jh240058 / jh250041

Study on the Real Effect of Non-Blocking Collective Communications

Takeshi Nanri (Kyushu U.), Kengo Nakajima (U. Tokyo), Richard Vuduc (Georgia Tech.), Takeshi Fukaya (Hokkaido U.), Hiroyuki Takizawa (Tohoku U.), Osamu Tatebe (U. Tsukuba), Daisuke Takahashi (U. Tsukuba), Toshihiro Hanawa (U. Tokyo), Shinji Sumimoto (U. Tokyo), Maddegedara Lalith (U. Tokyo), Rio Yokota (Tokyo Tech.), Takahiro Katagiri (Nagoya U.), Keiichiro Fukazawa (Kyoto U.), Susumu Date (Osaka U.), Takashi Soga (Osaka U.), Yoshiyuki Morie (Teikyo U.), Richard Graham (NVIDIA), Martin Schulz (TUM), Bengisu Elis (TU Munich), Dennis Herr (TUM), Hari Subramoni (Ohio State U.), Aamir Shafi (Ohio State U.), Kaushik Kandadi suresh (Ohio State U.), Nathaniel Shineman (Ohio State U.), Benjamin Michalowicz (Ohio State U.), Tu Tran (Ohio State U.), Shulei Xu (Ohio State U.), Bharath Ramesh (Ohio State U.), Felix Wolf (TU Darmstadt), Gerhard Wellein (NHR), Gerardo Cisneros-Stoianowski (NVIDIA), Brody Williams (NVIDIA), Yong Qin (NVIDIA), Fabian Czappa (TU Darmstadt), Ayesha Afzal (NHR), Takeo Narumi (Kyushu U.), Shoji Sakoda (Kyoto U.), Daichi Mukunoki (Nagoya U.), Remma Arisako (Kyushu U.)

Motivation

- Collective communication is the significant causes of scalability degradation in HPC.
- NBC (Non-Blocking Collective communication) is expected to be a means to overlap this collective communication with computation and hide the communication time, but its use is currently limited to a small number of applications.
- This project provides programmers with correct knowledge about the usage and performance characteristics of NBC and the effect of communication hiding in real applications.

Lessons learned in FY2024

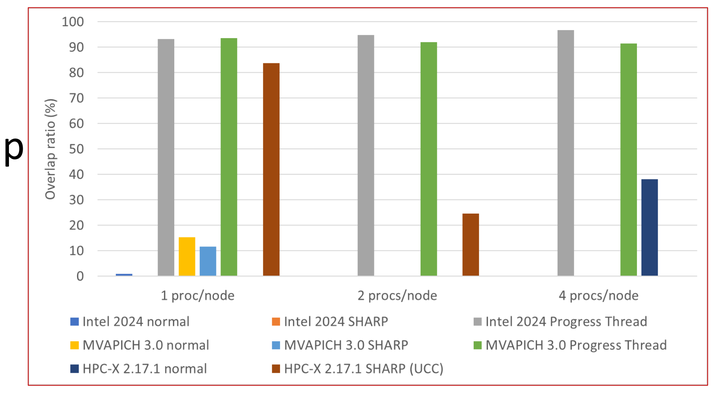
Topic 1: Available progress methods for NBC on each system

- Current availability of progress methods on JHPCN systems
  - SHARP required drivers and libraries to be matched
  - Tofu Barrier is not used in MPI NBC functions
  - Progress threads are available with MVAPICH and Fujitsu MPI
  - Currently, effect of overlapping is limited in most of the cases

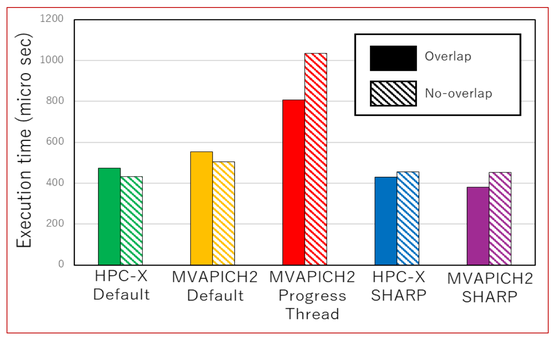
Method	AOBA-S	Wisteria-Oddysey	TSUBAME 4.0	Flow I	Flow II	Camphor 3	SQUID	GENKAI
SHARP	OK	-	OK	-	-	OK	OK	OK
Tofu Barrier	-	-	-	-	-	-	-	-
Assistant core	-	OK	-	OK	-	-	-	-
Progress thread	OK	OK	OK	OK	OK	OK	OK	OK

Topic 2: Trends of the effect of overlapping by NBC

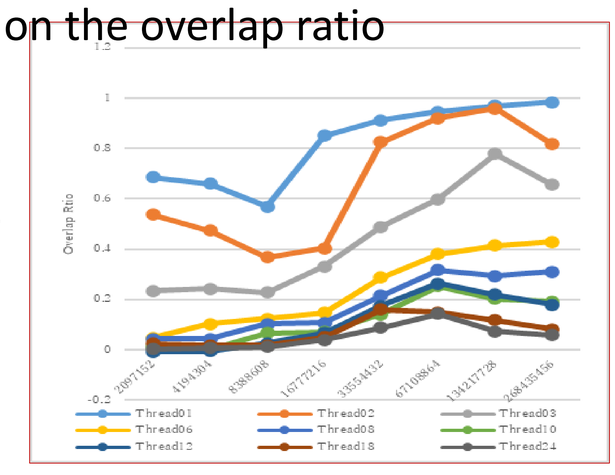
- Results with existing NBC benchmarks (OSU Micro Benchmarks, Intel MPI Benchmarks)
  - SHARP shows low overlap ratio with multiple procs/node
  - Progress thread shows high overlap ratio



