SPI, mapping, and site ordering in Lattice QCD code on Mira

Heechang Na and James Osborn
Argonne Leadership Computing Facility, ANL

MiraCon, ALCF, ANL, IL

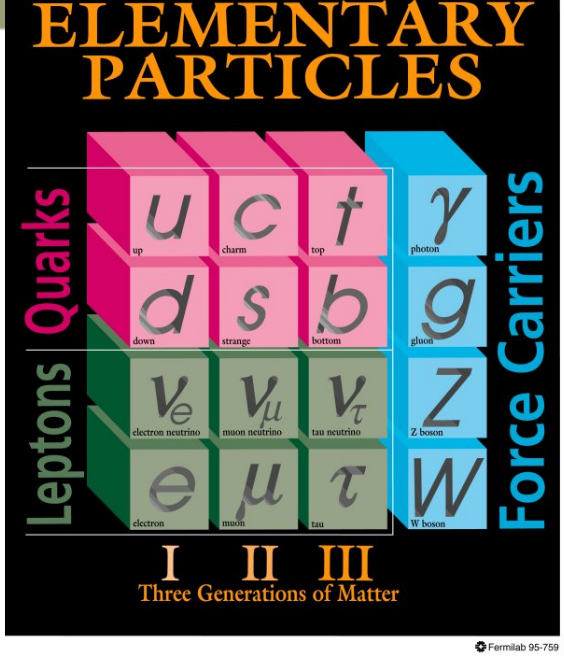
Mar.–7–2013

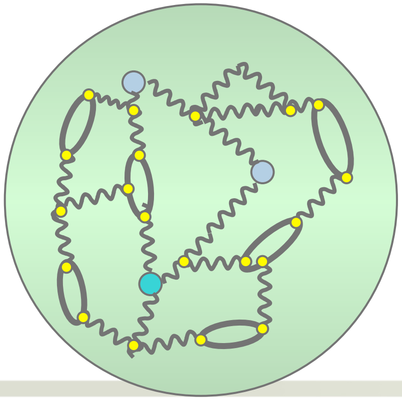
Outline

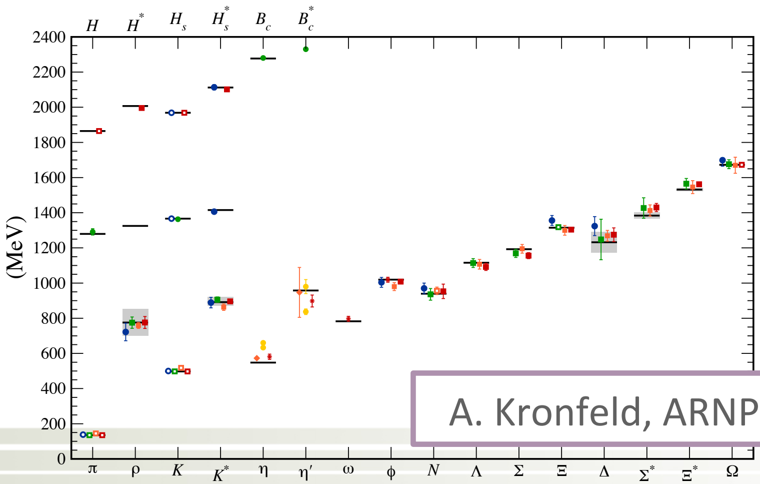
- Motivation
  - QCD and lattice QCD
  - Ising model and lattice QCD
- Communications with SPI
  - Resource structure
  - qspi
  - Benchmarks
- Mapping
- Site ordering
- Overall benchmarks

• Motivation: QCD and Lattice QCD

- QCD: Quantum Chromodynamics
- The Standard Model
  - Electro-magnetic force (QED)
  - Strong force (QCD)
  - Weak force
- Lattice QCD
  - Realization of QCD in a finite grid (lattice)
  - Provide numerical solutions for non-perturbative quantities.
- Applications on high energy particle physics (flavor physics, spectroscopy, beyond the Standard Model, etc) and nuclear physics (periodic table, quark-gluon plasma, equation of state of QCD, etc..)



