Fully non-blocking communication-computation overlap using assistant cores towards Exa-scale computing

SC17, Nov. 12-17th 2017, Denver

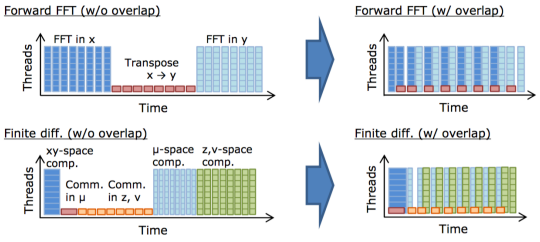
Motoki Nakata[1], Masanori Nunami[1], Shinsuke Satake[1], Yoshihiro Kasai[2], Shinya Maeyama[3], Tomo-Hiko Watanabe[3], and Yasuhiro Idomura[4]
[1]National Institute for Fusion Science, [2]Fujitsu Limited, [3]Nagoya University, [4]Japan Atomic Energy Agency

Motivation: Issues in Comm.-Comp. overlap in Many-core system

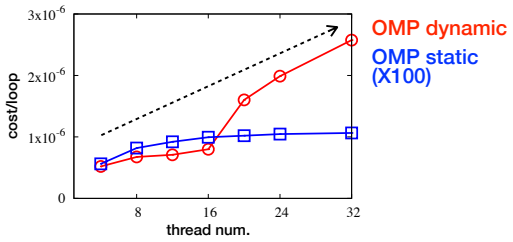
- **Pipelined overlap technique[1,2] using OpenMP** (Comm: Master thread + Comp: other threads with dynamic scheduling) has been applied to spectral (FFT) and finite difference(FD) schemes in 5D gyrokinetic Eulerian code GKV, which provides plasma simulations for the future fusion reactors.

—> Strong scaling over 600k cores with 99.99994% efficiency [2] on K computer (8cores/node)[3].

Pipelined Overlap with OpenMP (dynamic)



Overhead costs of scheduling in OpenMP (FX100)



- However, in Many-core system, the **dynamic scheduling overhead in thread parallelizations** are crucial [4], and Comm.-Comp. overlap performance will degrade.

—> In this work, the above issue is exemplified on Fujitsu FX100 (32 cores + 2 Assist. cores /node), and a solution with optimized overlap technique is proposed.

Non-blocking comm. by Assistant cores on SMaC architecture

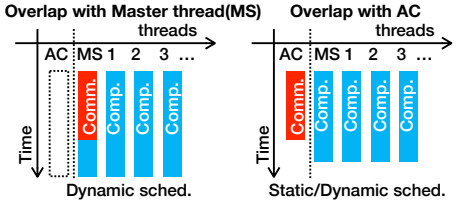
- In Japanese FLAGSHP2020 project, **Post-K computer** is being developed towards Exa-scale computing, and Scalable Many Core (SMaC) architecture [5] with "**Assistant cores (AC)**" will be applied.

- **Assistant cores (AC)** is useful not only for reducing OS Jitter, but also for improved performance[6]:

—> Optimized Comm.-Comp. overlap without Master thread comm.

—> Fully non-blocking ISend/IRecv, IAllreduce, IAlltoAll. (for PIC, FD, and Spectral codes)

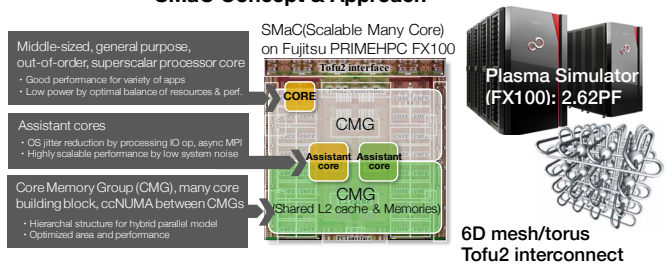
—> Static or chunk-size-optimized Dynamic scheduling (More efficient calc. to mask overheads)



Comm. performance of MS and AC

(8MPI proc)	ISend/IRecv	IAll_reduc	IAll_to_All
MS	0.12305s	0.70994s	0.70609s
AC	0.12304s	0.67314s	0.68144s

