Domain decomposition method on GPU cluster for Lattice QCD

Yusuke Osaki
Ken-Ichi Ishikawa
Hiroshima Univ.

0. Outline

1. Motivation of GPU cluster & Domain–Decomposition preconditioner

   I will give our motivation for GPU cluster and Domain–Decomposition preconditioner.

2. Domain–Decomposition preconditioner

   There are 2 kind of approaches for the Domain–Decomposition Schwarz preconditioner. One is the mulitiplicative Schwarz preconditioner, and another is the additive Schwarz preconditioner.

   2. We will give some features for both approaches

3. Results

   1. We implemented the additive Schwarz preconditioned mixed–precision Bi–CGStab solver for GPU cluster. And we investigated the performance of the additive Schwarz preconditioner. I will show some results on the performance.

4. Summary & conclusion

   1. I will summarize my talk.

1.1 Previous studies

- There are many studies using GPU in Lattice QCD simulations.
  - G.I.Egri, Z.Fodor, C.Hoelbling, S.D.katz, D.Nogradi, and K.K.Szabo, hep-lat/0611022
  - M.A.Clark, R.Babich, K.Barros, R.C.Brower, C.Rebbi, hep-lat/0911.3191
  - K.Z.Ibrahim, F.Bodin, O.Pene, Journal of Parallel and Distributed Computing, 68,1350
  - C.Chen, E.Dzienkowski, J,Giedit, hep-lat/1005.3276
  - V.Anselmi, G.Conti, F.D.Renzo, hep-lat/0811.2111
  - M.A.Clark
  - etc.
  - And many contributions in this conference.

I'm sorry for not correctly citing your contribution

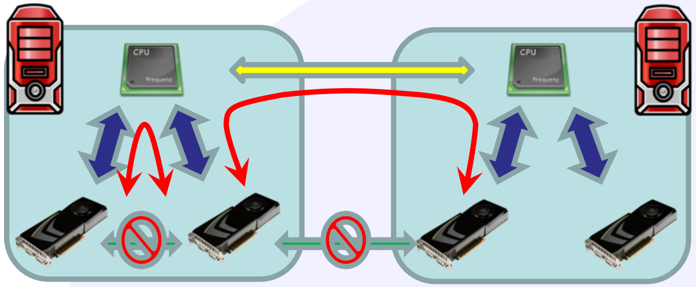
- It has been reported that the GPU can achieve **over 100 GFlops** in single precision using single GPU card for the Wilson kernel of quark solvers.
- The situation is still limited in single GPU case but….there is a limitation in lattice volume size because memory on single GPU is limited at a few GBytes and the total speed for realistic large volume simulations. Therefore we need multiple GPU simulations.

1.2 Motivation of GPU cluster and its difficulty

- There is a limitation in lattice volume size because memory on single GPU is limited at a few GBytes and the total speed for realistic large volume simulations. Therefore we need multiple GPU simulations.
- For example, to do more realistic lattice QCD simulations, such as 64^4, a^{-1} = 2 GeV, M_q = 5MeV, we need at least over O(10) TFlops sustained speed. This is impossible with single GPU.
- In this talk, I will show the results of the additive Schwarz preconditioner on multiple GPU cluster.

1.3 Difficulty in parallel computation on GPU cluster

- It is well known that, the parallel computation on GPU clusters is difficult because there is no capability to direct data communication between GPU cards.



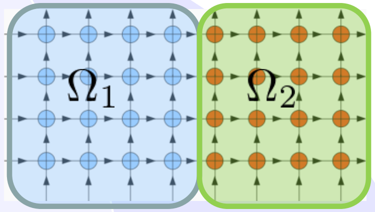
- Therefore, in order to exchange data among GPU cards, data should be exchanged through CPU memory and network devices. This means that we have to do 2 or 3 steps for data exchanging.
- This could be a bottleneck of parallel computation on GPU cluster.
- To remove the bottleneck, we investigated the Schwarz Domain-Decomposition preconditioner for quark solver on GPU cluster.

- I will briefly explain the details of the Schwarz Domain-Decomposition preconditioner.