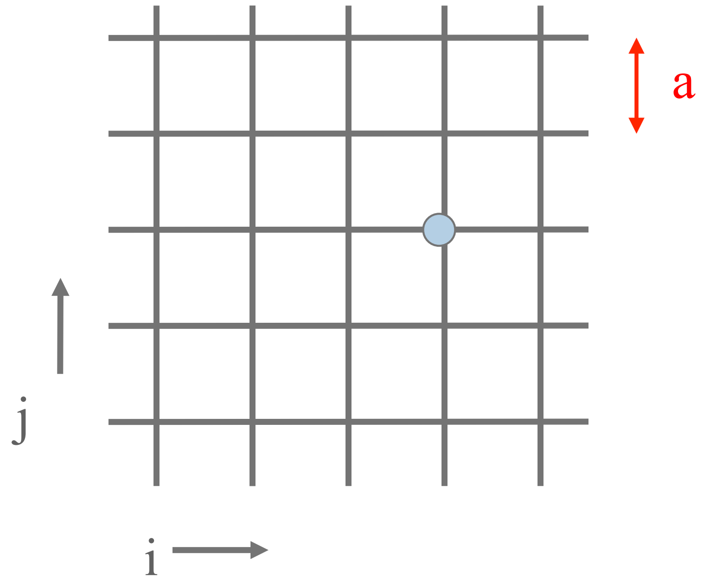




A. Kronfeld, ARNPS 64 (2012) 265

• Motivation: Ising model and Lattice QCD

- Ising model is a model for ferromagnetism.
- Local spin interaction model in quantum mechanics.
- Monte Carlo method can be used for numerical study.



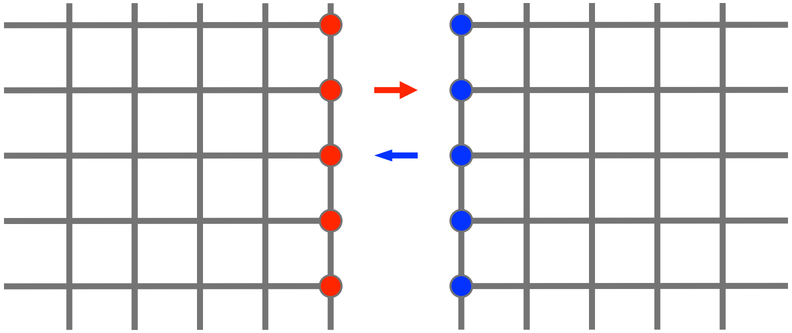
$$H = -\sum J_{ij}\sigma_i\sigma_j - \sum h_j\sigma_j$$

$$L = \bar{\psi}_i(i\gamma^\mu D_\mu - m)\psi_j - \frac{1}{4}G_{\mu\nu}G^{\mu\nu}$$

- Lattice QCD provides numerical solutions for QCD (Quantum Chromodynamics)
- QCD is a local Quantum field theory.
- Monte Carlo method is the main method of Lattice QCD.

• Motivation: Ising model and Lattice QCD

- Local interactions are important.
  - Communications between near neighbor are more important
  - All boundaries are sending messages at the same time
    -> Better latency -> MU SPI communication
  - Reducing the surface area and number of hops for each message
    -> mapping
  - Reducing the number of chunks of memory -> site ordering



- Ising model:
  - Two column vector at each site (up or down)
- Lattice QCD:
  - Bosonic field: 3x3 complex matrix times 4 directions
  - Fermionic field: 3x3x4x4 complex matrix times N flavors
    - Typical size of lattice: $96^3\times192$, typical size of the matrix: $(4\times10^9)^2$

• Communications with the Message Unit (MU) SPI

- Resource structure in a node:
  - 16+1 groups, 4 sub-groups per group: 64 sub-groups per node
  - 8 FIFOs per sub-groups: 32 FIFOs per group or 512 FIFOs per node
    - same numbers of injection FIFOs and reception FIFOs

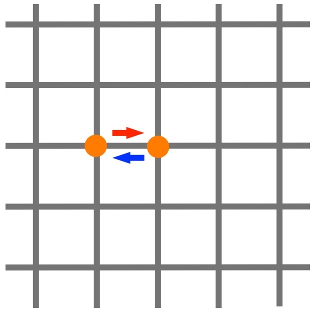
- 5D torus network: 10 optical cables
  - 10 FIFOs per process would be optimal.
  - For c64 mode, one process has 8 FIFOs.
    - And, have to think about MPI

- MU SPI provides:
  - Point-to-point network
  - Collective network
    - Memory fifo, direct put, and remote get for each network
  - We utilize the point-to-point direct put for our qspi communication library.

