Towards Efficient Two-Level Preconditioned Conjugate Gradient on the GPU

International Conference on Preconditioning Techniques for Scientific and Industrial Applications

May 16-18, 2011, Bordeaux, France

Kees Vuik, Rohit Gupta

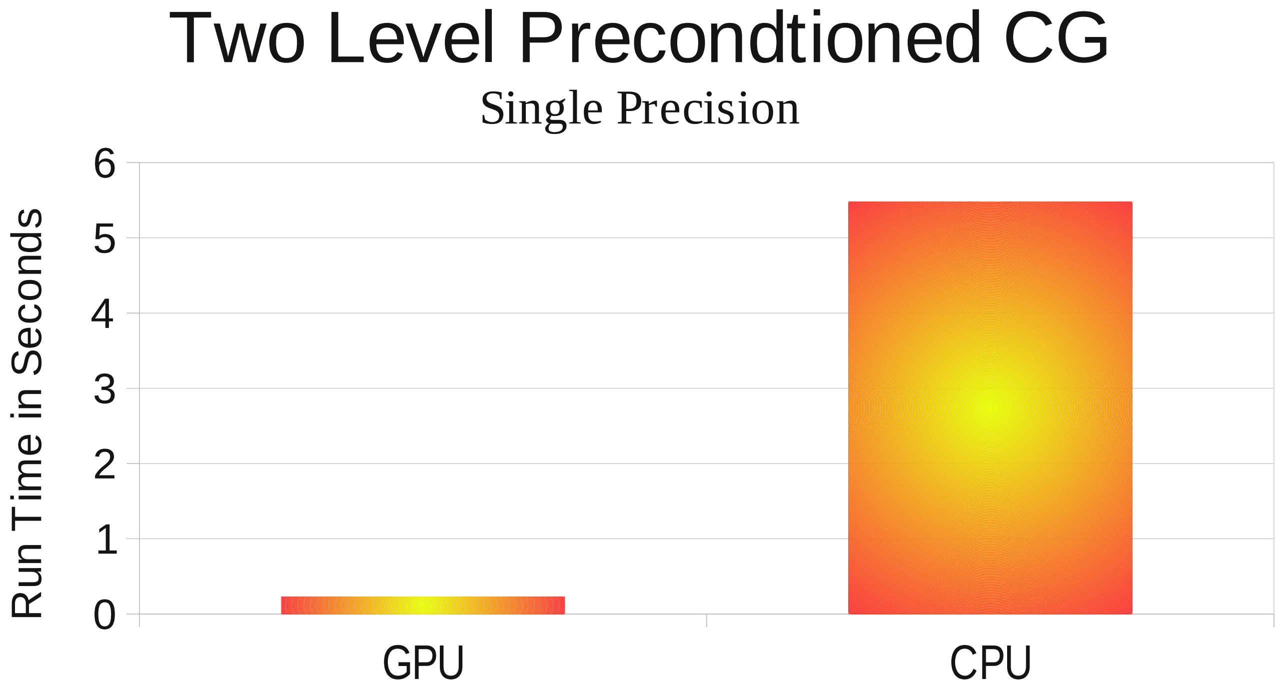
16-05-2011

Content

- Introduction
- GPU Computing
- Solver - Overview
- Preconditioning
- Deflation
- Results
- Conclusions

Bird's Eye View

Comparison of CPU optimized code with respect to GPU code



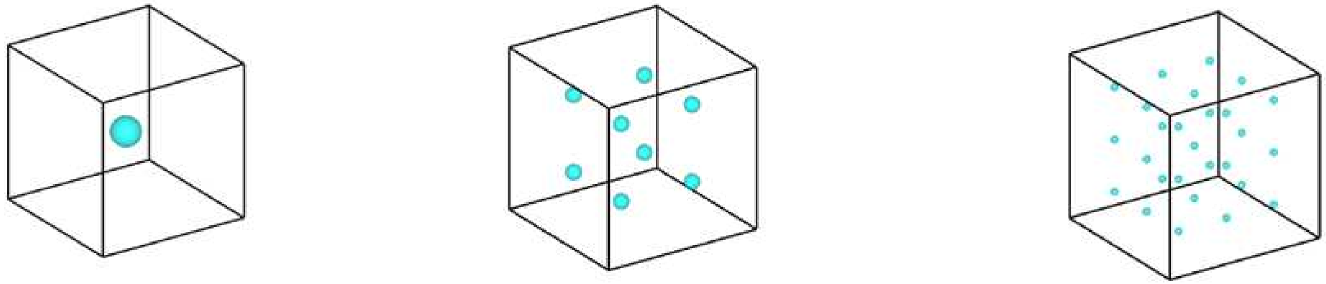
Bird's Eye View

Comparison of CPU optimized code with respect to GPU code

23x

Problem Description

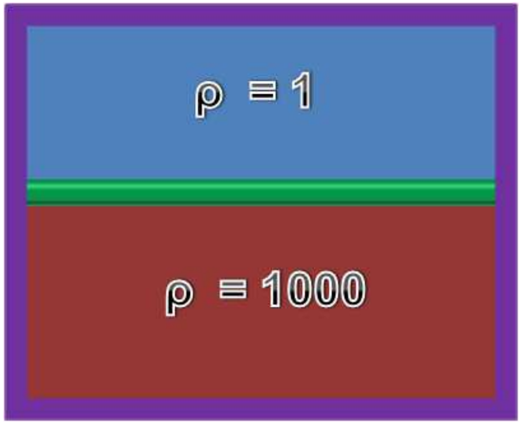
Mass-Conserving Level Set Method to Solve the Navier Stokes Equation.

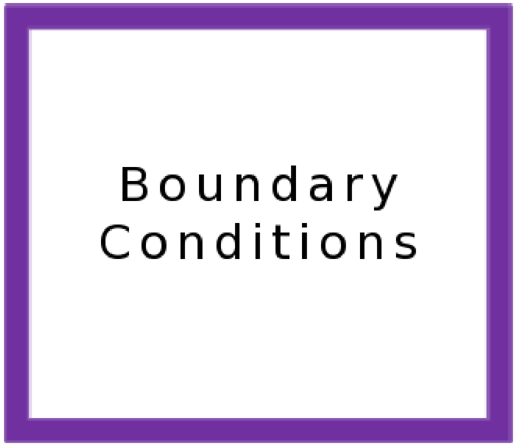


Air bubbles rising in Water.