2.1 The Schwarz Domain-Decomposition preconditioner

- We want to solve the equation $D\psi = b$, where D is the Wilson–Dirac operator, ψ & source b are quark fields. for example we decompose the fields into two domains.
- The above equation can be written in the following 2 × 2 form.

$$\begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix} \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

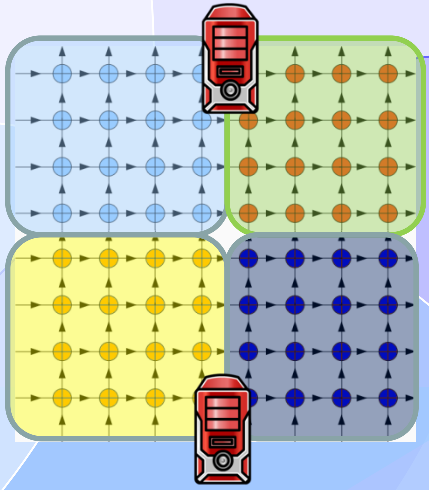


- The multiplicative Schwarz preconditioner, K, is as follows

$$K = \begin{pmatrix} D_{11}^{-1} & 0 \\ -D_{22}^{-1} D_{21} D_{11}^{-1} & D_{22}^{-1} \end{pmatrix}$$

```
$x = 0, r = 0$
for
    $v = Kr$
    $x = v; r = r - Dv$
end for
$x \to D^{-1}b$
```

```
$x = 0, r = 0$
for
    $v = D_{11}^{-1} r$
    $x = v; r = r - Dv$
    $v = D_{22}^{-1} r$
    $x = v; r = r - Dv$
end for
$x \to D^{-1}b$
```

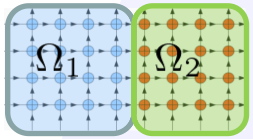


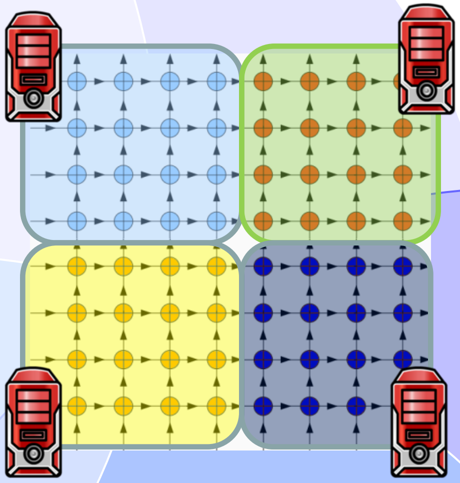
  - This method has been introduced by Lüsher [CPC 156 (2004)].
- The data communication between domains can be reduced with the node assignment as in this figure.
  - To avoid the node idling, each node should contain both of the Ω_1 and the Ω_2 domains.

2.2 The additive Schwarz Domain-Decomposition preconditioner

- The multiplicative Schwarz preconditioner has a dependency between the Ω_1 and the Ω_2 domains. This preconditioner corresponds to the Block Gauss–Seidel iteration.
- There is another version of the Schwarz Domain–Decomposition preconditioner, it is the additive Schwarz preconditioner which corresponds to the Block Jacobi iteration. In this case, the dependence of between Ω_1 and Ω_2 is removed.
- The additive Schwarz preconditioner, *K*, is as follows.

$$K = \begin{pmatrix} D_{11}^{-1} & 0 \\ \underline{0} & D_{22}^{-1} \end{pmatrix}$$





```
$x = 0, r = 0$
for
    $\underline{v = Kr}$
    $x = v; r = r - Dv$
end for
$x \rightarrow D^{-1}b$
```

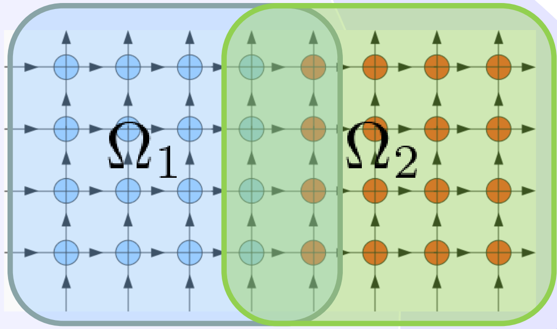
```
$x = 0, r = 0$
for
    $\underline{v = D_{11}^{-1}r + D_{22}^{-1}r}$
    $x = v; r = r - Dv$
end for
$x \rightarrow D^{-1}b$
```