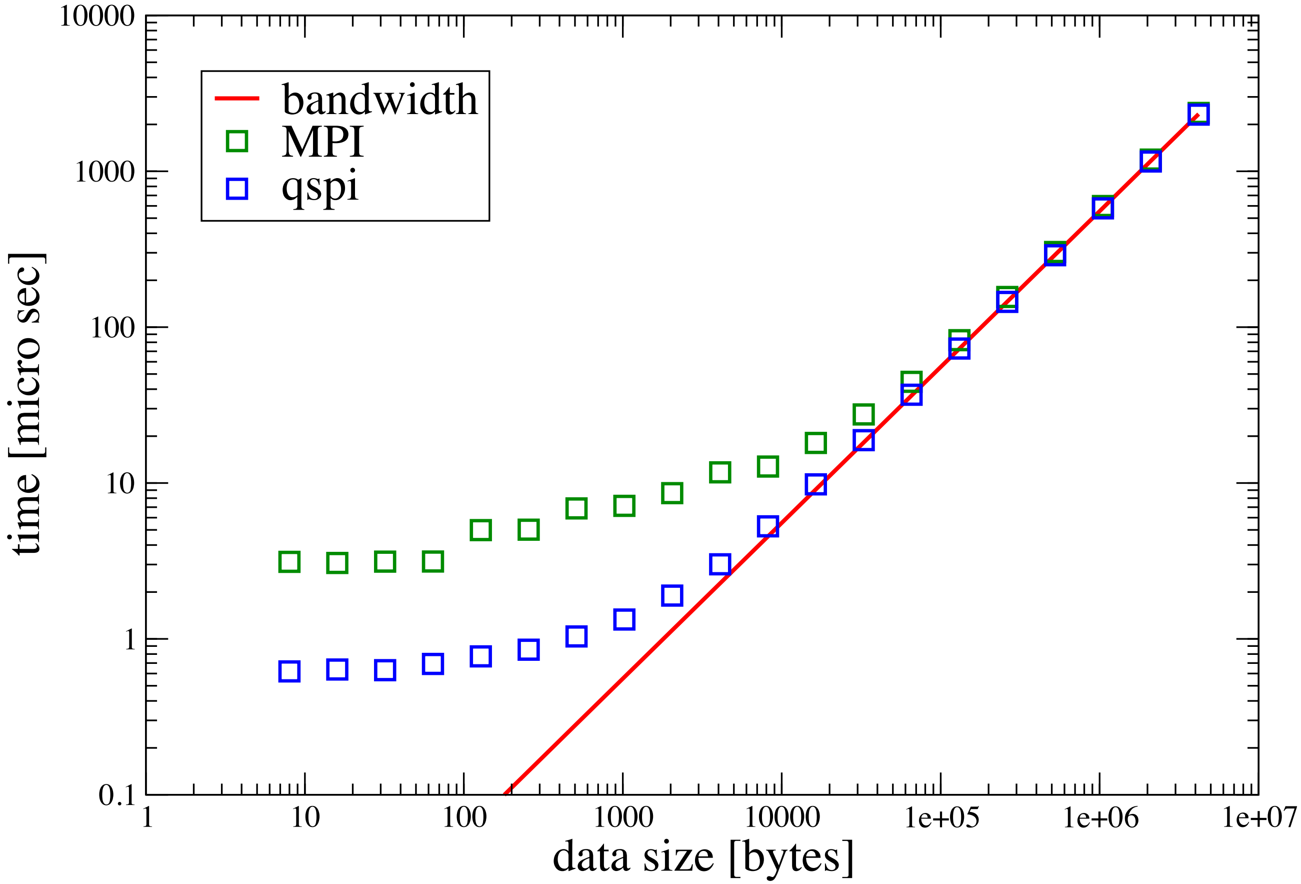
• qspi communication library

- void qspi_init(void);
  - allocate FIFOs and base address tables
- void qspi_set_send(int dest, void *buf, size_t size, qspi_msg_t send_msg);
- void qspi_set_recv(int src, void *buf, size_t size, qspi_msg_t recv_msg);
  - prepare the handle variables
- void qspi_prepare(qspi_msg_t msgs[], int num);
  - exchange the handles between senders and receivers (using MPI)
  - set the descriptors
- void qspi_start(qspi_msg_t msg);
  - inject the descriptors
- void qspi_wait(qspi_msg_t msg);
  - waiting until zero receive counter
- void qspi_finalize(void);

