PGAS implementation of Sparse Matrix Vector Multiplication and Lattice Boltzmann Method using GPI

Faisal Shahzad
Outline

- Motivation
- GPI Introduction
- Experimental framework
- SpMVM with GPI, results
- LBM with GPI, results
- Conclusion
Motivation & Introduction

- Can communication be more efficient?
- Can communication be made truly asynchronous?

GPI: Global address space Programming Interface
- A PGAS model API developed by Fraunhofer ITWM, Kaiserslautern, Germany
- Why GPI (our motivation)
  - GPI targets to incorporate fault tolerant behavior
  - Fault tolerance: Node failures do not crash the whole GPI application

Focus in this talk: performance comparison between GPI and MPI variants of the program
Two memory parts

- Local: only local to the GPI process (and its threads)
- Global: Available to other processes for reading and writing.

Hybrid approach (like MPI/OpenMP)

- Each Node(NUMA) domain can have only one GPI process
- MCTP/OpenMP threads within node(NUMA) domain
Experimental Setup

- **Applications:**
  - A prototype CFD solver based on a Lattice Boltzmann Method (LBM)
  - Sparse Matrix Vector Multiplication (SpMVM) algorithm

- **Approaches compared:**
  - Blocking MPI communication
  - Non-blocking MPI communication
  - Non-blocking MPI communication with explicit non-blocking communication support (APSM library; *details next slide*)
  - Synchronous GPI communication
  - Asynchronous GPI communication

- **Cluster:**
  - LiMa (Erlangen): 500 nodes (Dual socket Intel Xeon 5650 “Westmere”), QDR Infiniband, Lustre based PFS Bandwidth ~ 3GB/s
Asynchronous Progress Support for MPI (APSM) Library

- Motivation - MPI’s non-blocking calls are not necessarily asynchronous

Dec. 16th '13
SpMVM with GPI (I)

\[ \overrightarrow{y} = \mathbf{A} \overrightarrow{x} \]

\[ y_i = \sum_j (A)_{i,j} \cdot x_j. \]

- Each process has RHS corresponding to its matrix rows.
- Each process requires remote RHS values e.g. 0th process: 2,3,4,5,7
- Resultant vector can be seen as summation of two components
  1. local-part
  2. remote-part

Where \( \mathbf{A} \) is an \( nxn \) matrix and \( x,y \) are \( n \) dimensional vectors.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>11</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16</td>
<td>17</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SpMVM with GPI (II)

\[
\tilde{y} = A \tilde{x}
\]

\[y_i = \sum_j (A)_{i,j} \cdot x_j.\]

- Before each iteration of SpMVM, the required RHS values need to be fetched from remote processes.

- The communication can be hidden behind local part of the SpMVM computation.

Where \(A\) is an \(n \times n\) matrix and \(x, y\) are \(n\) dimensional vectors.
SpMVM with GPI (III): Synchronous communication case

```c
SpMVM_sync(mat, rhs, res)
{
    rhs->communicate_remote_RHS_blocking();
    spMVM(mat, rhs, res);
}
```

The SpMVM function with synchronous MPI communication

- Each process gathers (read/write) all its required RHS elements before the result vector is computed.
- Local and remote parts are calculated together.
- Synchronization is necessary for one-sided communication (GPI).

```c
SpMVM_sync(mat, rhs, res)
{
    rhs->communicate_remote_RHS_one_sided();
    rhs->wait();
    sync();
    spMVM(mat, rhs, res);
    sync();
}
```

The SpMVM function with synchronous GPI communication
SpMVM with GPI (IV): Asynchronous communication case

```c
SpMVM_async(mat, rhs, res)
{
    rhs->communicate_remote_RHS_non-blocking();
    spMVM_local(mat, rhs, res);
    rhs->wait();
    spMVM_remote(mat, rhs, res);
}
```

- Local and remote parts are calculated separately.
- Communication is done asynchronously with the local computation part.
- For one-sided communication (GPI), a barrier is essential before and after communication.
SpMVM benchmark results (I)

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Dimension</th>
<th>Avg. NNZ per row</th>
<th>size in MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRZE3</td>
<td>$6.2 \cdot 10^6$</td>
<td>19</td>
<td>1530</td>
</tr>
<tr>
<td>DLR1</td>
<td>$2.8 \cdot 10^5$</td>
<td>144</td>
<td>642</td>
</tr>
<tr>
<td>HV15R</td>
<td>$2 \cdot 10^6$</td>
<td>140</td>
<td>4545</td>
</tr>
<tr>
<td>RM07R</td>
<td>$3.8 \cdot 10^5$</td>
<td>98</td>
<td>602</td>
</tr>
</tbody>
</table>

The performance depends on the structure of the matrix.

