ***Send-Receive-Rebuild*** procedure, based on Object-Oriented design, for transfer blocks with their geometry, physics and contact attributes.

Methods

- Transfer blocks between subdomains



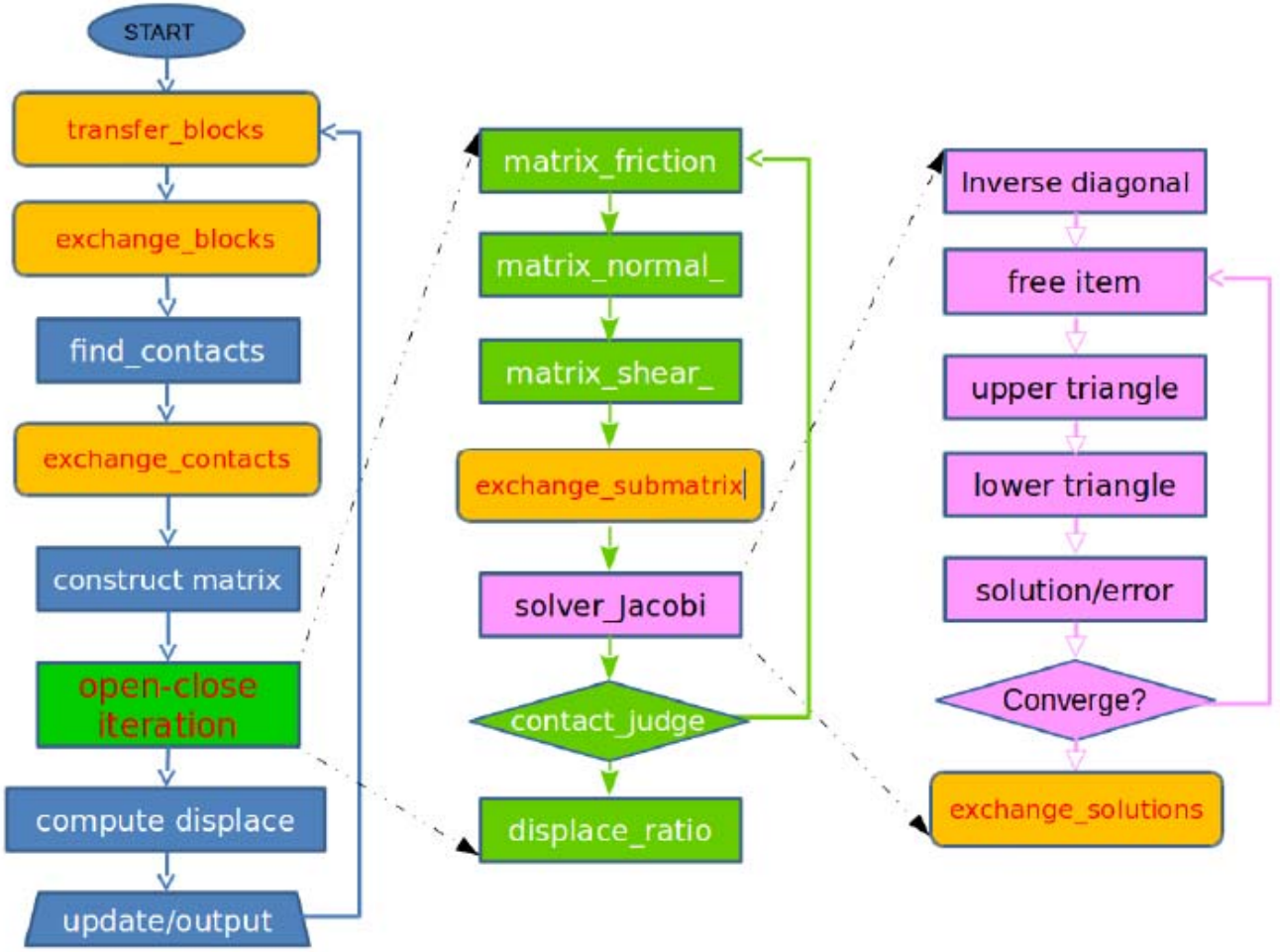
Methods

- Contact Detection
  - Multi-domains make it more complex for contact in DDA.
  - Contact detection is required both within local subdomain and inter-subdomains.
  - The concept of ***Ghost Blocks*** was adopted and ***Three Rules*** were proposed for contact detection based on local index and global index.

