

A bird's eye view of Matrix Distributed Processing

Massimo Di Pierro

School of Computer Science, Telecommunications and Information Systems
DePaul University, 243 S. Wabash Av., Chicago, IL 60604, USA

Abstract. We present Matrix Distributed Processing, a C++ library for fast development of efficient parallel algorithms. MDP is based on MPI and consists of a collection of C++ classes and functions such as `lattice`, `site` and `field`. Once an algorithm is written using these components the algorithm is automatically parallel and no explicit call to communication functions is required. MDP is particularly suitable for implementing parallel solvers for multi-dimensional differential equations and mesh-like problems.

1 Introduction

Matrix Distributed Processing (MDP) [1] is a collection of classes and functions written in C++ to be used as components for fast development of efficient parallel algorithms. Typical algorithms include solvers for partial differential equations, mesh-like algorithms and various types of graph-based problems. These algorithms find frequent application in many sectors of physics, engineering, electronics and computational finance.

MDP components can be divided into two main categories:

- Non parallel components: Linear Algebra components (class `mdp_complex`, class `mdp_array`, class `mdp_matrix`) and Statistical Analysis components (class `Measure`, class `Jackboot`)
- Parallel components: (class `mdp_lattice`, class `mdp_site`, class `mdp_field`, etc.)

In this paper we will focus exclusively on the Linear Algebra and the Parallel components¹.

MDP is based on MPI and can be used on any machine with an ANSI C++ and support for the MPI communication protocol. No specific communication hardware is required but a fast network switch is suggested. MDP has been tested on Linux PC clusters, SUN workstations and a Cray T3E.

The best way to introduce MDP is to write a program that solves a typical problem:

¹ The parallel components can interoperate with other third party C/C++ linear algebra packages and can be used to parallelize existing applications with minimal effort.

Problem: Let's consider the following differential equation:

$$\nabla^2 \varphi(x) = f(x) \quad (1)$$

where $\varphi(x)$ is a field of 2×2 Complex matrices defined on a 3D space (**space**), $x = (x_0, x_1, x_2)$ limited by $0 \leq x_i < L_i$, and

$$\begin{aligned} L &= \{10, 10, 10\}, \\ f(x) &= A \sin(2\pi x_1 / L_1), \\ A &= \begin{pmatrix} 1 & i \\ 3 & 1 \end{pmatrix} \end{aligned} \quad (2)$$

The initial conditions are $\varphi_{initial}(x) = 0$. We will also assume that $x_i + L_i = x_i$ (torus topology).

Solution: In order to solve eq. (1) we first discretize the Laplacian ($\nabla^2 = \partial_0^2 + \partial_1^2 + \partial_2^2$) and rewrite it as

$$\sum_{\mu=0,1,2} [\varphi(x + \hat{\mu}) - 2\varphi(x) + \varphi(x - \hat{\mu})] = f(x) \quad (3)$$

where $\hat{\mu}$ is a unit vector in the discretized space in direction μ . Hence we solve it in $\varphi(x)$ and obtain the following a recurrence relation

$$\varphi(x) = \frac{\sum_{\mu=0,1,2} [\varphi(x + \hat{\mu}) + \varphi(x - \hat{\mu})] - f(x)}{6} \quad (4)$$

The following is a typical MDP program that solves eq. (1) by recursively iterating eq. (4). The program is parallel but there are no explicit call to communication functions:

```

00  #include "mdp.h"
01
02  void main(int argc, char** argv) {
03      mdp.open_wormholes(argc,argv); // open communications
04      int L[]={10,10,10};           // declare volume
05      mdp_lattice      space(3,L);   // declare lattice
06      mdp_site          x(space);    // declare site variable
07      mdp_matrix_field phi(space,2,2); // declare field of 2x2
08      mdp_matrix        A(2,2);      // declare matrix A
09      A(0,0)=1;  A(0,1)=I;
10      A(1,0)=3;  A(1,1)=1;
11
12      forallsites(x)                  // loop (in parallel)
13          phi(x)=0;                   // initialize the field
14      phi.update();                   // communicate!
15
16      for(int i=0; i<1000; i++) {     // iterate 1000 times

