Optimizations of a spectral/finite difference gyrokinetic code for improved strong scaling toward million cores

Shinya Maeyama[1)], Tomohiko Watanabe[2)], Yasuhiro Idomura[1)], Motoki Nakata[1)], Masanori Nunami[2)], and Akihiro Ishizawa[2)]

[1)] Japan Atomic Energy Agency, Rokkasho, Japan; [2)] National Institute for Fusion Science, Toki, Japan

SC13
Nov. 17-22, 2013
Denver, Colorado

This work is supported by HPCI Strategic Programs Field No. 4 and G8 Research Council Initiative (NuFuSE).

Motivation

One of the key issues in development of magnetic fusion devices is plasma turbulence, which extends over temporal and spatial scales of electrons and ions. However, multi-scale plasma turbulence simulations based on 5D gyrokinetic model require huge computations.

In this work, we present optimization of inter-node communications and computation-communication overlaps on massively parallel platforms such as the K computer[1] and the BX900 (Nehalem-EP cluster), and report their impact on the strong scaling of the gyrokinetic Eulerian code GKV[2].

[1] http://www.aics.riken.jp/en/kcomputer/ [2] T.-H. Watanabe and H. Sugama, Nucl. Fusion 46, 24 (2006).

Model and parallelization of GKV

- Solve the time evolution of the turbulent electromagnetic fields and of the plasma distribution functions in 5D space (x, y, z, v, μ) with multi species of plasma (s).
- Multi-dimensional domain decomposition in x (or y), z, v, μ and s; hybrid parallelization with MPI/OpenMP.

→ Multi-dimensional CFD calculations consist of
- Spectral method in x, y (including data transpose in x, y for parallel FFTs)
- Finite difference in z, v, μ (including point-to-point communications in z, v, μ for stencil calculations)
- Field solver (including reduction communications over v, μ, s)

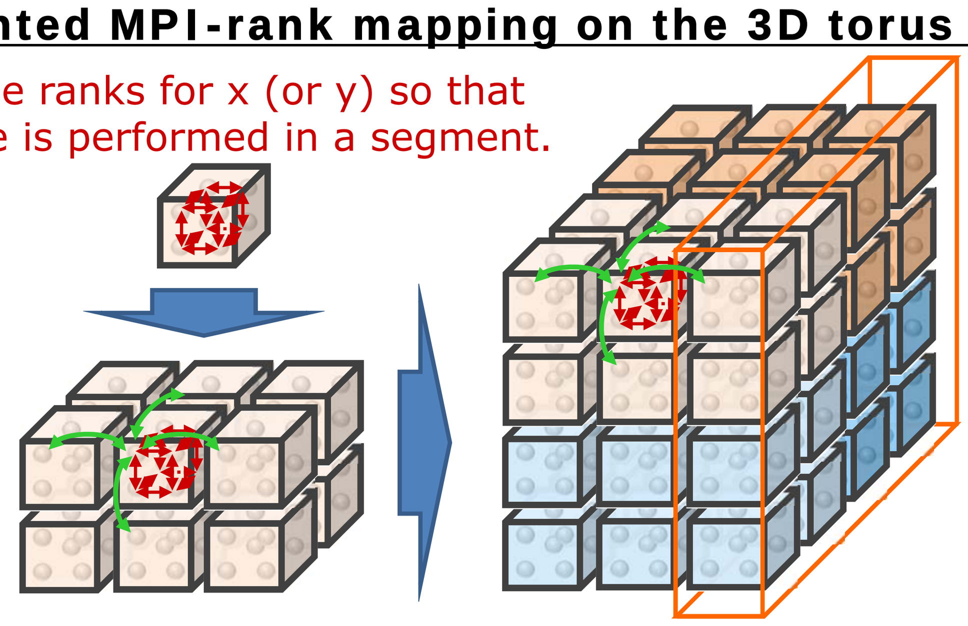
Optimization of inter-node communications

Torus fusion (Tofu) Inter-connect on the K computer

- 3D torus network as a user view
- Multi-way simultaneous communications (4 send + 4 receive)
- Highly-optimized collective communication algorithm[3]

Segmented MPI-rank mapping on the 3D torus network

① Arrange ranks for x (or y) so that transpose is performed in a segment.



③ Arrange rank for s so that reduction is performed through a cross section.