- Results with our new benchmark NBC-Eff
  - Fixed computation amount to enable comparison among different progress methods
  - Progress thread causes low speed of fundamental communication

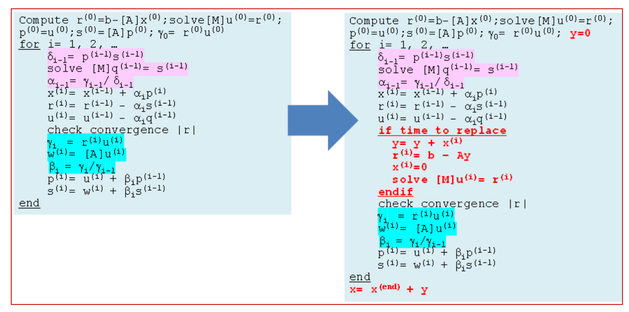


- Effect of memory bandwidth on the overlap ratio
  - Overlap ratio decreases as the number of compute threads increases.
  - Caused by the upper limit of memory bandwidth by overlapping computation and communication



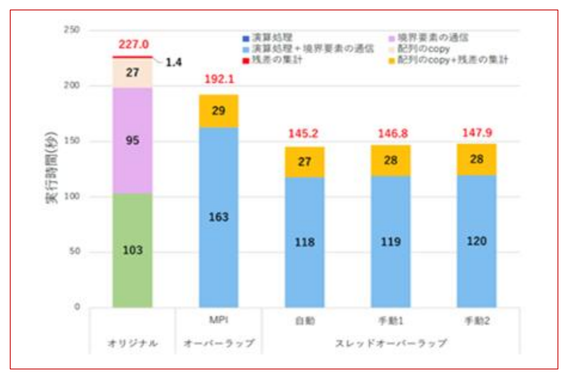
Topic 3: Investigation of communication hiding algorithms with NBC

- Pipelined CG
  - Need to prevent propagation of rounding errors
  - Explore two stabilization techniques:
    - Residual Replacement (RR)
    - Iterative Refinement (IR)
  - IR-RR enables low-precision robust convergence



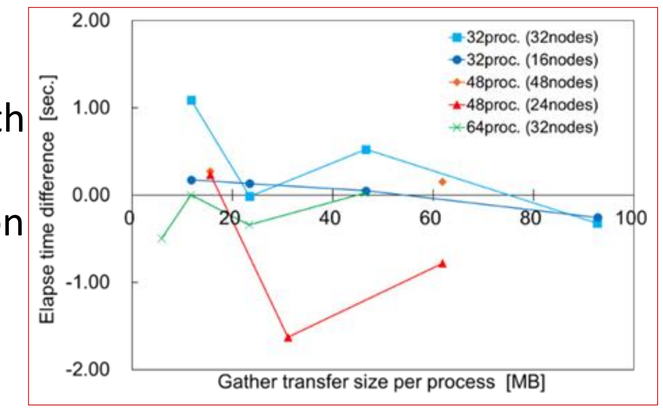
Residual-Refinement (RR) for Pipeline CG

- Jacobi method
  - Applied two communication overlapping methods (thread-based and MPI_Isend/Irecv) for halo-exchange
  - Thread-based overlapping showed higher performance



Comparison of overlapping methods between MPI_Isend/Irecv and thread

- MHD (Magneto-Hydro-Dynamics)
  - Overlap MPI_Igatherv with MPI_File_write
  - Showed overlap effects with small communication sizes
    - Opposite from anticipation (to be studied in future)



- FFT
  - Communication hiding effect of three-dimensional FFT with MPI_Ialltoall was confirmed.
  - Progress thread was used as the progress method
  - Overlapped version showed 10% higher performance than original on 32 nodes of Genkai A
    - Original: 8 MPI procs/node, 15 compute threads/proc
    - Overlapped: 8 MPI procs/node, 14 compute threads/proc
- WaitIO
  - Conducted an initial evaluation of communication performance with blocking collectives in WaitIO
  - Hierarchical MPI_Allreduce that combined WaitIO with MPI showed x27.7 speedup compared to the standalone MPI_Allreduce of WaitIO on Flow I

Plans for FY2025

- Topic 1: Study on the effective usage of NBCs
  - Examine hierarchical algorithm with offloading techniques to enable communication hiding with larger PPNs
  - Effective usage of NBCs on GPU/vector clusters
- Topic 2: Study trends of communication hiding by NBCs
  - Enhance our benchmark program NBC-Eff for GPU/vector clusters and persistent collective communications
- Topic 3: Investigation of communication hiding algorithms with NBC
  - Continue examining the effect of communication hiding in applications including:
    Kryrlov solvers, Generalized CP decomposition, FFT, CFD