Computational Model

Computational Hotspot - Solution of the Pressure Correction equation in 2D





$$-\nabla.(\frac{1}{\rho(x)}\nabla p(x)) = f(x),\ x \in \Omega \tag{0}$$

$$\frac{\partial}{\partial n} p(x) = g(x),\ x \in \partial\Omega \tag{0}$$

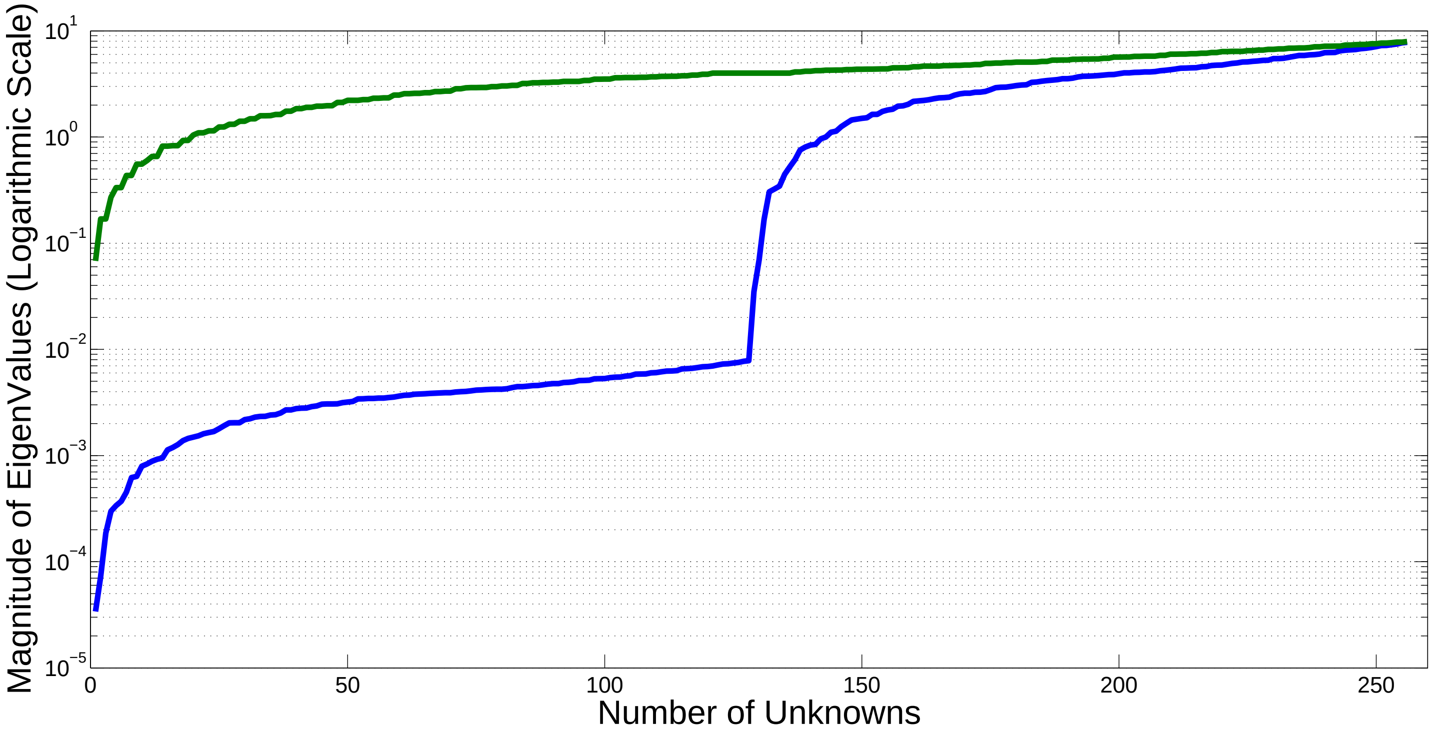
Nature of the Coefficient Matrix

$$Ax = b, \text{ where } x^T A x > 0 \quad (0)$$

A is a sparse matrix with a 5-point stencil. It has a very large condition number due to a huge jump in the density.

- Condition Number $\rightarrow \kappa(A) := \frac{\lambda_n}{\lambda_1}$
- Stopping criterion $\rightarrow \frac{\|b - Ax_k\|_2}{\|r_0\|} \leq 10^{-6}$

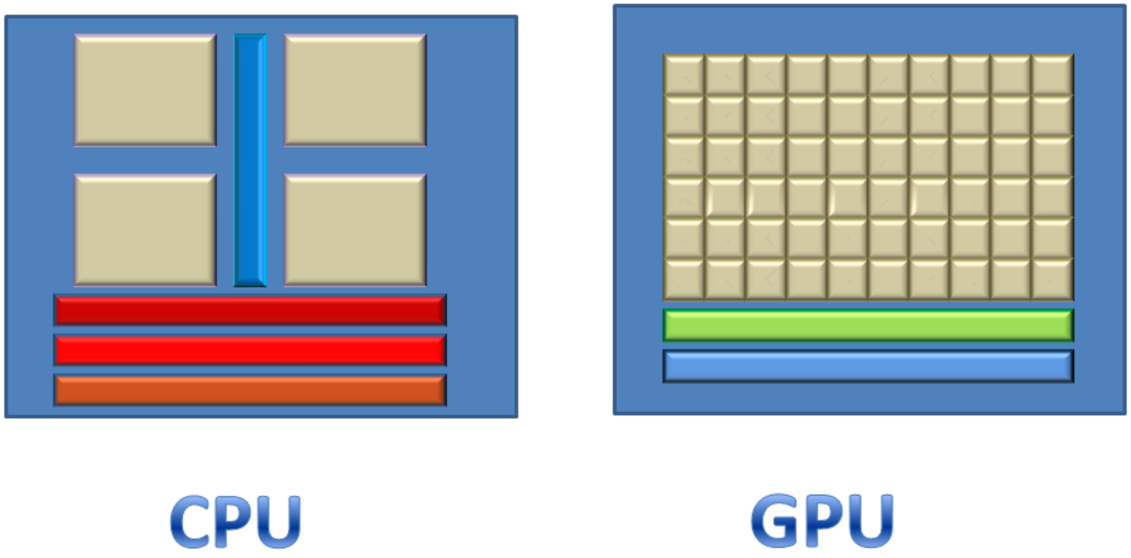
Nature of the Coefficient Matrix



Huge jump at the interface due to contrast in densities.