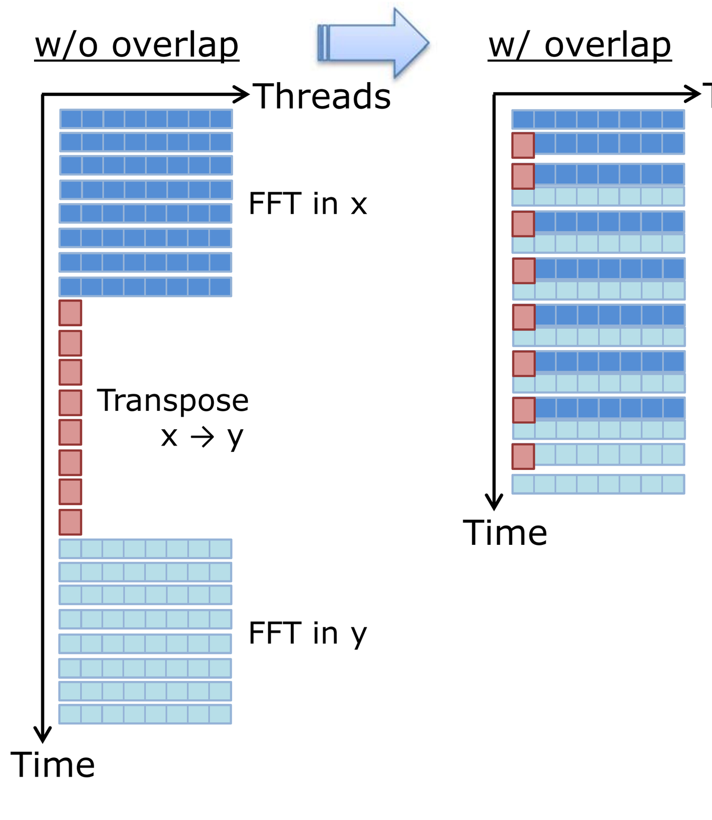
② Arrange ranks for z, v, μ so that point-to-point communications are performed between adjacent segments.

- Fully utilize the bi-section bandwidth and minimize cost of transpose.
- Reduce collisions of the messages of simultaneous point-to-point communications by using multi-way channels.

[3] T. Adachi, et al., Comput. Sci. Res. Dev. 28, 147 (2013).

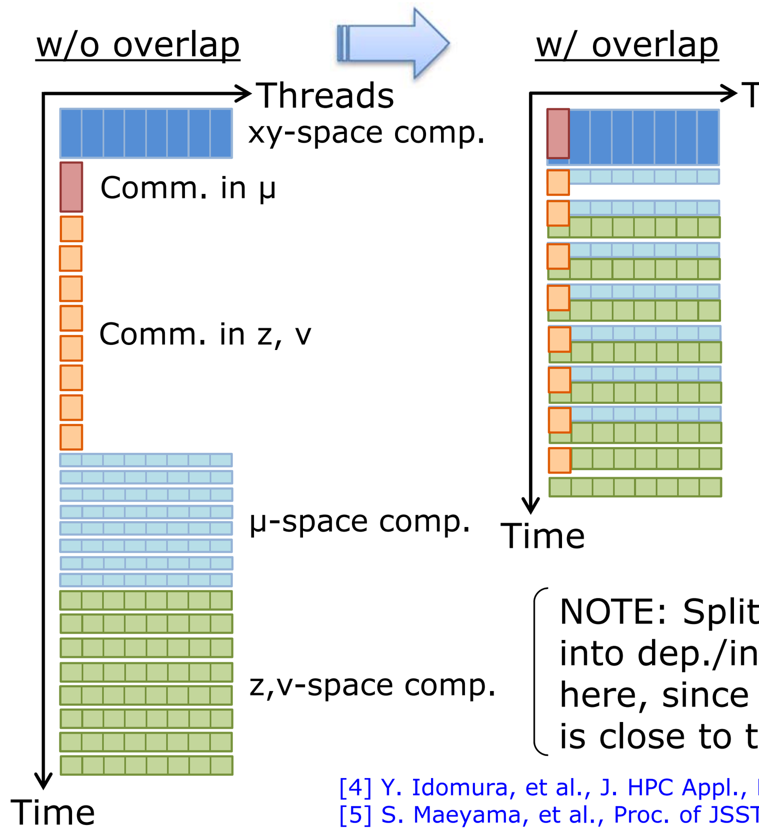
Computation-communication overlaps

Pipelined overlap of spectral calculation



① Decompose nonlinear xy-spectral calculations (buffer copy, data transpose, FFTs, pseud-spectral computation) in the independent loop (μ).
② Overlap the data transpose and the other computations with different μ-indexes using a communication thread implemented on the MASTER thread[4,5].