SMaC Concept & Approach



References
[1] Idomura et al., J. HPC Appl.(2013), [2] Maeyama et al., SC13 & Parallel Comp.(2015), [3] http://www.aics.riken.jp/en/k-computer/about/, [4] Ouermi et al., PEARC17(2017), [5] Shimizu (Fujitsu Ltd.), SC15 / SC16, [6] Similar to redundant core in Blue Gene/Q, [7] Nakata et al., Phys. Rev. Lett. (2017)

Summary

- A fully non-blocking optimized Communication-Computation overlap technique is presented:
(1) Assistant cores (AC) have no degradations of the comm. performance in comparison to that on the master thread, and enable us to use fully non-blocking MPI communications in the code.
(2) In Comm.-Comp. overlap by AC, all the OpenMP threads concentrate their computations, and the usage of Static scheduling leads to ~22.5% improvement of the numerical performance.
(3) Chunk-size optimization on AC enables the flexible use of Static/Dynamic OMP in Many-core system.

Effects of the optimizations with Assistant cores

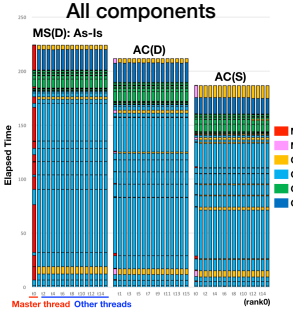
compile options
FC = mpifrtpx, FFLAGS= -Kfast,parallel,ocl
FFLAGS_OMP = -Kopenmp LIB = -SSl2, -lfftw3 -lm

- The numerical performance is compared among 3 types of the overlap:
MS(D): Master thread comm. with Dynamic sched.(chunk size=1, As-Is),
AC(D): Assistant core comm. with Dynamic sched. (chunk size=1),
AC(S): Assistant core comm. with Static sched.

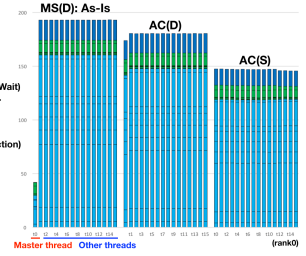
~22.5% improvement of the performance

Performance of GKV kernel on FX100: 4x12x18nodes
(~10 billion grid points, 432 MPI proc. with 16SMP)

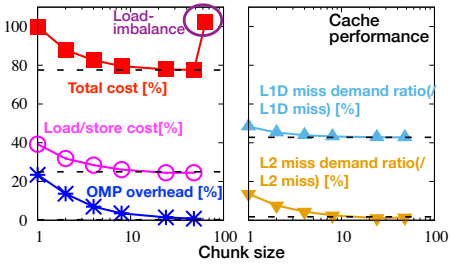
All components



Extracting Comp. part



Chunk size dependence in OMP dynamic
Dashed lines: Static case (equiv. to chunk size 48)



(i) Comm. on MS is successfully masked by AC: **~16/15 = 6.25% improve. of computation.**

(ii) Static scheduling leads to not only reduction in OMP overhead*, but also improved load/store and cash performance(incl. prefetch) with the continuous memory access for larger chunk size: **>15% improve.**

*OMP overhead is evaluated by the difference between OMP loop and non-OMP loop with no computations.
*Improved performance is similar for the other MPI ranks.

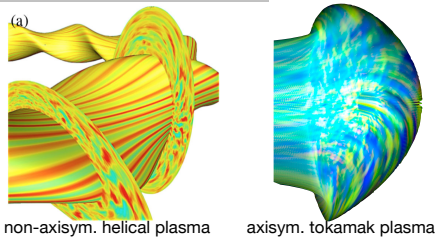
~22.5% improve. in total

—> When the chunk size is optimized in AC(D) [around the value in Static], the comparable performance to AC(S) can be achieved: enabling the flexible use of AC(S) and AC(D).

Application: 5D gyrokinetic plasma turbulence simulations

- By applying the Comm.-Comp. overlap, we realized a 5D turbulence simulations for complex geometry plasma with both deuterium ions and electrons, for the first time [7].

—> High performance computing significantly accelerates plasma and fusion plasma research.



non-axisym. helical plasma

axisym. tokamak plasma

Acknowledgment: this work is supported by NIFS-Fujitsu collaborations on Plasma Simulator (FX100), and by the MEXT Grant for Post-K project: Development of innovatives clean energy, Core design of fusion reactor(6-D).