```

```

17         forallsites(x)                                // loop (in parallel)
18             phi(x)=(phi(x+0)+phi(x-0)+
19                 phi(x+1)+phi(x-1)+
20                 phi(x+2)+phi(x-2)-
21                 A*sin(2.0*Pi*x(1)/L[1]))/6; // equation
22         phi.update();                                // communicate!
23     }
24     phi.save("field_phi.mdp");                        // save field
25     mdp.close_wormholes();                            // close communications
26 }

```

Notes:

- Line 00 includes the MDP library.
- Lines 03 and 25 respectively open and close the communication channels over the parallel processes.
- Line 04 declares the size of the box $L = \{L_0, L_1, L_2\}$
- Line 05 declares a lattice, called `space`, 3-dimensional, on the box L . MDP supports up to 10-dimensional lattices. By default a lattice object is a mesh with torus topology. It is possible to specify an alternative topology, boundary conditions and any parallel partitioning for the lattice. Notice that each lattice object contains a parallel random generator.
- Line 06 declares a variable site, called `x`, that will be used to loop over lattice sites (in parallel).
- Line 07 declares a field of 2×2 matrices, called `phi`, over the lattice `space`. MDP is not limited to fields of matrices. It is easy to declare fields of any user-defined structure or class.
- Lines 08 through 10 define the matrix `A`.
- Lines 12 and 13 initialize the field `phi`. Notice that `phi` is distributed over the parallel processes and `forallsites` is a parallel loop.
- Line 14 performs communications so that each process becomes aware of changes in the field performed by other processes (*synchronization*).
- Lines 16 through 24 perform 1000 iterations to guarantee convergence. In real life applications one may want to implement some convergence criteria as stopping condition.
- Line 17 loops over all sites in parallel.
- Lines 18 through 21 implement eq. (4). Notice the similarity in notation. Here `phi(x)` is a 2×2 complex matrix
- Line 22 performs *synchronization*.
- Line 24 saves the field. Notice that any field, including the user defined ones, inherit methods `save` and `load` from a basic class `mdp_field`.
- It should also be noted that all MDP classes and functions are both type and exception safe. Moreover MDP components can be used without knowledge of C pointers and pointer arithmetics.

2 Linear Algebra

MDP includes a Linear Algebra package. The basic classes are:

- class `mdp_real`, that should be use in place of float or double.
- class `mdp_complex`, (just another implementation of complex numbers).
- class `mdp_array`, for vectors and/or multidimensional tensors.
- class `mdp_matrix`, for any kind of complex rectangular matrix.

The most notably difference between our linear algebra package and other existing packages is its natural syntax.

For example:

```
mdp_matrix A,B;
A=Random.SU(3);
B=exp(inv(A))*hermitian(A+5);
```

reads like

$$\begin{aligned} A \text{ and } B &\text{ are matrices} \\ A &\text{ is a random } SU(3) \text{ matrix} \\ B &= e^{(A^{-1})}(A + 5 \cdot \mathbf{1})^H \end{aligned} \tag{5}$$

Notice that each matrix can be resized at will and is resized automatically when a value is assigned.

3 Lattice, Site and Field

An `mdp_lattice` is a container for *topology* and *partitioning* information about the sites. In more abstract terms a lattice is any collection points (vertices) embedded in a multi-dimensional space and connected with directional links. The set of links determines the lattice topology and the boundary conditions. The term partitioning refers to the function that assigns each site (vertex) to one of the parallel process. A lattice, by default, is a mesh.

Each lattice is partitioned over the parallel processes at runtime. There is a default topology and default partitioning but it is possible pass any topology and partitioning functions to the `mdp_lattice` constructor.

A lattice also contains a parallel random number generator: each site of each lattice has its own independent random number generator.

On each lattice it is possible to allocate one or more fields. Some fields are built-in, for example: `mdp_complex_field`, `mdp_vector_field`, `mdp_matrix_field`, etc. All of them extend (inherit from) `mdp_field<mdp_complex>`.

Class `mdp_complex` can be used to declare any type of field. For example:

```
class W {
public: int w[10];
};
int L[]={30,30};
mdp_lattice plane(2,L);
mdp_field<W> psi(plane);
```

declares a 30×30 lattice (`plane`) and a field (`psi`), that lives on the `plane`. The field variables of `psi`, `psi(x)` assuming `x` is an `mdp_site` of `plane`, belong to class `W`.

Each user-defined field can be saved:

```
psi.save("filename");
```

loaded

```
psi.load("filename");
```

and synchronized

```
psi.update();
```

as any of the built-in fields.