Pipelined overlap of finite difference calculation



① Comm. in μ is overlapped with the xy-space comp., which is independent of comm. data.
② Overlaps of comm. in z and v are implemented by pipelining in the μ loop.

NOTE: Split of the stencil calculation into dep./indep. parts[4] is inefficient here, since the employed array size is close to the buffer size.

[4] Y. Idomura, et al., J. HPC Appl., DOI:10.1177/1094342013490973 (2013).
[5] S. Maeyama, et al., Proc. of JSST 2012, Kobe, Japan, 252 (2012).

Summary

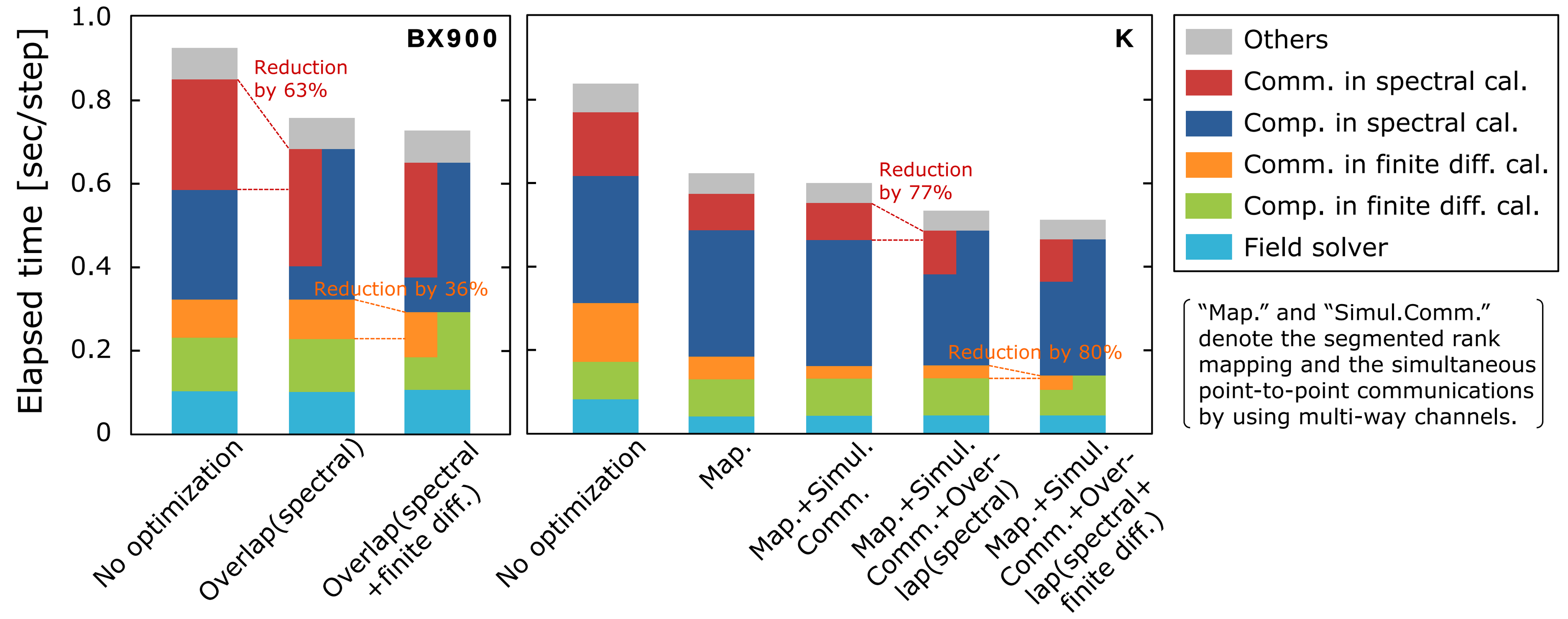
The present optimizations improve scaling of GKV and enable multi-scale plasma turbulence simulations.
- Segmented rank mapping utilizes the bi-section bandwidth for data transpose and reduces collisions of the messages for simultaneous point-to-point communications, and therefore, saves communication cost.
- Pipelined overlaps of computations and communications efficiently mask the communication cost.
- Thanks to the above optimizations, GKV shows excellent strong scaling up to ~600k cores and achieves high parallel efficiency ~99.99994% on the K computer.