A Brief Introduction to the GPU

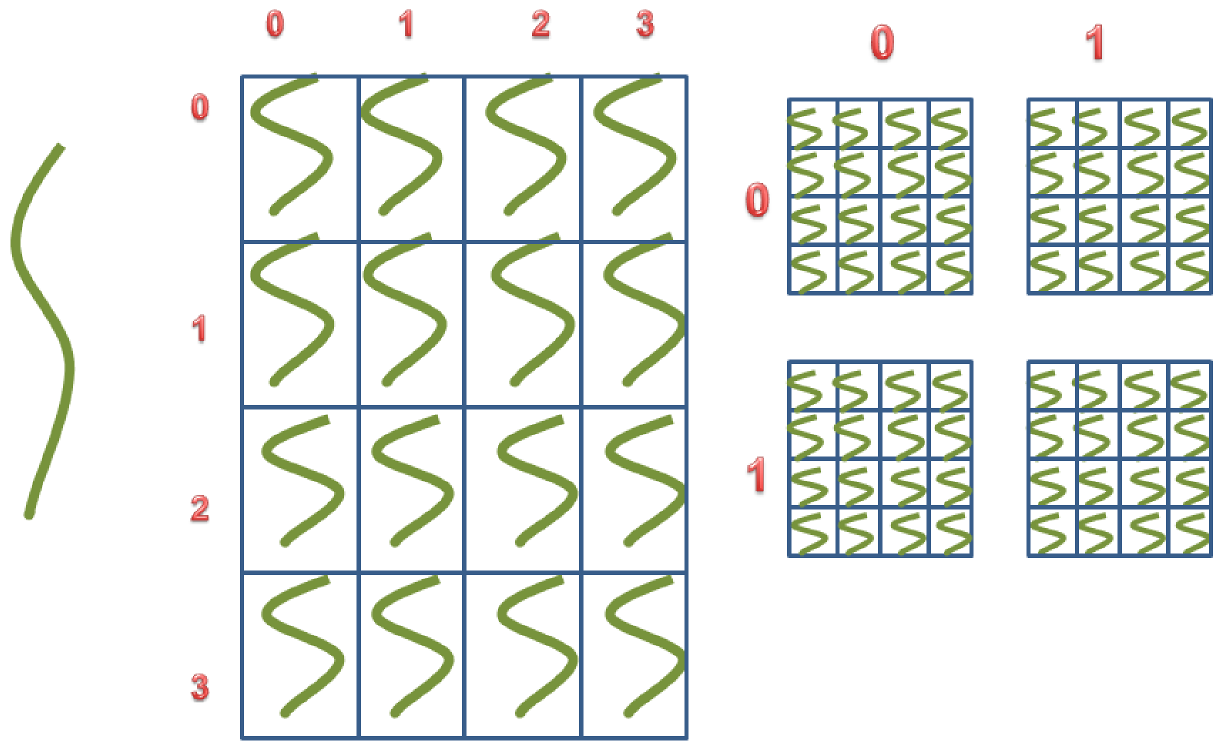
- SIMD Architecture
- Large Memory Bandwidth
- User Managed Caches



A brief Intro.... Contd.

- Basic Unit of Execution - Thread
- Each Thread Executes a Kernel
- Aggregates of Threads = Blocks
- Shared Memory within a block

A brief Intro.... Contd.



A brief Intro.... Contd.

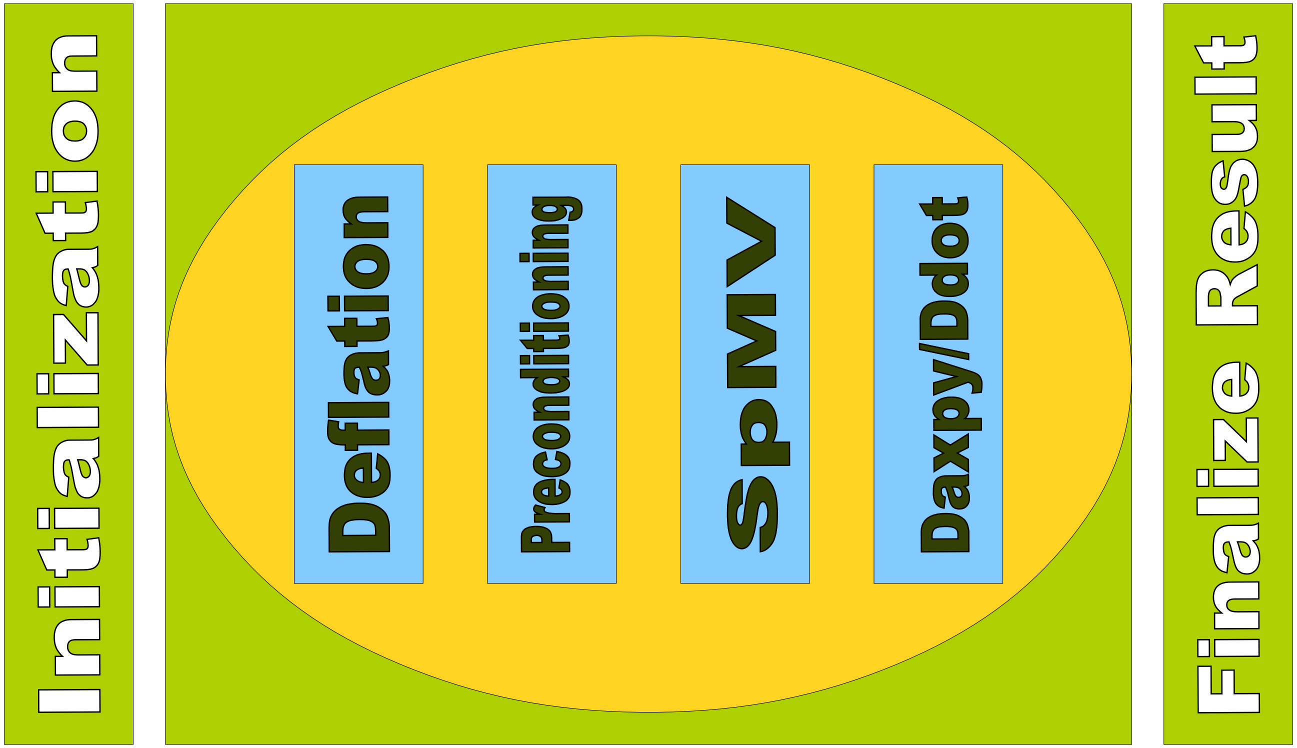
Architecture	TeslaC1060	TeslaC2070(Fermi)
Number of Compute Cores	240 cores	448 cores
Memory Bandwidth	102Gb/s	144 Gb/s
Double Precision Throughput(Peak)	78 Gflops/s	515 Gflops/s
Memory	4GB	6GB
Shared Memory / L1cache configurability	No	Yes

A brief Intro.... Contd.