Methods

- Exchange of submatrix
  - The submatrices of confirmed contacts on remote subdomains are required by the local solver.
  - The remote contact submatrices are stored and recorded by the Ghost Block with the same local global index as its original block.
  - Before solving the equations, all the submatrices need to be sent back to local subdomain by MPI.

Methods


START
transfer_blocks
exchange_blocks
find_contacts
exchange_contacts
construct matrix
open-close iteration
compute displace
update/output
matrix_friction
matrix_normal_
matrix_shear_
exchange_submatrix
solver_Jacobi
contact_judge
displace_ratio
Inverse diagonal
free item
upper triangle
lower triangle
solution/error
Converge?
exchange_solutions

MPI functions

- Virtual Topology
  - MPI_Dist_graph_create_adjacent
- Configuration
  - MPI_Bcast
- Exchange data
  - MPI_Isend &MPI_Irecv
  - MPI_Waitall
- Synchronization
  - MPI_barrier

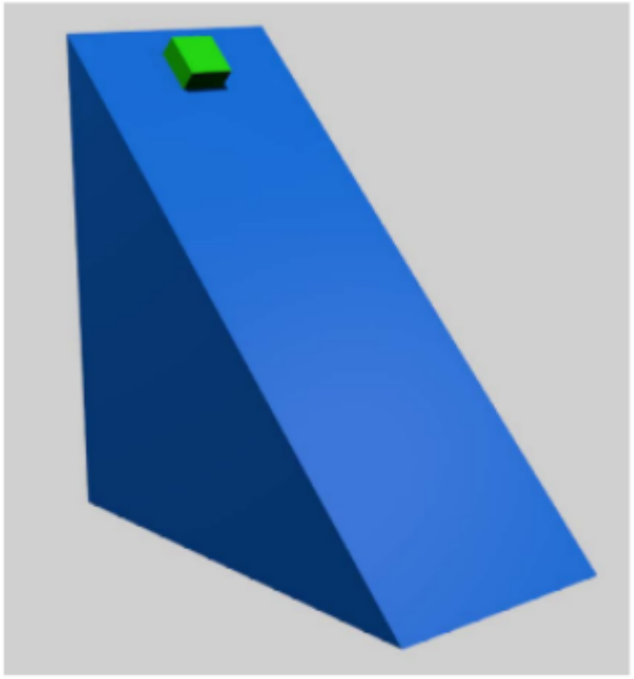
Conclusion

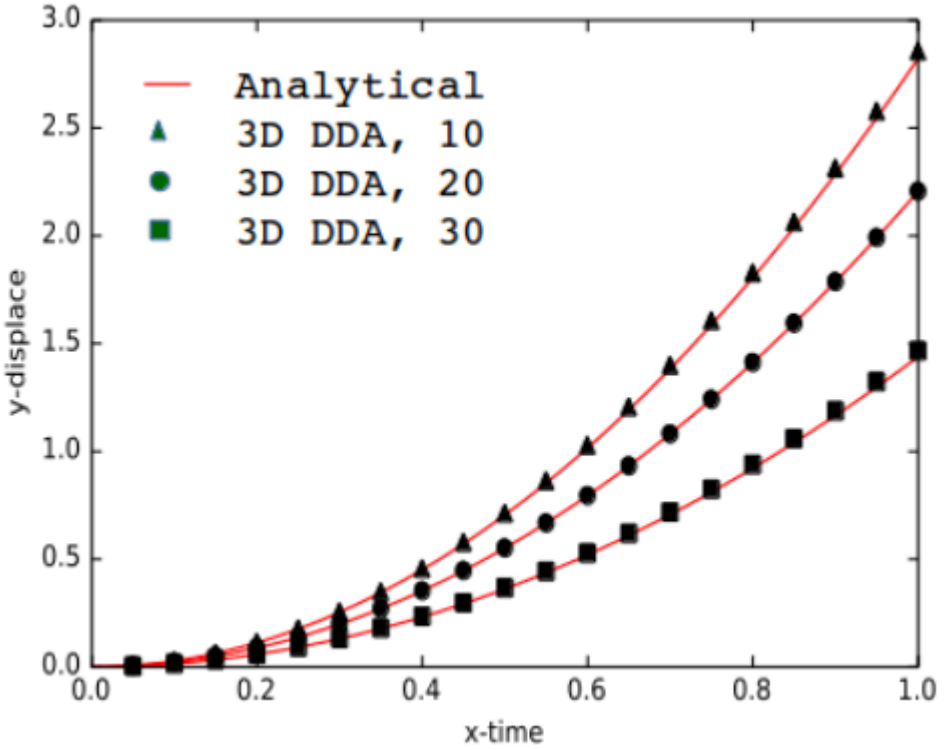
- Domain-Decomposition
  - Virtual Graph
- Contact between sub-domain
  - Ghost blocks
- Blocks transfer between sub-domain
  - Send-recv-rebuild
- Parallel solver
  - SOR → parallel Jacobi

Content

- Method and Technique
- Profiling of Serial Program
- Parallelization
- **Verification and Testing**
- Run on Supercomputers
- Visualization