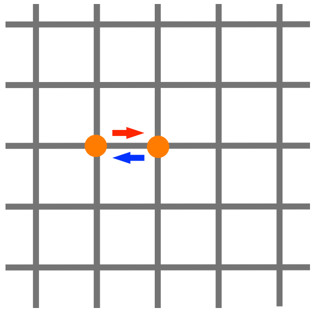
• Benchmarks between qspi and MPI



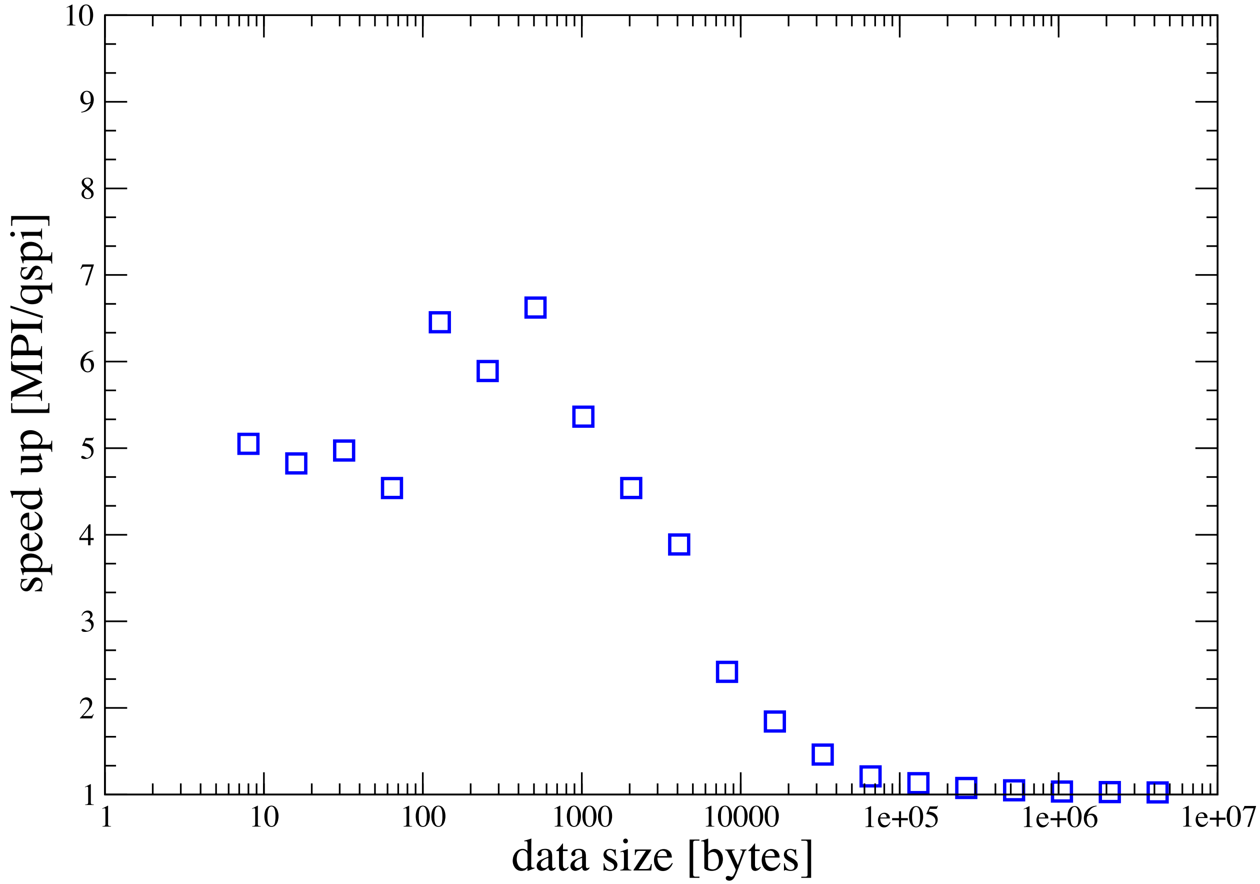
- Ping-pong between the nearest neighbors in c1 mode
    - Latency: qspi: 0.6 micro-sec,   MPI: 3 micro-sec