Effects of the optimizations

System: BX900 (2.93 GHz, 11.72 GFlops/core, Memory BW 6.4 GB/s/cores, 4 cores/node, InfinibandQDR with Fat Tree, Interconnect BW 4 GB/s × 2, 2134 nodes; LinPack 191.4 Tflops), K (2 GHz, 16 GFlops/core, Memory BW 8 GB/s/core, 8 cores/node, Tofu interconnect with 6D mesh/torus, Interconnect BW 5 GB/s × 4, 88128 nodes; LinPack 10.51 PFlops).

Cost analysis and cross-platform versatility

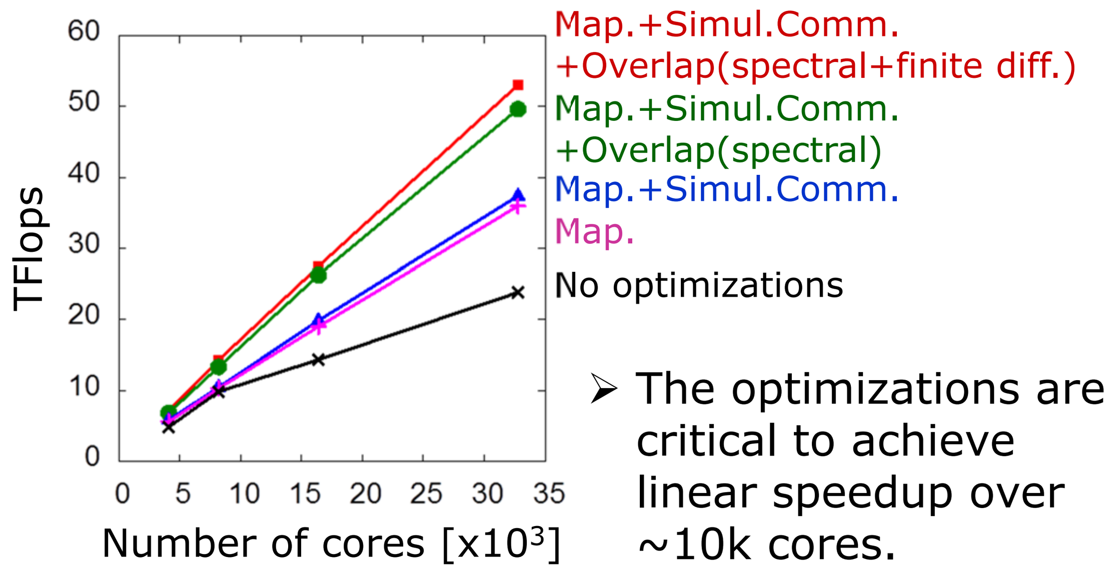
Problem size: $(n_x, n_y, n_z, n_v, n_\mu, n_s)$ = (256, 256, 32, 32, 32, 2)
Parallelization: $(N_{xy}, N_z, N_v, N_\mu, N_s, N_{threads})$ = (4, 4, 4, 8, 2, 4) on BX900, (4, 4, 4, 4, 2, 8) on K



- The results prove the versatility of the pipelined overlaps.
- The pipelined overlaps ideally reduce the communication overhead to ~$1/N_{threads}$.
- The segmented rank mapping on the torus network reduces communication cost.