Most Important Optimizations for GPU Code

- Reduce Host-GPU Transfers
- Maximize use of Memory Bandwidth
- Minimize Thread Divergence
- Utilize Shared Memory/ L1 cache based on kernel

Control Flow in the Algorithm



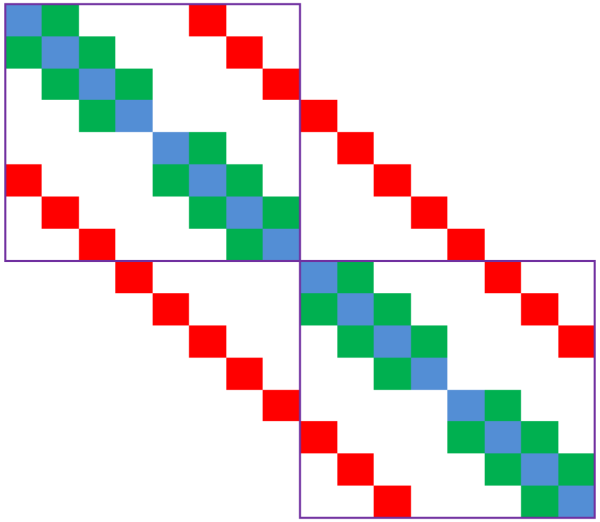
Conjugate Gradient with Two Level Preconditioning

Preconditioning

- Diagonal Preconditioning
- Block Incomplete Cholesky
- Incomplete Poisson(IP)
- Modifications based on IP

Preconditioning

Block Incomplete Cholesky Preconditioning [a] [b]



Within blocks the computation is sequential

[a] An Iterative Solution Method for Linear Systems of Which the Coefficient Matrix is a Symmetric M-Matrix. J.A. Meijerink, H.A.van der Vorst (1977).Math. Comp. (American Mathematical Society)

[b] Iterative methods for sparse linear systems. 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, 2003. Yousef Saad

Preconditioning

Incomplete Poisson

$$M = K * K^T, \text{ where } K = (I - L * D). \quad (0)$$

$$Stencil for A = (-1, -1, 4, -1, -1). \quad (0)$$

$$Corresponding Stencil for M^{-1} = (\frac{1}{4}, \frac{1}{16}, \frac{1}{4}, \frac{9}{8}, \frac{1}{4}, \frac{1}{16}, \frac{1}{4}) \quad (0)$$

Drop the lowest terms (i.e. $\frac{1}{16}$).

M^{-1} has the same sparsity pattern as A.

Degree of Parallelism for $M^{-1} * r$ is N.

A Parallel Preconditioned Conjugate Gradient Solver for the Poisson Problem on a Multi-GPU Platform, M. Ament. PDP 2010

Preconditioning

Incomplete Poisson Variants [a]

Scaling of A matrix. $\hat{A} = D^{-\frac{1}{2}} * A * D^{-\frac{1}{2}}$

As parallel as IP and as effective as Block-IC

Slightly more computations compared to IP.

---

[a] A vectorizable variant of some ICCG methods. Henk A. van der Vorst, SIAM Journal of Scientific Computing. Vol. 3 No. 3 September 1982.

Deflation

Operations involved in deflation [a] [b]

- $b = Z^T * x$
- $m = E^{-1}b$
- $w = A * Z * m$
- $w = x - w$

where, E is the Galerkin Matrix and Z is the matrix of deflation vectors.

---

[a] Efficient deflation methods applied to 3-D bubbly flow problems. J.M. Tang, C. Vuik Elec. Trans. Numer. Anal. 2007.

[b] An efficient preconditioned CG method for the solution of a class of layered problems with extreme contrasts in the coefficients. C. Vuik, A. Segal, J.A. Meijerink J. Comput. Phys. 1999.

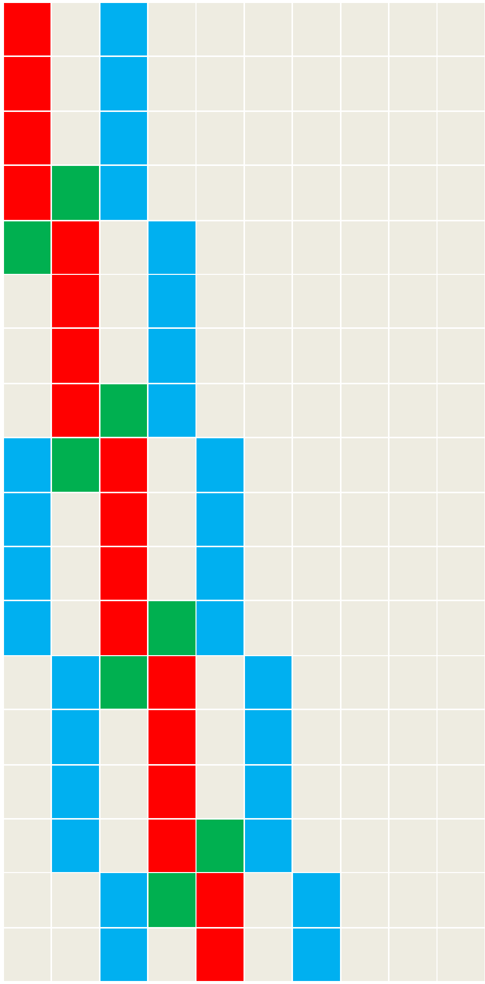
Deflation

Stripe-Wise Domains

57	58	59	60	61	62	63	64
49							56
41							48
33							40
25							32
17							24
9							16
1	2	3	4	5	6	7	8