• Benchmarks between qspi and MPI

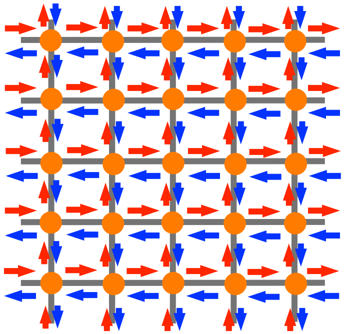


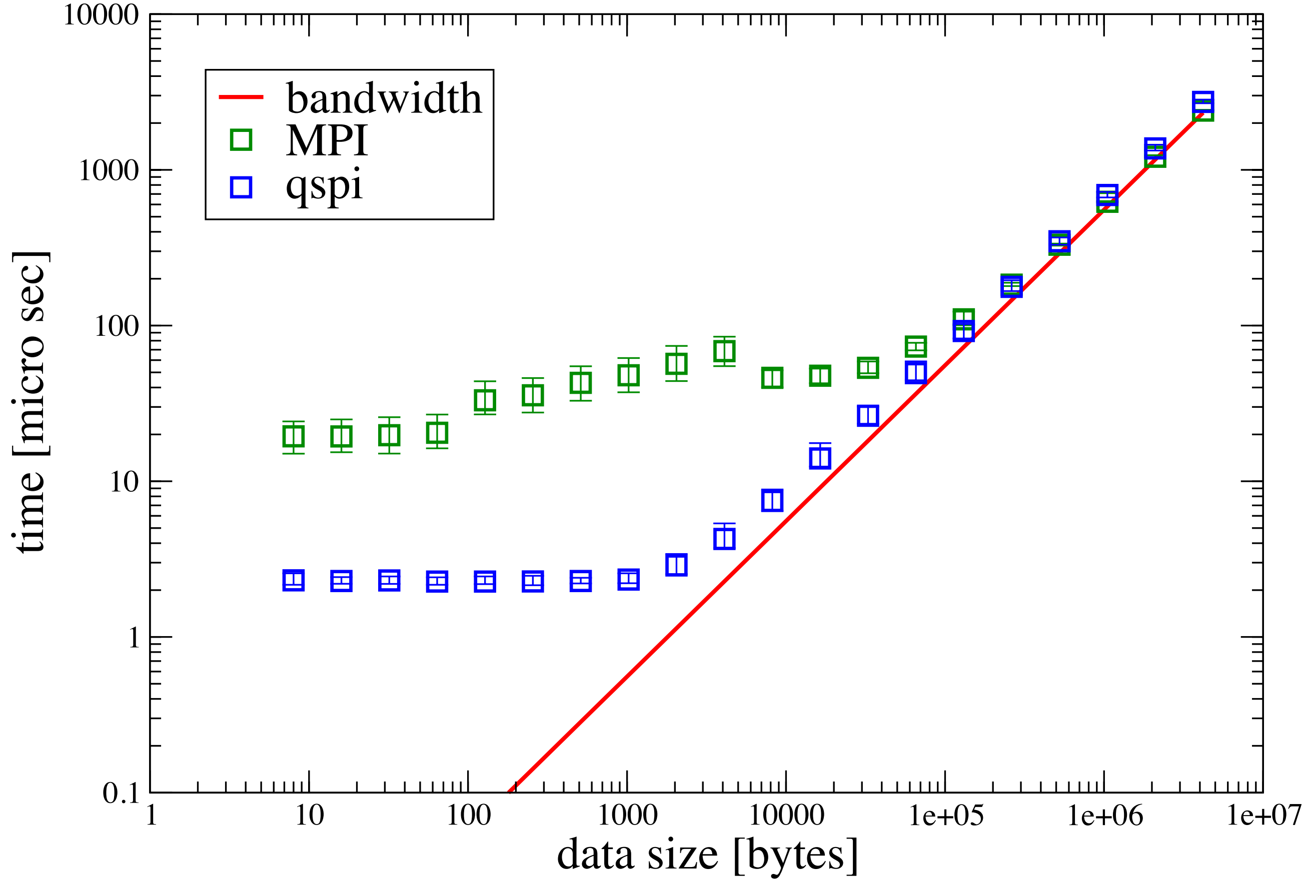
- Ping-pong between the nearest neighbors in c1 mode
  - Latency: qspi: 0.6 micro-sec, MPI: 3 micro-sec