- In this case, one node contains one domain as in right figure. The degree of parallelization is higher than that of the multiplicative case.
- However, it is well–known that multiplicative one has better performance than that of additive one (This is similar to the relation between the Gauss–Seidel and Jacobi preconditioner)

2.3 Overlapped Domain-Decomposition

- We can also overlap the decomposed domains as in the following figure
  - In the additive Schwarz preconditioner, the pseudo−code is as follows.

```
$x = 0, r = 0$
for
   Recive the data of overlapped sites
   $v = D_{11}^{-1} r + D_{22}^{-1} r$   with restriction to the original domain
   $x = v; r = r - Dv$
end for
x $\rightarrow D^{-1} b$
```



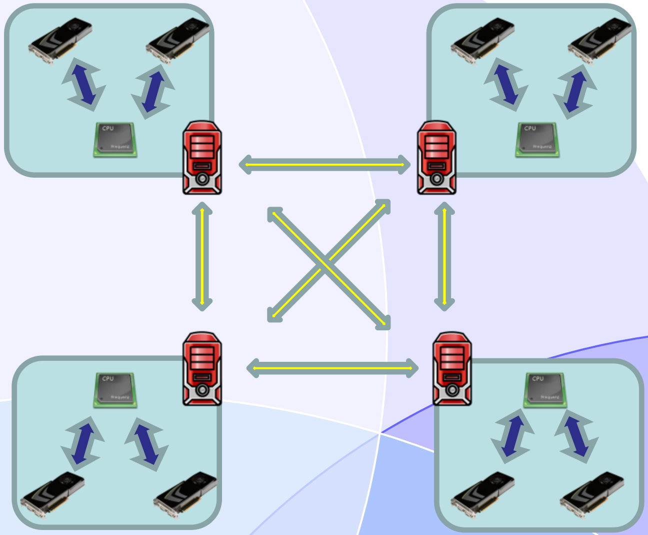
  - This method is called Restricted Additive Schwarz preconditioner.
- We can reduce the iteration count by overlapping domains.
  - Matrix vector multiplication count is reduced.
  - But the domain inversion requires extra data exchanging for overlapped sites.
  - There are tradeoff between communication reduction from iteration reduction and communication increment from overlapping domains.
  - So it is not trivial that overlapping domains improves the performance.

2.4 The Multiplicative & Additive Schwarz preconditioners

- The data communication between domains can be reduced with the Schwarz Domain-Decomposition preconditioner.
- We would like to compute the preconditioner on multiple GPUs in parallel.
- It is well known that larger domain size is preferable for GPU computation. We can assign larger domain size to a GPU card with the additive Schwarz preconditioner than with the multiplicative preconditioner by a factor two.
- Therefore, we investigated the timing and flop performance of the mixed-precision Bi-CGStab solver with the additive Schwarz preconditioner on multiple GPU cluster.

3.1 Tested machine

- We investigated the timing and flop performance of the mixed-precision Bi-CGStab solver with the additive Schwarz preconditioner on multiple GPU cluster.