Delft Institute of Applied Mathematics

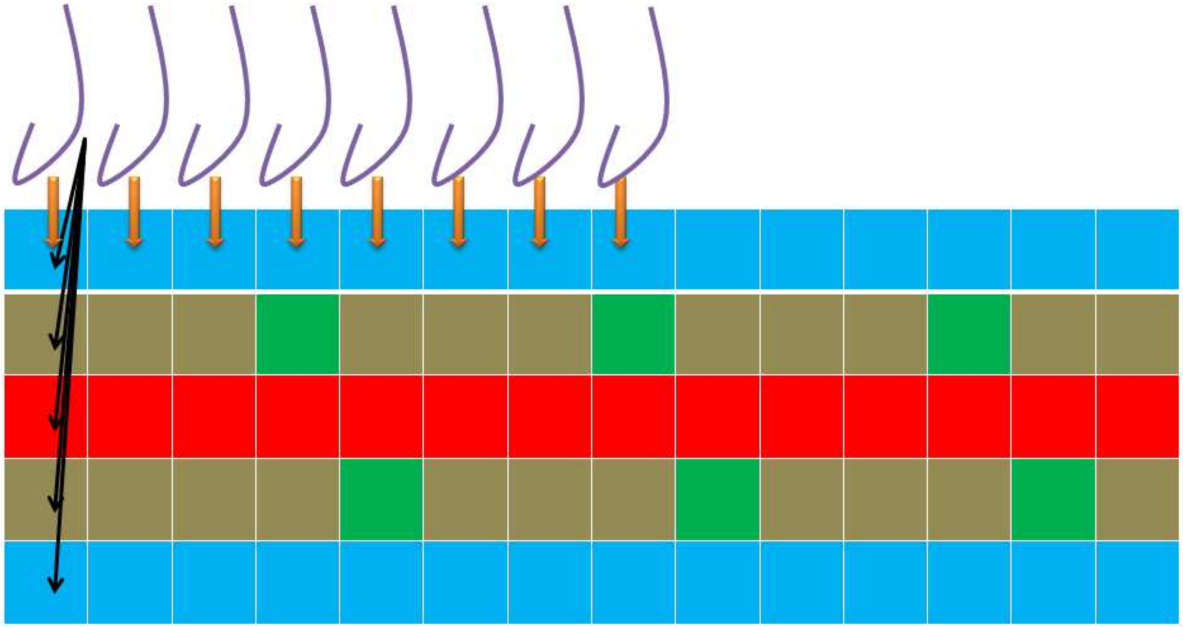
Deflation



Efficient Data Structures

Deflation

Efficient Data Structures



This data Structure has the advantages of the DIA Storage format[a].

[a] Efficient Sparse Matrix-Vector Multiplication on CUDA. N. Bell and M. Garland, 2008 , NVIDIA Corporation, NVR-2008-04

Deflation

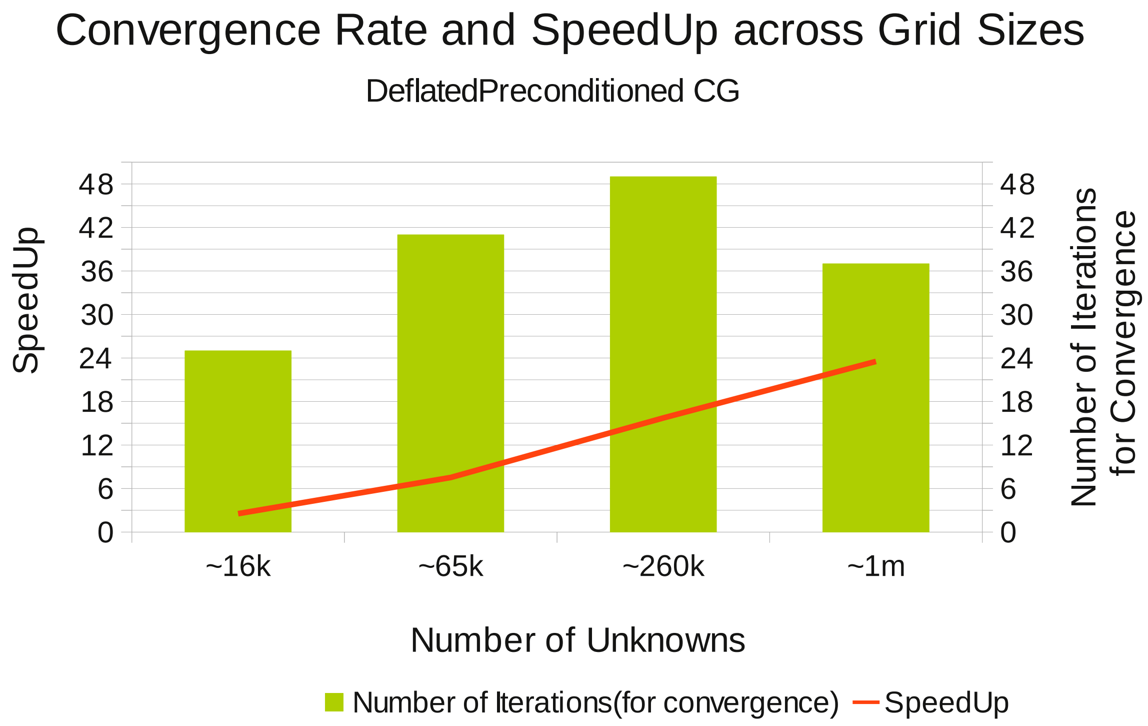
- Breaking Up of Operations
- Stripe-Wise Domains
- Efficient Data Structures