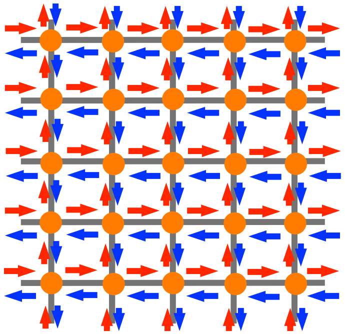
• Benchmarks between qspi and MPI

- 5D exchange in c1 mode
  - Latency: qspi: 2.3 micro-sec, MPI: 19 micro-sec

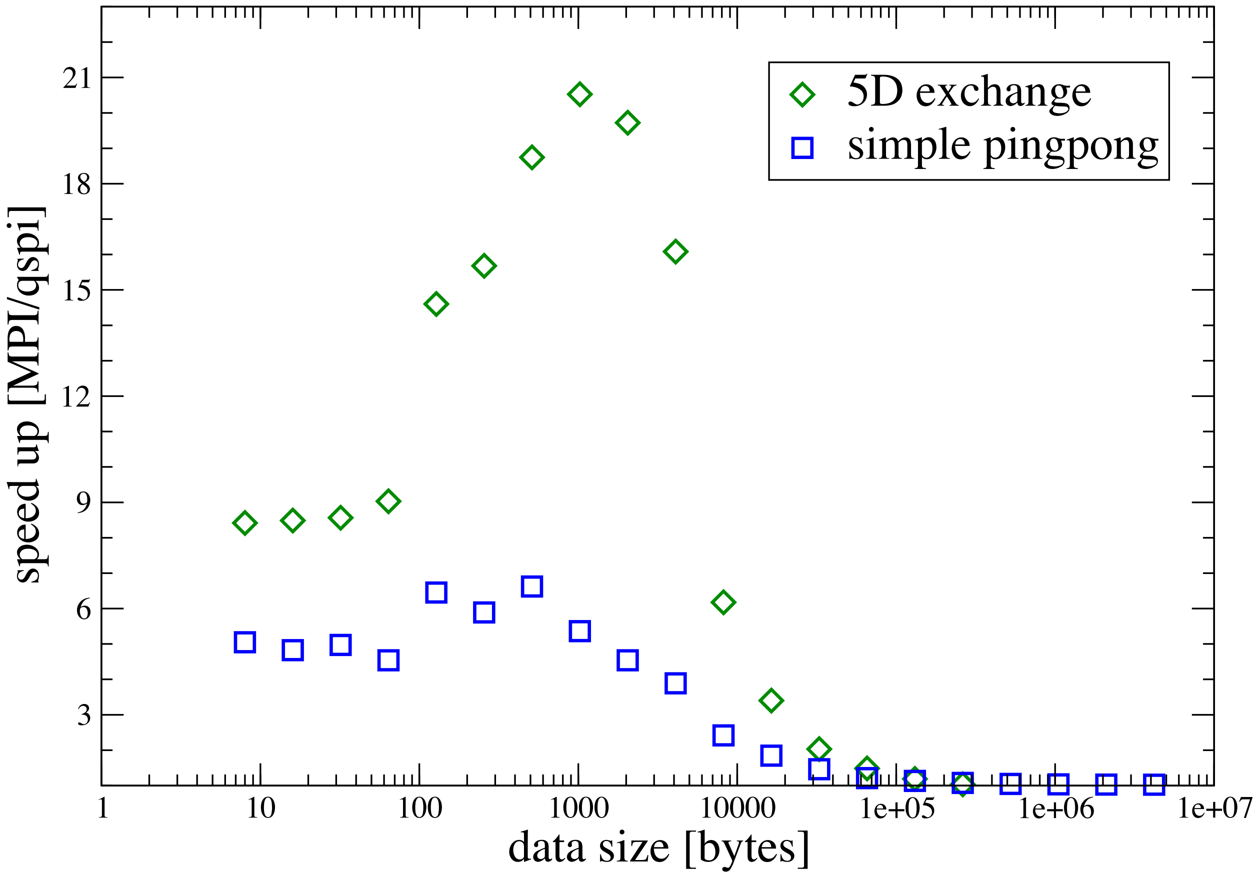




• Benchmarks between qspi and MPI

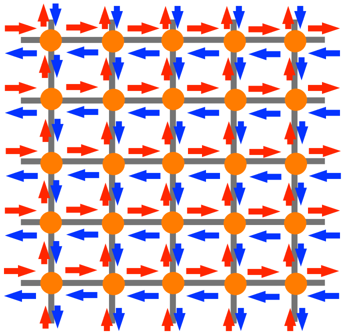