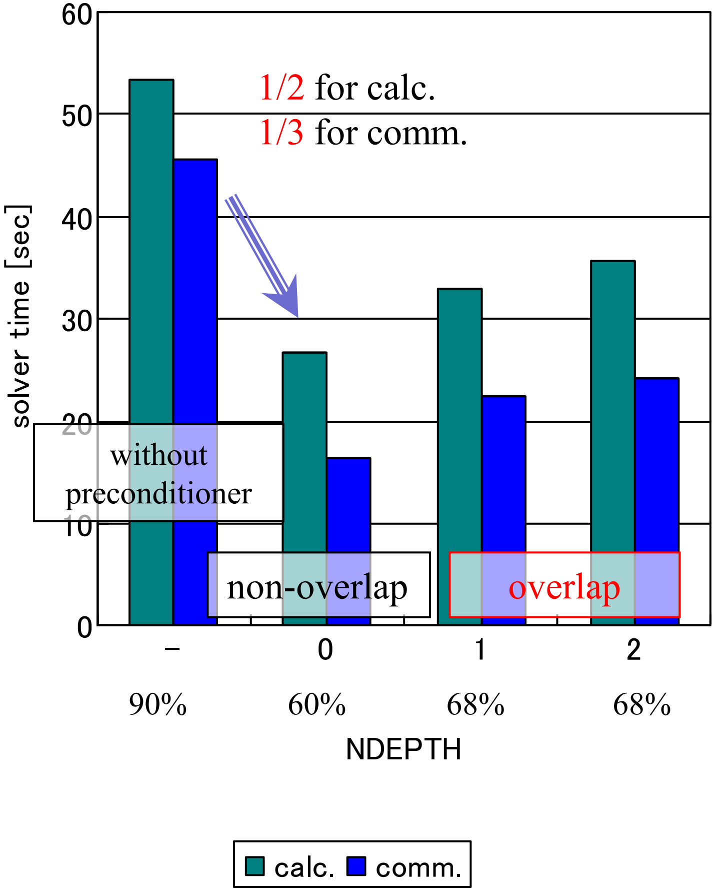
- We employed the following PC cluster.
  - GPU : GeForce GTX 285 × 2 × 4
  - CPU : intel Core i7 920 × 4
  - RAM : 6 GByte DDR3 memory ×4
  - LAN Adapter : intel Gigabit ET Quard Port Server
- We used the cheep Gigabit ethernet cards, instead of expencive network cards such as Infiniband, Myrinet, or 10 GBit ethernet cards.
  - However, to improve MPI performance, we used Open-MX protocol instead of TCP/IP protocol. The Open-MX library is freely available.
    - Open-MX : http://open-mx.gforge.inria.fr/

3.2 Tested Parameters

- We employed the following softwares.
  - OS : Cent OS 5.3.
  - programing language : intel Fortan & Nvidia CUDA 2.3 (version two point three.)
  - OpenMPI. (library for data communication which is enhanced with the Open-MX protocol)
- The details of GPU solver are as follows.
  - The host code is the mixed-precision BiCGStab solver
  - The GPU code is the single-precision BiCGStab solver with the addive Schwarz preconditiner (after even/odd site preconditioning.)
  - We solved the O(a)-improved Wilson quark propagator on a random gauge configuration.
  - We measured the timing and the performance on muliple GPU cluster in the parameters as follows.
    - kappa = 0.126
    - csw = 1.0
    - solver stopping condition is $|b - Dx|/|b| < 10^{-14}$

3.3 Results Using Domain-Decomposition method

The timing of the solver convergence and the communication on a 32^4 lattice.



- We observed that improvement the total convergence time with the additive Schwarz preconditioner.
- The reason of the this improvement mostly relies on the reduction of the communication time.
- The additive Schwarz preconditioner acctually reduces the communication overhead.
- The overlapping domain region does not work well.
- These results are obtained on a 32^4 lattice. We also observed similar results on a 16^4 lattice.
- However the improvement is slightly weaker on the 16^4 lattice.

4.1 Summary & Conclusion (1)

- We studied the parallelization of Lattice QCD solver on the multiple GPU cluster.
  - Communication between GPUs is a bottleneck of parallelization on multiple GPU clusters, because there is no direct communication device on commodity GPU cards.
- We investigated the Domain-Decomposition preconditioner to reduce the communication overhead without expensive network cards.
  - We tried the Restricted Additive Schwarz preconditioner with non-overlapping and overlapping domains.
  - (We can reduce the communication time with the Schwarz method.)
  - One can keep larger lattice volume on each GPU with the additive Schwarz method than that with the multiplicative Schwarz method. Thus higher GPU efficiency is expected.
- Using 4 node PC cluster attached with 2 GPU cards on each PC, we measured the timing and flop performance of the mixed-precision Bi-CGStab solver with the additive Schwarz preconditioner (on a random 32^4 gauge configuration with the clover fermion action).

4.2 Summary & Conclusion (2)

- The timings of the solver are compared in two cases, those with and without the Schwarz preconditioner.
  - The total solver timing was reduced by a factor 2, and the ratio of communication time over total time was reduced to 60% from 90%, by using the additive Schwarz preconditioner.
  - We also investigated the effect of overlapping of domains in the additive Schwarz preconditioner. We did not observe any improvement by the overlapping.
- We can relax the communication bottleneck of multiple-GPU cluster by using the additive Schwarz preconditioner.
- The overlapping additive Schwarz preconditioner does not work.

Thank you !