Delft Institute of Applied Mathematics

Results

Single Precision Experiments

SpeedUp and Convergence across GridSizes - Poisson Type Problem



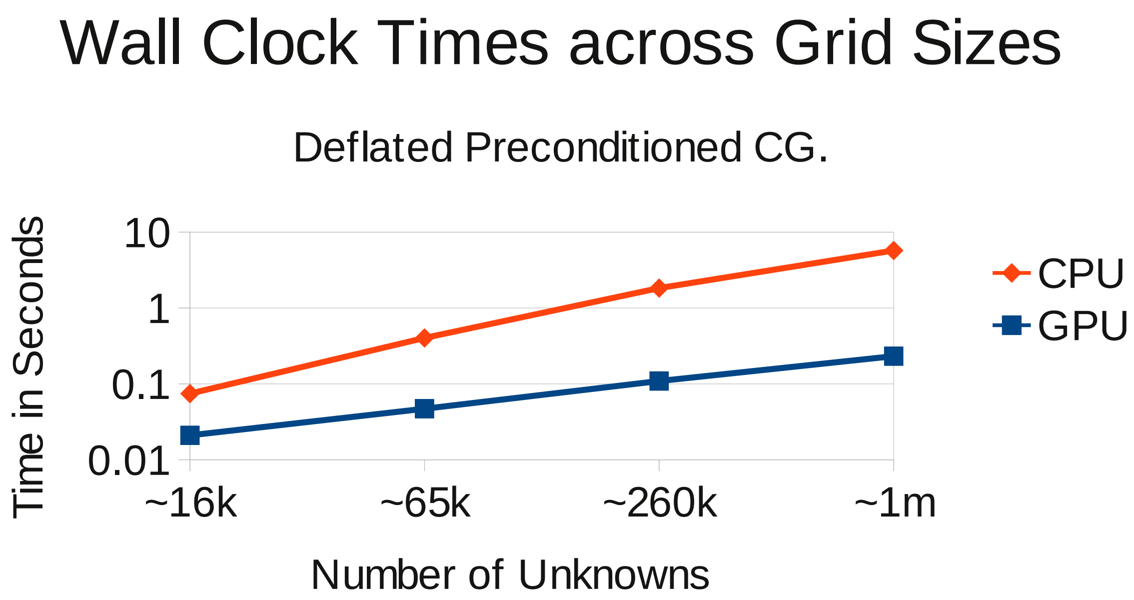
Deflation Vectors are of size $8n$. Precision Criteria is 10^{-5} till $260k$ and for $1m$ it is 10^{-4}

TUDelft

Results

Single Precision Experiments

Wall Clock Times across GridSizes - Poisson Type Problem

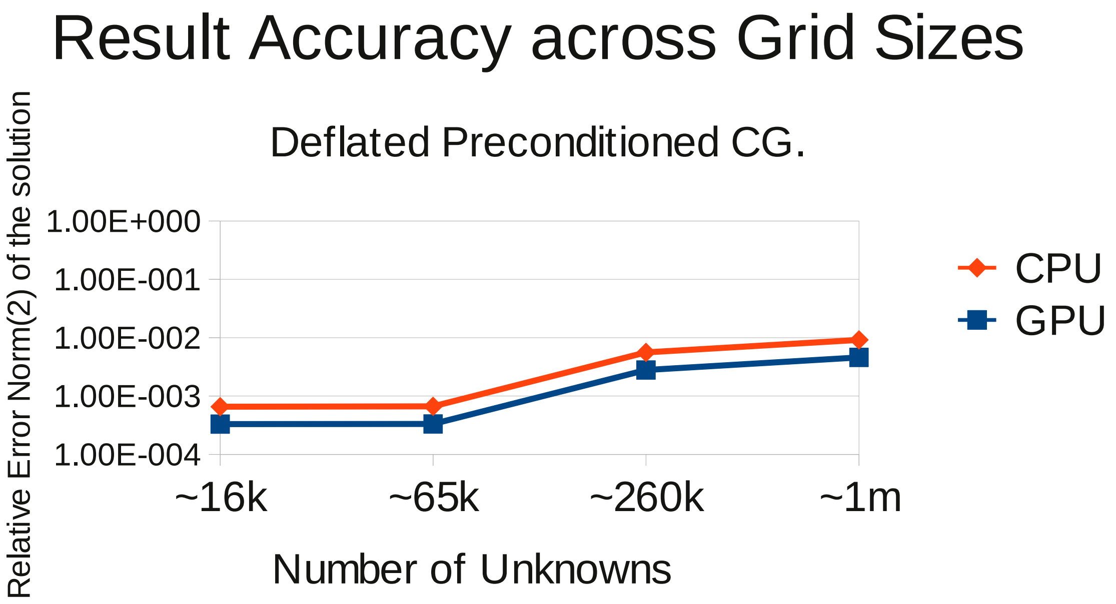


Deflation Vectors are of size $8n$. Precision Criteria is 10^{-5} till $260k$ and for $1m$ it is 10^{-4}

Results

Single Precision Experiments

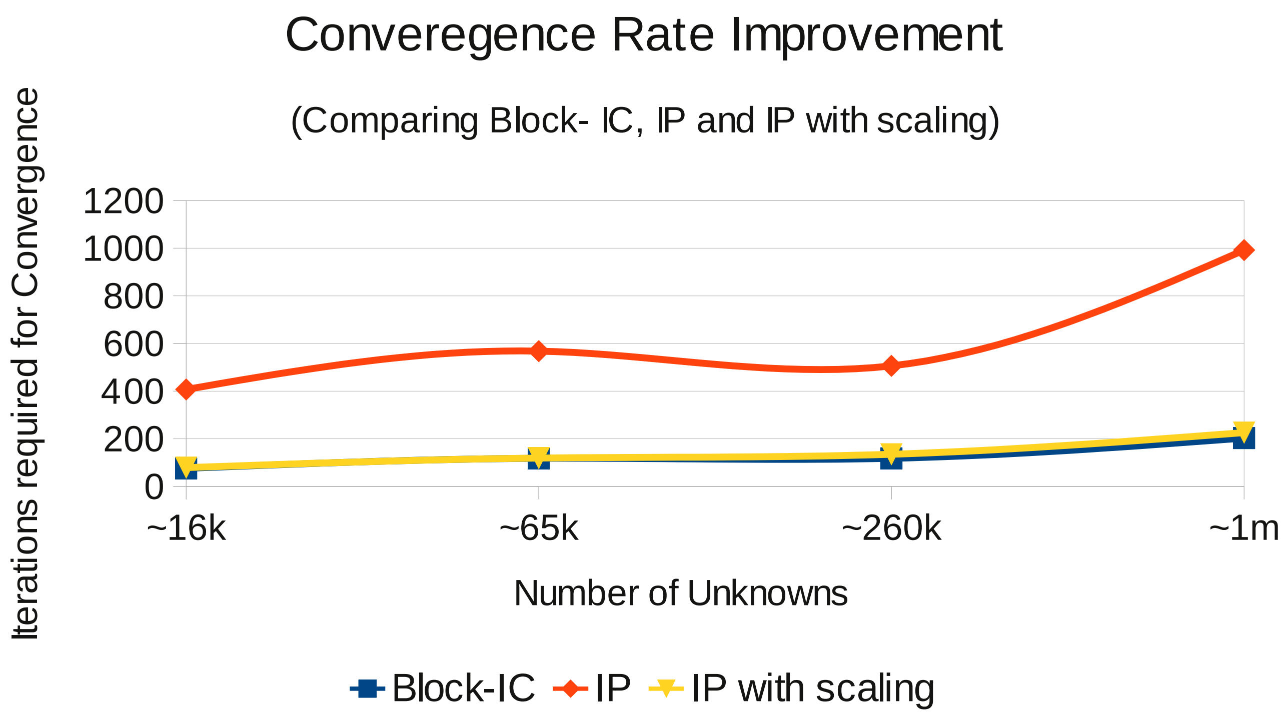
Accuracy across GridSizes - Poisson Type Problem