Impact on the strong scaling

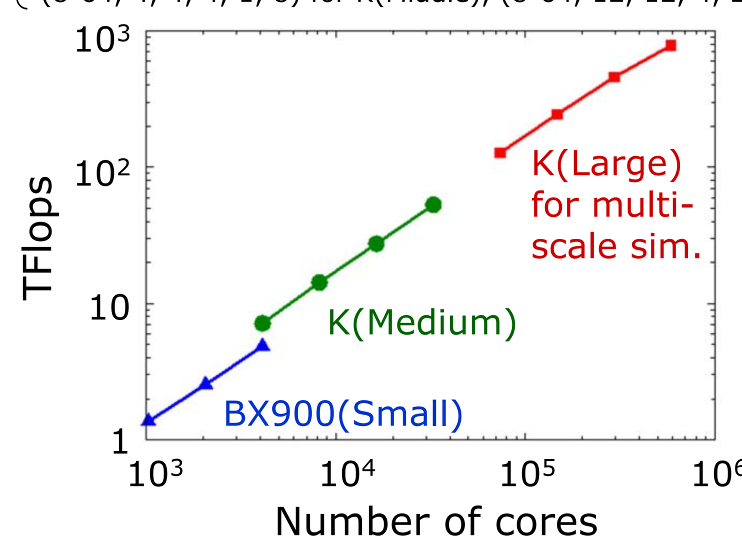
Problem size: $(n_x, n_y, n_z, n_v, n_\mu, n_s)$ = (1024, 1024, 32, 32, 32, 1)
Parallelization: $(N_{xy}, N_z, N_v, N_\mu, N_s, N_{threads})$ = (8-64, 4, 4, 4, 1, 8) on K



- The optimizations are critical to achieve linear speedup over ~10k cores.

Strong scaling for multi-scale simulations

Problem size: $(n_x, n_y, n_z, n_v, n_\mu, n_s)$ = (256, 256, 32, 32, 32, 2) for BX900(Small), (1024, 1024, 32, 32, 32, 1) for K(Middle), (1024, 1024, 96, 96, 32, 2) for K(Large)
Parallelization: $(N_{xy}, N_z, N_v, N_\mu, N_s, N_{threads})$ = (2-4, 4, 4, 4-8, 2, 4) for BX900(Small), (8-64, 4, 4, 4, 1, 8) for K(Middle), (8-64, 12, 12, 4, 2, 8) for K(Large)

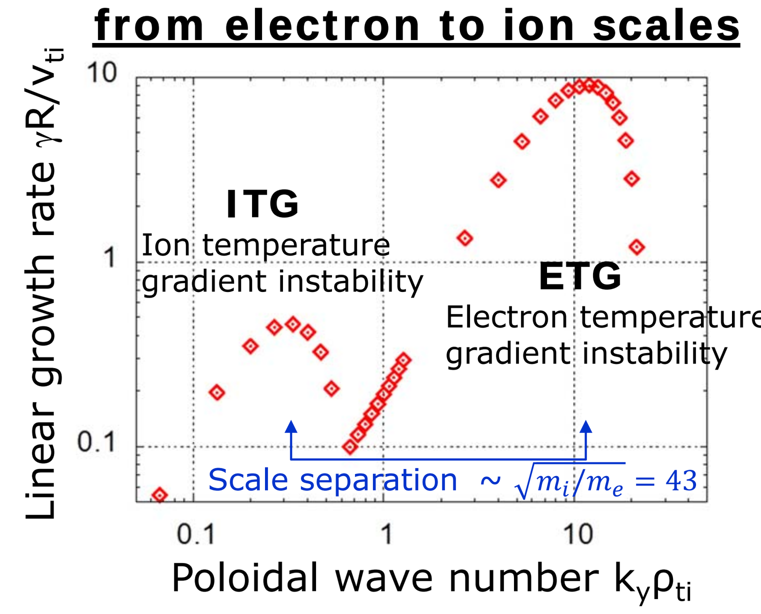


- GKV shows good scaling up to ~600k cores and achieves high parallel efficiency (~99.99994% estimated from the Amdahl's law), where Flops/Peak is 8.3%-10.8%.

Multi-scale plasma turbulence simulations

The optimizations strongly accelerate the plasma turbulence simulations covering scales of electrons and ions.
- Multi-scale plasma turbulence simulation using the real electron-to-ion mass ratio succeeds for the first time.
- High-growth-rate electron turbulence is suppressed by ion-scale turbulence which dominates heat transport.

Linear instabilities from electron to ion scales



Evolution of the electrostatic potential in multi-scale turbulence

ETGs rapidly grow and saturate. → ITGs slowly grow up. → ITGs saturate and suppress ETGs.