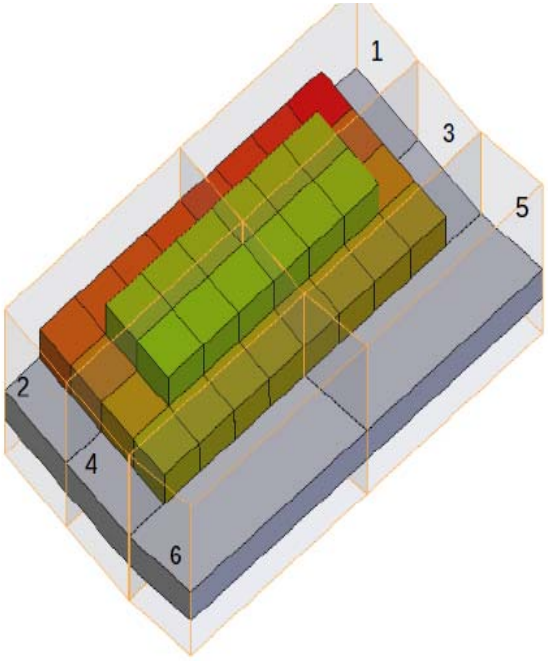
Verification

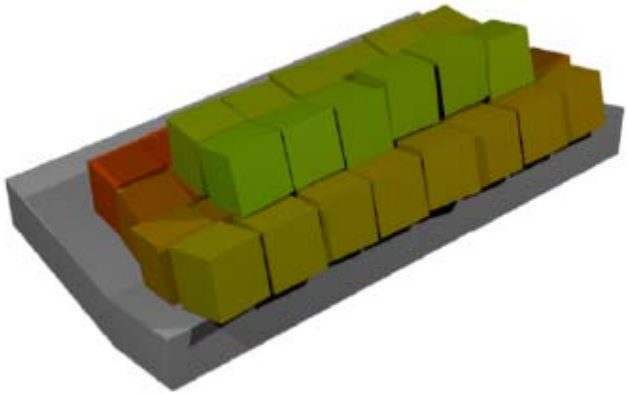




Test on workstation

- 44 blocks on 6 processors of Desktop Workstation – 2.6x speedup





Test on mini-cluster

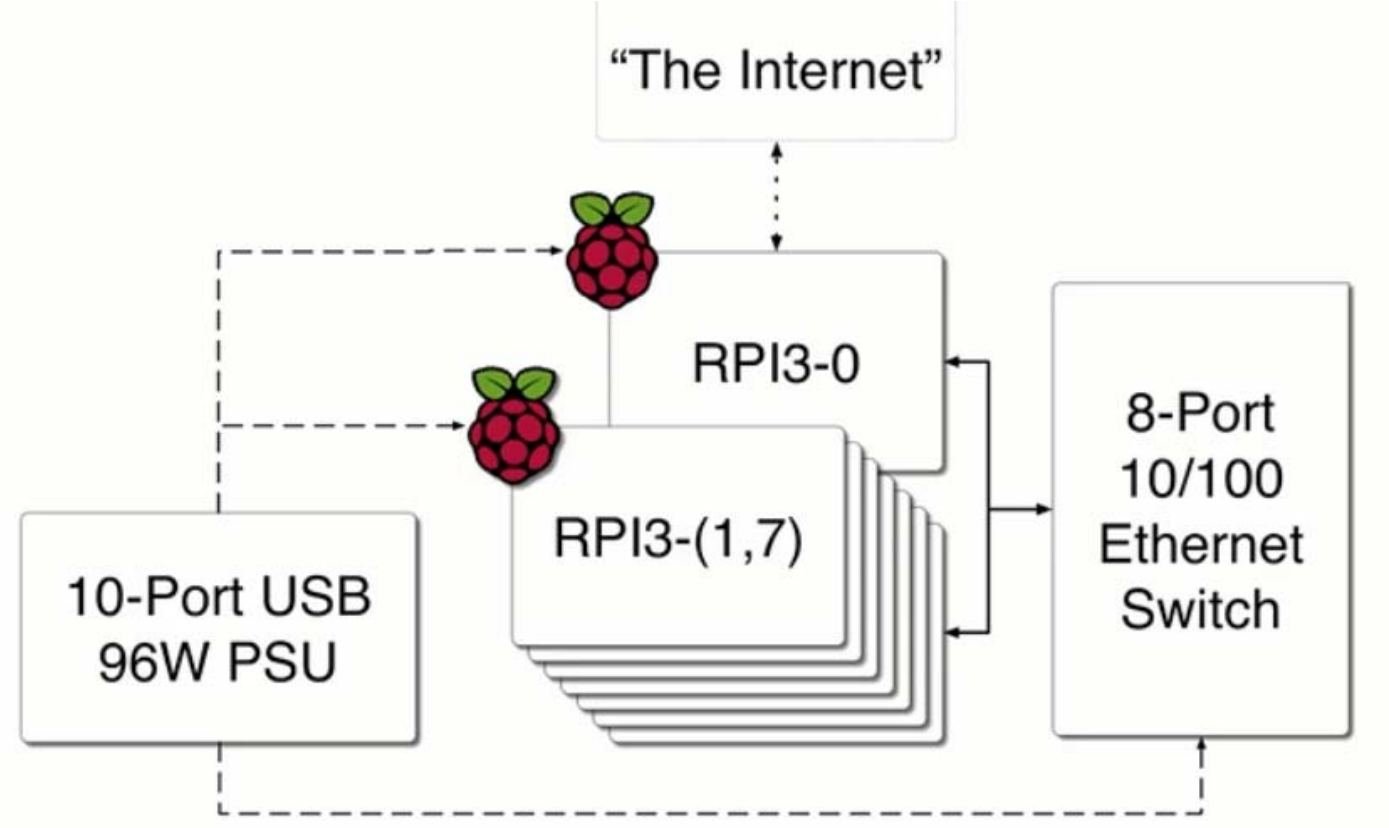
- Mini-cluster using Raspberry Pi

Raspberry Pi 3 Model B+

- 1.4GHz 64-bit quad-core processor
- dual-band wireless LAN,
- Bluetooth 4.2/BLE,
- faster Ethernet,
- Power-over-Ethernet support (with separate PoE HAT)