Deflation Vectors are of size $8n$. Precision Criteria is 10^{-5} till $260k$ and for $1m$ it is 10^{-4}

Results

IP Variants   For two Phase Problem - Double Precision

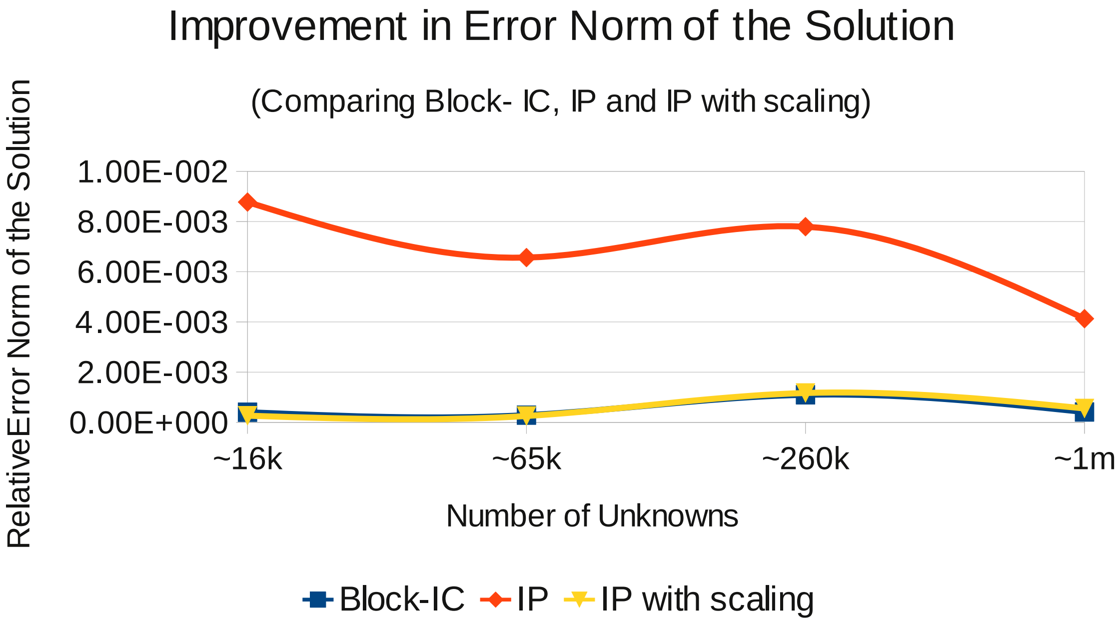


Deflation Vectors are of size $2n$. Precision Criteria is 10^{-6}.

Block size is $2n$ for $n \times n$ grid where $N = n \times n$ and N is the number of unknowns

Results

IP Variants   For two Phase Problem - Double Precision



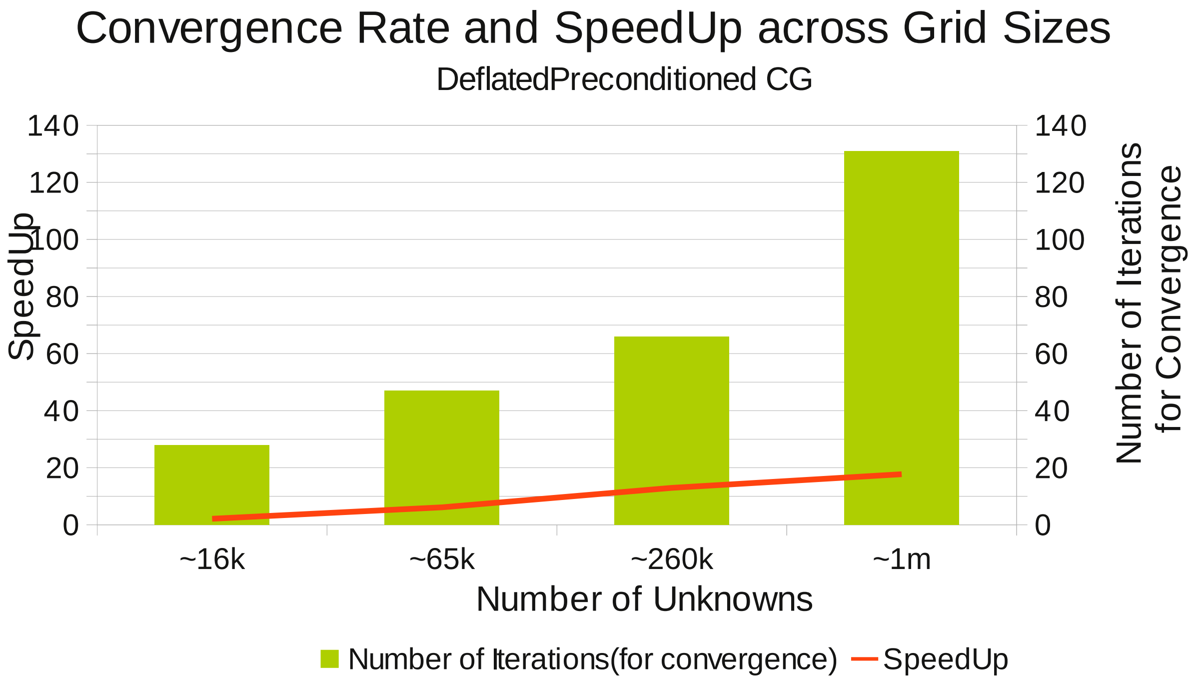
Deflation Vectors are of size $2n$. Precision Criteria is 10^{-6}.