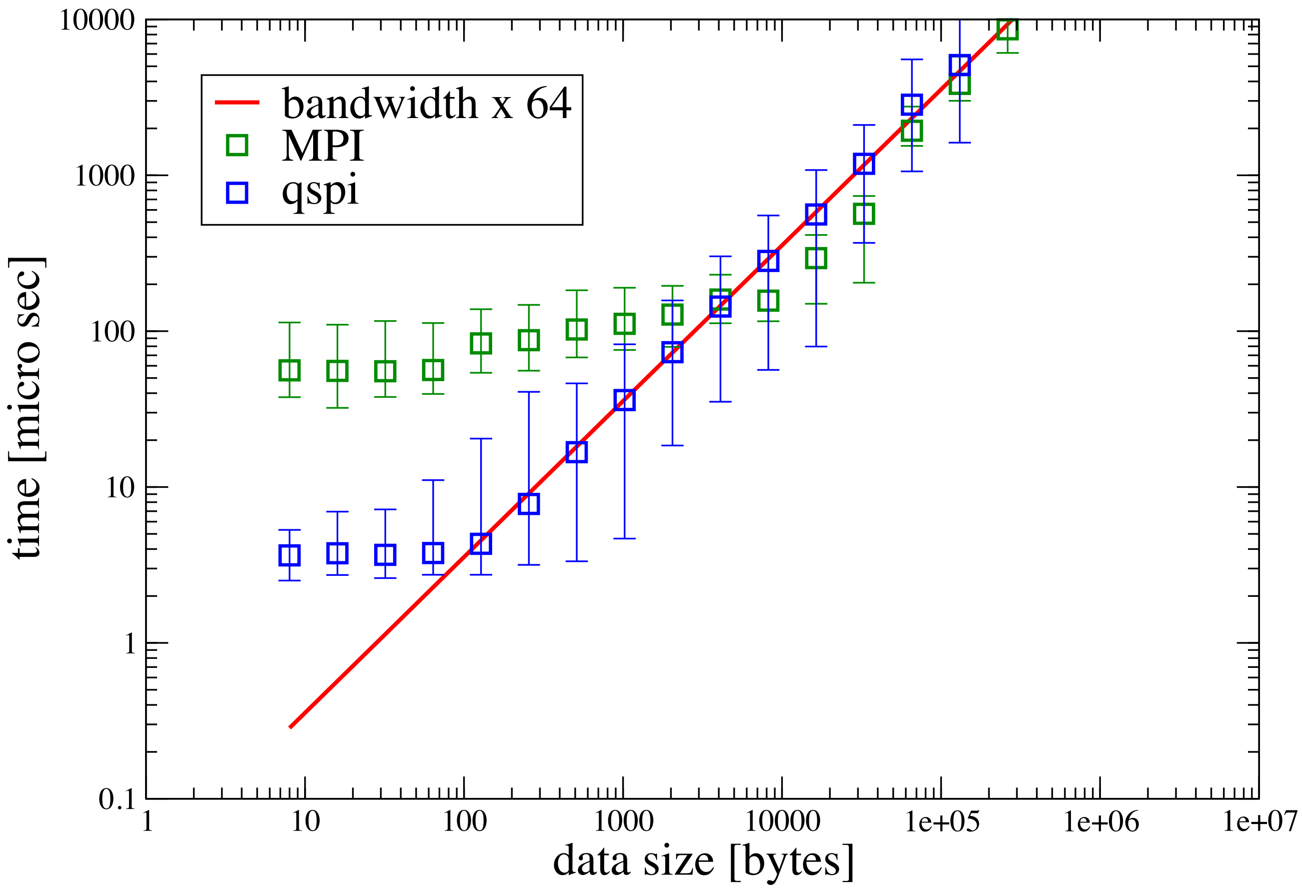
- 5D exchange in c1 mode
  - Latency: qspi: 2.3 micro-sec, MPI: 19 micro-sec



• Benchmarks between qspi and MPI

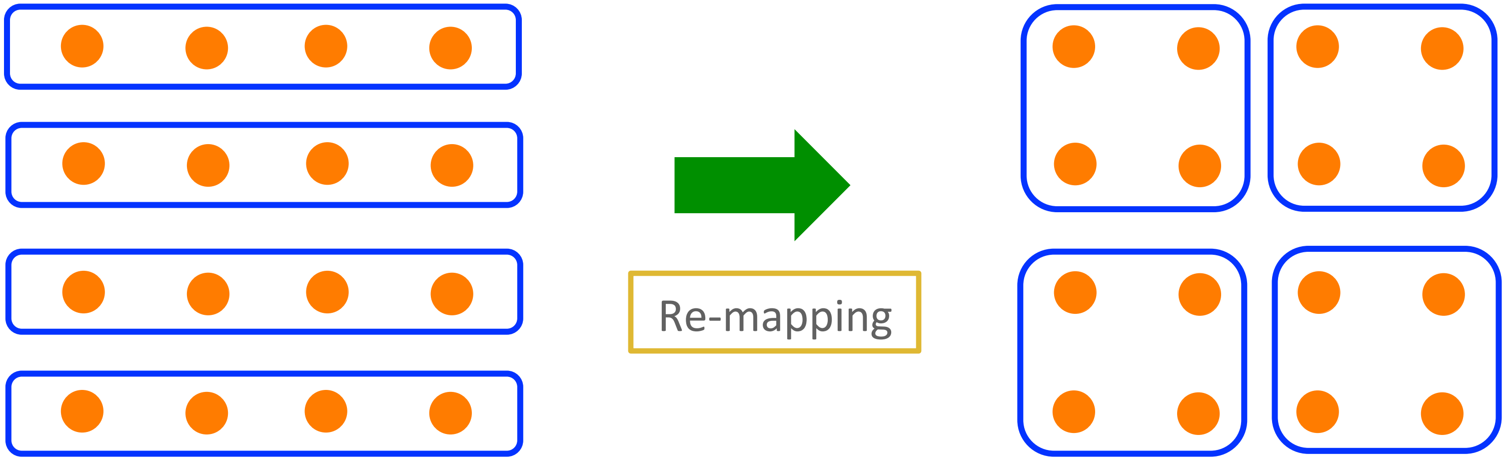
- 5D exchange in c64 mode
    - Latency: qspi: 3.6 micro-sec,   MPI: 55 micro-sec