Test on mini-cluster

- Mini-cluster using Raspberry Pi



Test on mini-cluster

- Mini-cluster using Raspberry Pi

Content

- Method and Technique
- Profiling of Serial Program
- Parallelization
- Verification and Testing
- **Run on Supercomputers**
- Visualization

国家超级计算广州中心
NATIONAL SUPERCOMPUTER CENTER IN GUANGZHOU

- 天河二号
  - 拥有约17920个计算节点
  - 每节点配备两颗Xeon E5系列12核心的中央处理器、三个Xeon Phi 57核心的协处理
  - 总内存容量约1.4PB，全局存储总容量约12.4PB

- http://www.nscc-gz.cn/

用户申请

- 试用（免费）
  - 限时限量
- 正式使用（签订合同）
  - 按机时收费，报价单独咨询
- 机时的计算以节点为基本单元
  - 比如使用2结点共12核运行1小时的程序
  - 机时为2*24*1=48（核*小时）
  - 而不是2*6*1=12 （核*小时）

登录

- 使用专用VPN通过 ssh 登录

- Step1:
  - 从管理员处获取认证文件。
- Step2：
  - 使用终端工具连接，通过使用系统管理员提供的Private Key文件认证登录。

文件传输

- 使用sftp
  - 可使用客户端例如
  - Xmanager、FileZilla、WinScp、FlashFTP
- 天河2文件系统被分为两个区
  - /HOME用于存储代码和程序编译
  - /WORK用于数据存放和运行作业
- 大量文件需要下载时，使用tar命令先行打包.

环境配置

- 天河2号使用module：
  - 通过配置modulefile支持环境变量的动态修改，
  - 能够控制软件不同版本对环境变量的依赖关系。
- module avail：
  - 查看可用的模块的列表
- module load [modulesfile]：
  - 能够加载需要使用的modulefiles
  - module load intel-compilers/13.0.0
  - module load OpenFoam/2.2.2

编译

- 天河二号系统已配置GNU和Intel编译器.
- 支持OpenMP和MPI两种并行编程模式:
  - OpenMP仅能在一个计算结点内并行;
  - MPI 可在一个或者若干个结点上并行。

编译(2)

- 编译器：
  - Intel
    - 已配置3个版本的Intel编译器
    - intel 11, intel 13, intel 14
    - which icc　　icc -V
  - GNU
    - 默认版本是4.4.6

编译(3)

- MPI编译环境：
  - 天河二号采用了自主互连的高速网络，因此底层MPI为自主实现，基于Intel编译器和GNU编译器进行编译。
  - 基于Intel编译器的mpi版本安装目录在/usr/local/mpi3下，为自主实现的mpi版本，默认版本基于intel14，静态库。
  - 用户也可使用Mpich或openMPI，但性能会低。

提交作业(1)

- 查看节点状态
  - 整体资源使用情况
  - yhinfo 或 yhi
- 查看作业状态
  - 所提交作业运行情况
  - yhq 或 yhq -a

提交作业(2)

- 交互式作业提交方式
  - yhrun
  - yhrun –n 48 –N 2 –p work ./program
  - yhcancel jobid

提交作业(3)