Block size is $2n$ for $n \times n$ grid where $N = n \times n$ and N is the number of unknowns

Results

Double Precision Experiments

SpeedUp and Convergence across GridSizes - Two-Phase Problem

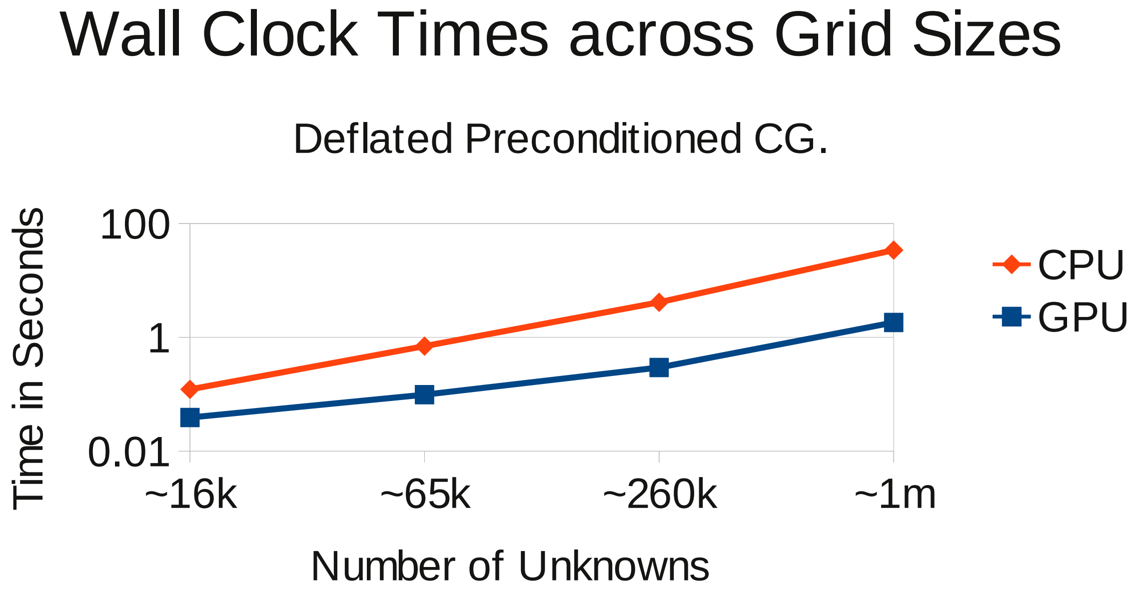


Deflation Vectors are $8n$. Density Contrast $(1000:1)$. Precision Criteria is 10^{-6}.

Results

Double Precision Experiments

Wall Clock Times across GridSizes - Two-Phase Problem

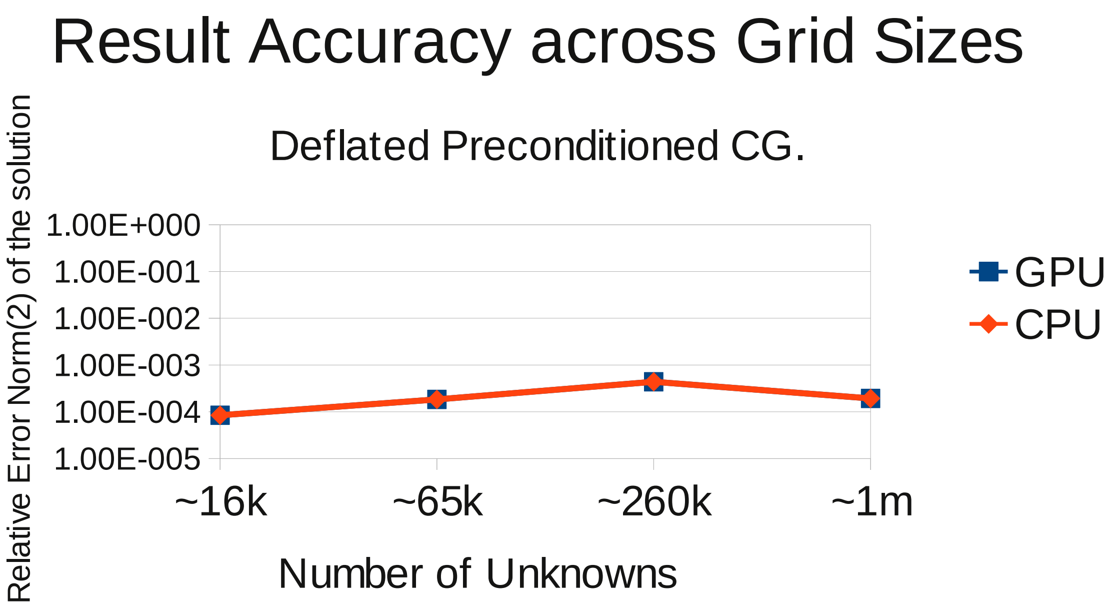


Deflation Vectors are $8n$. Density Contrast $(1000 : 1)$. Precision Criteria is 10^{-6}.

Results

Double Precision Experiments

Accuracy across GridSizes - Two-Phase Problem



Deflation Vectors are $8n$. Density Contrast $(1000 : 1)$. Precision Criteria is 10^{-6}.

Conclusions

- Deflation is highly parallelizable.
- Suitable Preconditioning for GPU must be used.
- More optimizations in order for Double Precision results.

Further Information

- Masters Thesis of Rohit Gupta
  http://ta.twi.tudelft.nl/users/vuik/numanal/gupta_eng.html
- GPU page
  http://ta.twi.tudelft.nl/users/vuik/gpu.html
- Open Source GPU software
  http://ta.twi.tudelft.nl/users/vuik/gpu.html#software