Num. of nodes: 32
LBM with GPI (I)

- A CFD solver based on a simplified kinetic approach derived from the Boltzmann equation

- **Time and space discretization (D3Q19 Lattice)**
  - Fluid particles are positioned in certain lattice sites
  - May move only in certain, fixed directions
  - The probability of fluid particles to move in certain direction (distribution function) is calculated in each timestep.

- **Algorithm**
LBM with GPI (II)

for (int t=1; t <= timesteps; ++t)
{
    update_cells();
    barrier();
    exchange_ghost_cells();
}

LBM iteration loop with synchronous communication

for (int t=1; t <= timesteps; ++t)
{
    update_boundary_cells();
    exchange_ghost_cells_begin();
    update_inner_cells();
    exchange_ghost_cells_end();
}

LBM iteration loop with asynchronous communication

- LBM iteration loop with synchronous communication.
- Barrier is essential before one-sided communication is performed.
- LBM iteration loop with asynchronous communication.
- Communication is overlapped with computational step on the inner cells.
- Barrier is essential after one-sided communication.
LBM iteration loop with asynchronous communication with relaxed sync.

- Flags ("boundary_ready") are used to for inter-process synchronization to avoid the need of a global synchronization.

- Communication is overlapped with computational step on the inner cells.
GPI: Weak scaling  
Cluster: LiMa

Weak scaling: Problem size (domain) increases with the same factor as number of nodes.

- For 96 nodes, the best async. GPI case performs 30% better than naive MPI case.
LBM benchmark results (II)

Communication overlap fraction

\[ \mu = \frac{T_{sync.} - T_{async.}}{T_{comm}} \]

- \( T_{sync.} \) = Runtime with sync. comm.
- \( T_{async.} \) = Runtime with async. comm.
- \( T_{comm} \) = Communication time

LBM Overlap Fraction: MPI vs. GPI

- MPI, nonblocking
- MPI, non-blocking, APSM
- GPI, asynchronous

Number of nodes vs. overlap fraction
Conclusion

- We compared LBM and SpMVM implementations of GPI with their respective MPI implementations.

- Algorithms must be adapted to leverage the PGAS languages.

- For LBM lazy synchronization was possible.

- For SpMVM a global synchronization can not be avoided.
Comments on the ease of implementation

- **LBM:**
  - A naive conversion from MPI to GPI code is very simple.
  - Only additional optimizations require more efforts.
  - Overlap between communication & computation.
  - Avoiding global synchronizations.

- **SpMVM:**
  - Setting up the communication structure is comparatively more time consuming.
  - The main communication routines are simple but require global synchronization (which limits performance).
From GPI to GPI2.0

- **LBM:**
  - Synchronization between neighbours done by GASPI notification messages.

- **SpMVM:**
  - Done: GPI -> GPI2.0
  - TODO: Use of notification messages to avoid global synchronization routines.
Thank you!

Questions?
Fault Tolerance with GPI-2

Faisal Shahzad
Outline

- Toy FT implementation with LBM
- DEMO
- Questions and discussion
Toy FT implementation with LBM

- **Idea:**
  - Reserving \(m\) idle processes at the start of program setup.
  
  - Program utilizes \(n-m\) processes for work (work-group)
  
  - In case of a failed process in 'work-group', a process from 'idle-group' is added to 'work-group'.
  
  - Processes in newly established 'work-group' restart the work from last checkpoint.
Program flow:
Health check routine

```c
void send_msg_to_check_state(gaspi_state_vector_t health_vec, gaspi_rank_t *avoid_list)
{
    for(int i=0; i<numprocs; ++i)
    {
        if(avoid_list[i]!=1)
        {
            ASSERT(gaspi_write(0, 0, i, 0, 0, sizeof(int), 0, GASPI_BLOCK));
        }
    }

    gaspi_return_t retval;
    retval = gaspi_wait(0, GASPI_BLOCK);
    ASSERT(gaspi_state_vec_get(health_vec));

    for(int i=0; i<numprocs; ++i)
    {
        if(health_vec[i]==1){
            avoid_list[i]=1;
        }
    }

    if(myrank==0) print_health_vec(health_vec);
}
```
Questions and discussion

- **Health check: simultaneous call by all processes?**
  - Race condition if not simultaneous
  - Multiple failures at the same time? (1-node, 2-processes failure)

- **Communication with `gaspi_timeout_t`:**
  - Time > 100000… call seems blocking? (at failure, process does not return from gaspi_read)
  - Usage with collective communication? (at failure, seem to be GASPI_BLOCK currently)
  - How can an optimized `gaspi_timeout_t` be determined.