• Mapping strategy for Lattice QCD

- Reducing the surface volume at the boundaries and the number of hops for each message are desirable.
- Proper mapping can reduce the surface volume and the number of hops.
  - For example, 2D mapping with 4 ranks/node:

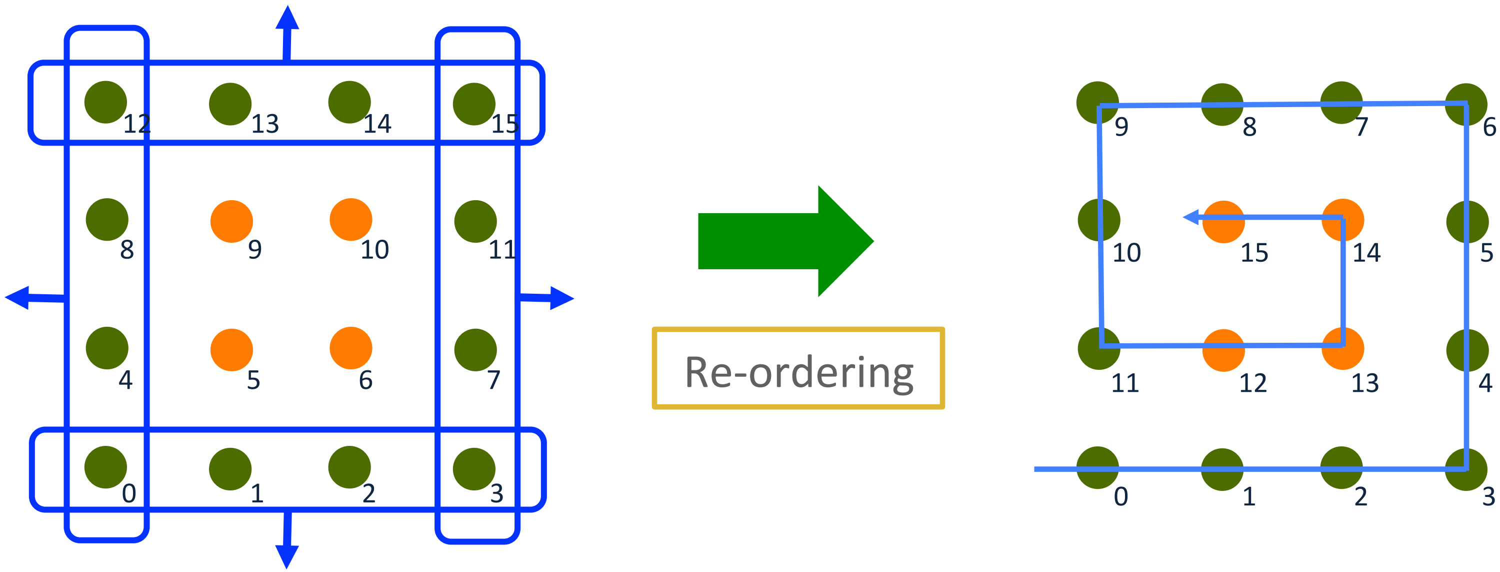


• Site ordering strategy for Lattice QCD

• Proper site ordering can reduce the number of fragmentations of sending messages.

• This is particularly important for the communications at the boundaries.

  • For example, site ordering of a 2D lattice:



• Overall banchmarks

- HISQ solver in the USQCD SciDAC modules
  - 12^4 volume/node on 128 nodes

qspi	Mapping	mode	Gflops/node
		c32	11.7
	✔	c64	12.9
✔		c32	13.8
✔	✔	c64	18.1