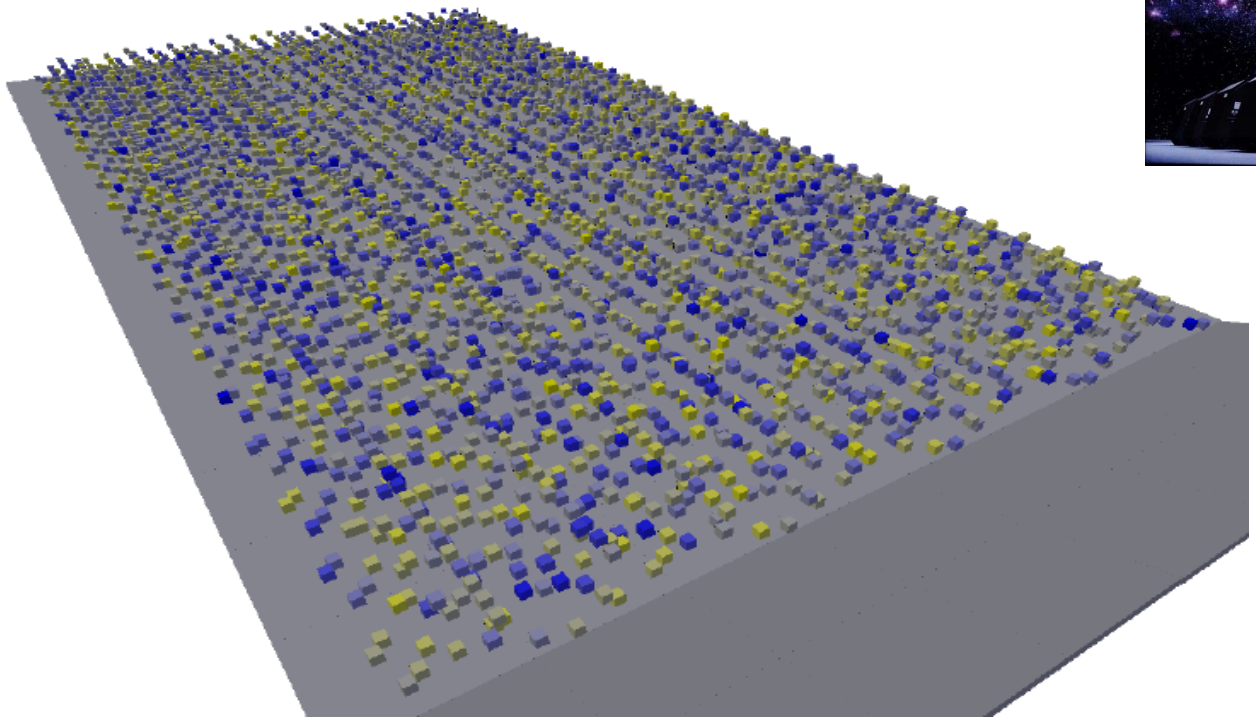
- 批处理作业提交方式
  - 指用户编写作业脚本，指定资源需求约束，提交后台执行作业：myjob.sh

```
#！/bin/bash
yhrun –n 48 –N 2 –p work  ./program
```