4 Optimization Issues

Once an `mdp_lattice` object is declared the constructor of class `mdp_lattice` performs the following operations:

- Declares a parallel random number generator associated to each site (it uses the Marsaglia random number generator).
- Builds tables containing topology and partitioning information that will be used by the fields to optimize (minimize) communication. Basically each site determines which other sites are its neighbors and where they are located (on which parallel process). When a new field is created on the lattice the field will use these tables to create buffers for the communications.

Once a field object is declared the constructor of class `mdp_field` (or derived field) performs the following operations:

- Each process allocates memory to store the local sites (i.e. sites that will be managed by the process itself).
- Each process loops over every other process and determines if the other process allocated sites that are neighbors of the local sites (this information is already stored in the tables maintained by the lattice object). If this occurs the two processes are said to *overlap*: they have sites in common that need to be synchronized.
- Each process allocates buffers to store copies of the sites that are not local but are neighbors of the local ones and need to be synchronized with the overlapping processes (in this paper we are assuming only next-neighbor synchronization but actually MDP supports also extended synchronization such as next-to-next-neighbor and more). Buffers are created according with some conditions: sites synchronized with the same overlapping process are stored contiguously in memory so that communication can be performed in a single *send/receive*. These buffers are created independently by each field.

Notice that two processes may be overlapping in respect to a given lattice and not overlapping in respect to a different lattice in the same program.

Every time a field changes, for example in a parallel loop such as

```
forallsites(x) phi(x)=0;
```

the program notifies the field that its values have been changed by calling

```
phi.update();
```

The method `update` performs all required communication to copy site variables that need to be synchronized between each couple of overlapping processes. These communications are optimal in the sense that:

- Each process, at each one time, is involved only in one send and one receive.
- Two different processes communicate only if they are overlapping in respect of the lattice associated to the field.
- If two different processes are overlapping, they perform a single send/receive of all sites variables that are synchronized between the two.
- Only the sending process needs to create a temporary buffer. The receiving process receives the site variables in the same buffer where they are normally stored without reordering (and without need for a temporary buffer).

We will refer to our set of communication rules as a “communication policy” (it is possible, in principle, to change this policy to deal with non-standard network solutions). Although our communication policy does not overlap communication with computation it has the advantage of minimizing network jam and calls to send/receive. Hence this communication policy is almost insensitive to network latency and is dominated by network bandwidth. Benchmarks are application dependent since parallel efficiency is greatly affected by the lattice size, by the amount of computation performed per site, processor speed and type of interconnection. In many typical applications, like the one described in the preceding example, the drop in efficiency is less than 10% up to 8 nodes (processes) and less than 20% up to 32 (our tests are usually performed on a cluster of Pentium 4 PCs (2.2GHz) running Linux and connected by Myrinet).

5 Conclusions

MDP is a powerful and reliable tool for developing efficient parallel numerical applications. Even if, on the one side, MDP is still undergoing development, on the other side, all of the features here described are fully functional and have been tested in real-life applications. For example MDP constitutes the core of the **FermiQCD** project [2] developed by the University of Southampton (UK) and Fermilab (Department of Energy). **FermiQCD** is collection of parallel algorithms for Quantum Chromo Dynamics computations. The typical **FermiQCD** problem is equivalent to solving iteratively a system of stochastic differential equations

in a 4-dimensional space. Typical field variables are vectors of complex matrices. **FermiQCD** programs are used in production runs in parallel on 8 or more nodes.

We believe MDP could be a useful tool for scientists developing parallel numerical applications. MDP version 2.0 (current) is open source and is free for research and educational purposes.

Project web pages:

- <http://www.pheonixcollective.org/mdp/mdp.html> (license and source code)
- <http://www.fermiqcd.net> (Lattice QCD applications)

References

1. M. Di Pierro, “Matrix Distributed Processing: ...”, *Computer Physics Communications*, **141** (2001), pp. 98-148 [<http://xxx.lanl.gov/abs/hep-lat/0004007>]. *Note: this paper describes version 1.3 of MDP. The current version is 2.0*
2. M. Di Pierro, “**FermiQCD**”, *Nucl. Phys. Proc. Suppl.* **106** (2002) 1034-1036 [<http://xxx.lanl.gov/abs/hep-lat/0110116>]