  - yhbatch -N 4 -p work ./myjob.sh
  - 计算开始后，工作目录中会生成以slurm开头的.out文件为输出文件

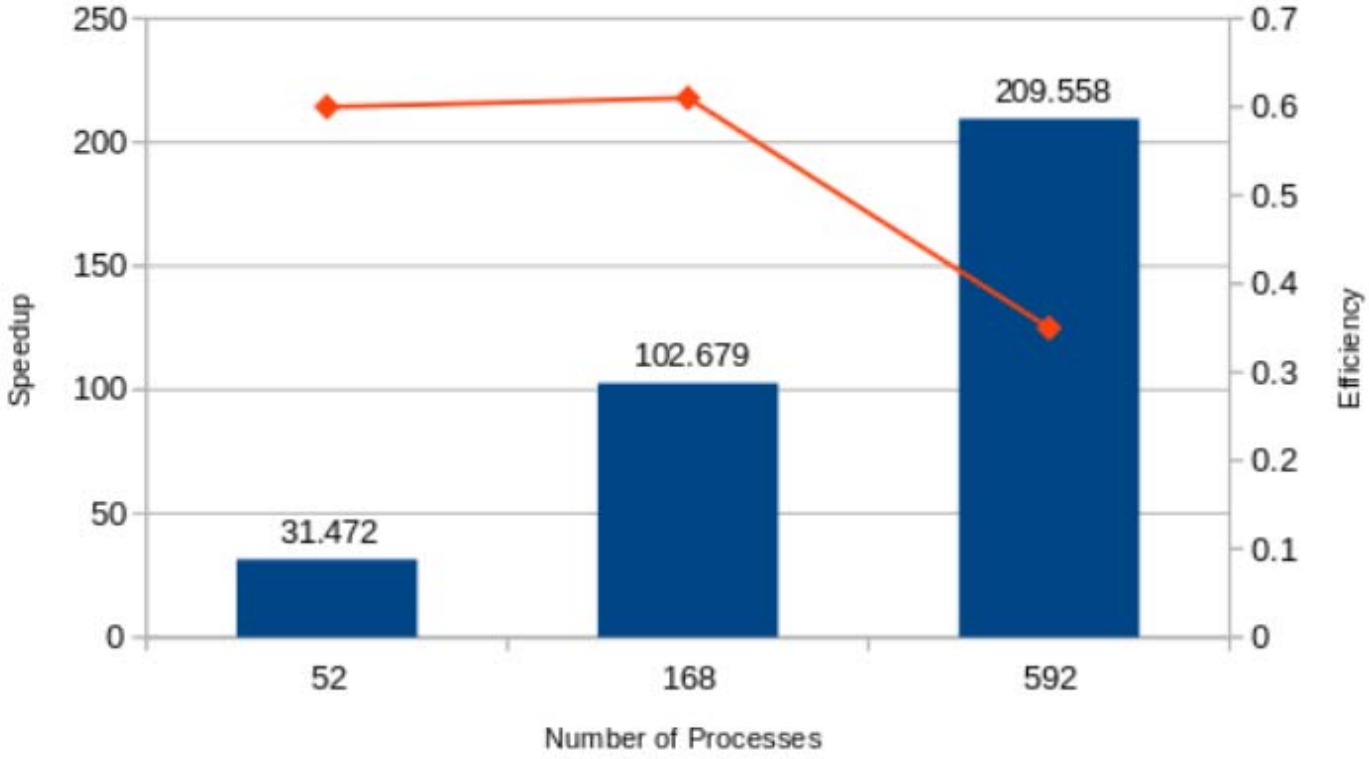
Test Examples

- 88,672 blocks of slope on Tianhe-2
- 592 processors – 209x speedup



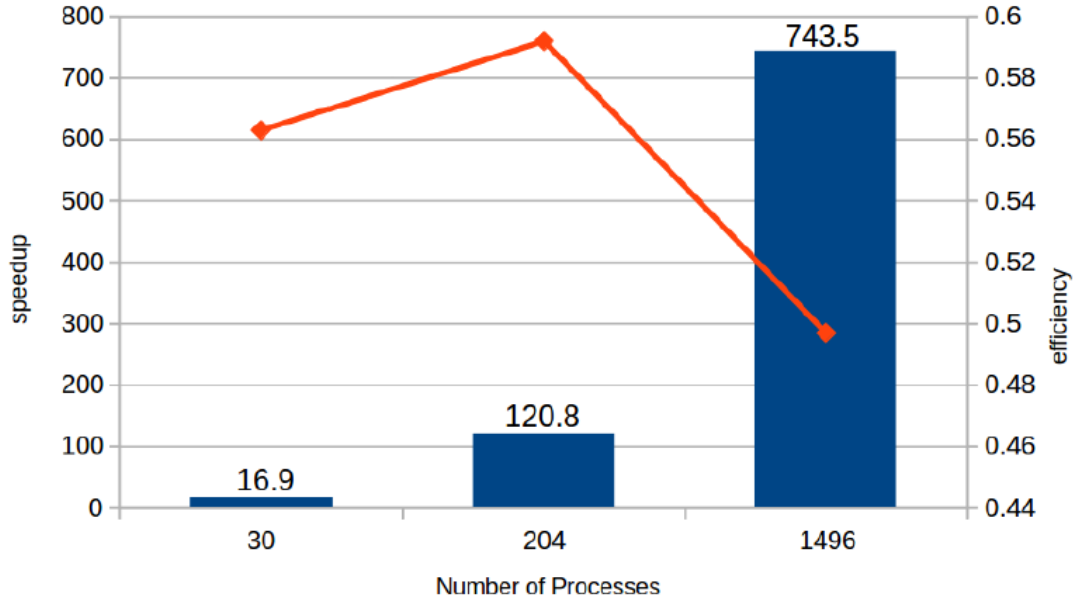
Test Examples

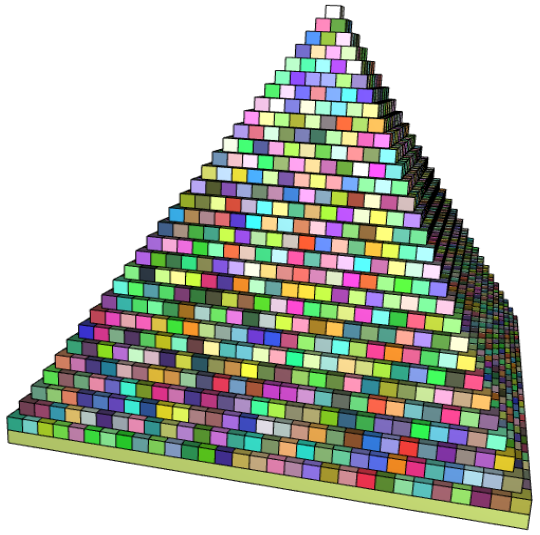
- 88,672 blocks of slope on Tianhe-2



Test Examples

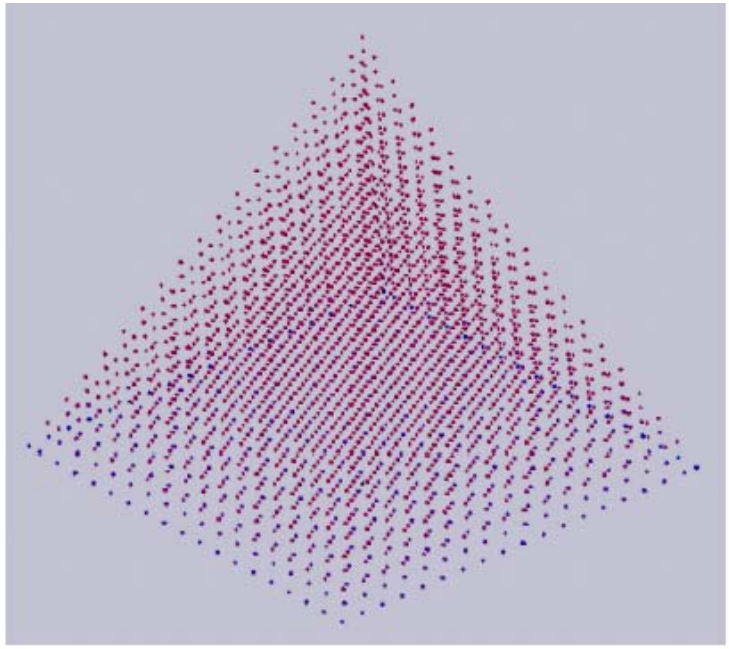
- 89,456 blocks of 64 layers Pyramid masonry structure on Tianhe-2 supercomputer – 743x speedup





Test Examples

- 2,304,776 blocks of full size Pyramid on Tianhe-2
- with 2,470 processors;
- 573 seconds (less than 10 minutes) for 10000 time steps.



Conclusions

- A parallel 3D DDA method using the strategy of domain decomposition is proposed and verified.
- The parallelization makes DDA capable of analysis large scale problems in three-dimensions.
- Test results have shown significant improvement in performance through tests on supercomputers.

Future works

- Better parallel contact algorithm based on E(A,B) theory;
- More efficient equation solver (Preconditioned Conjugate Gradient);
- More efficient visualization method for large scale problem;

Content

- Method and Technique
- Profiling of Serial Program
- Parallelization
- Verification and Testing
- Run on Supercomputers
- **Visualization**

Visualization

ParaView

- scales to allow visualization and analysis of even the largest scientific results.