TOHOKU UNIVERSITY
Cyberscience Center

Orchestrating a brighter world NEC

Performance Evaluation of Quantum ESPRESSO on SX-ACE

REV-A Workshop held on conjunction with the IEEE Cluster 2017
Hawaii, USA
September 5th, 2017

Osamu Watanabe	**(NEC Corp. | Tohoku Univ.)**
Akihiro Musa	**(NEC Corp. | Tohoku Univ.)**
Hiroaki Hokari	**(NEC Corp.)**
Shivanshu Singh	**(NEC Technologies India)**
Raghunandan Mathur	**(NEC Technologies India)**
Hiroaki Kobayashi	**(Tohoku Univ.)**

Outline

Introduction

Quantum ESPRESSO Overview

SX-ACE Overview

Performance Evaluation

Conclusions

Introduction

- Background
- Motivation

Background

Recent trends of computer simulation codes in HPC

- Various computer simulation codes have been developed as open-source software (OSS).
- Major processors adopt a concept of a vector processing to further improve the computational performance.
- Computer simulation codes need to follow the vector processing manner to have a benefit of such high performance potential.

Motivation

Evaluation and analysis of OSS codes on vector systems (NEC SX-ACE)

- OSS code: Quantum ESPRESSO
  - Quantum ESPRESSO (QE) is one of the major applications in materials science.
  - QE is widely used as a first-principle calculation application.

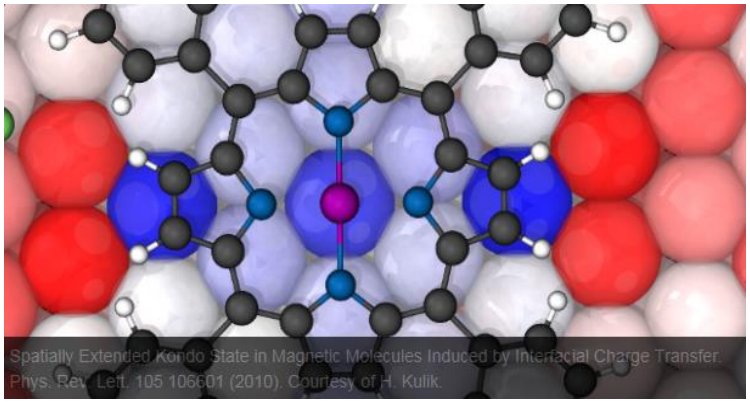
Quantum ESPRESSO Overview

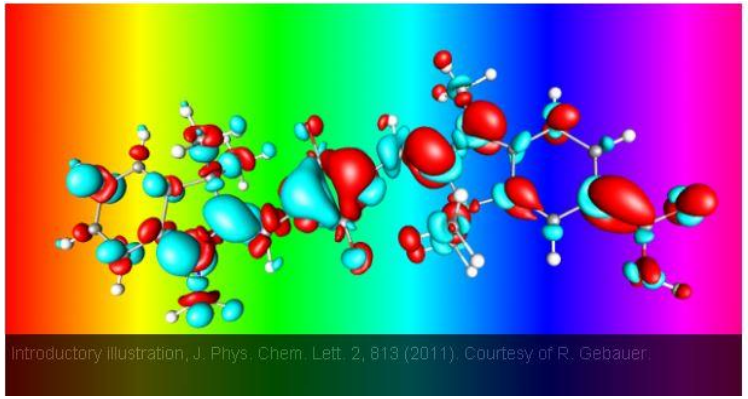
- Quantum ESPRESSO
- Packages and Plug-ins
- Programming language and parallelization
- Parallelization levels in QE

Quantum ESPRESSO

(opEn-Source Package for Research in Electronic Structure, Simulation, and Optimization)

- An integrated suite of Open-Source computer codes for electronic-structure calculations and materials modeling at the nanoscale
- Based on density-functional theory, plane waves, and pseudopotentials



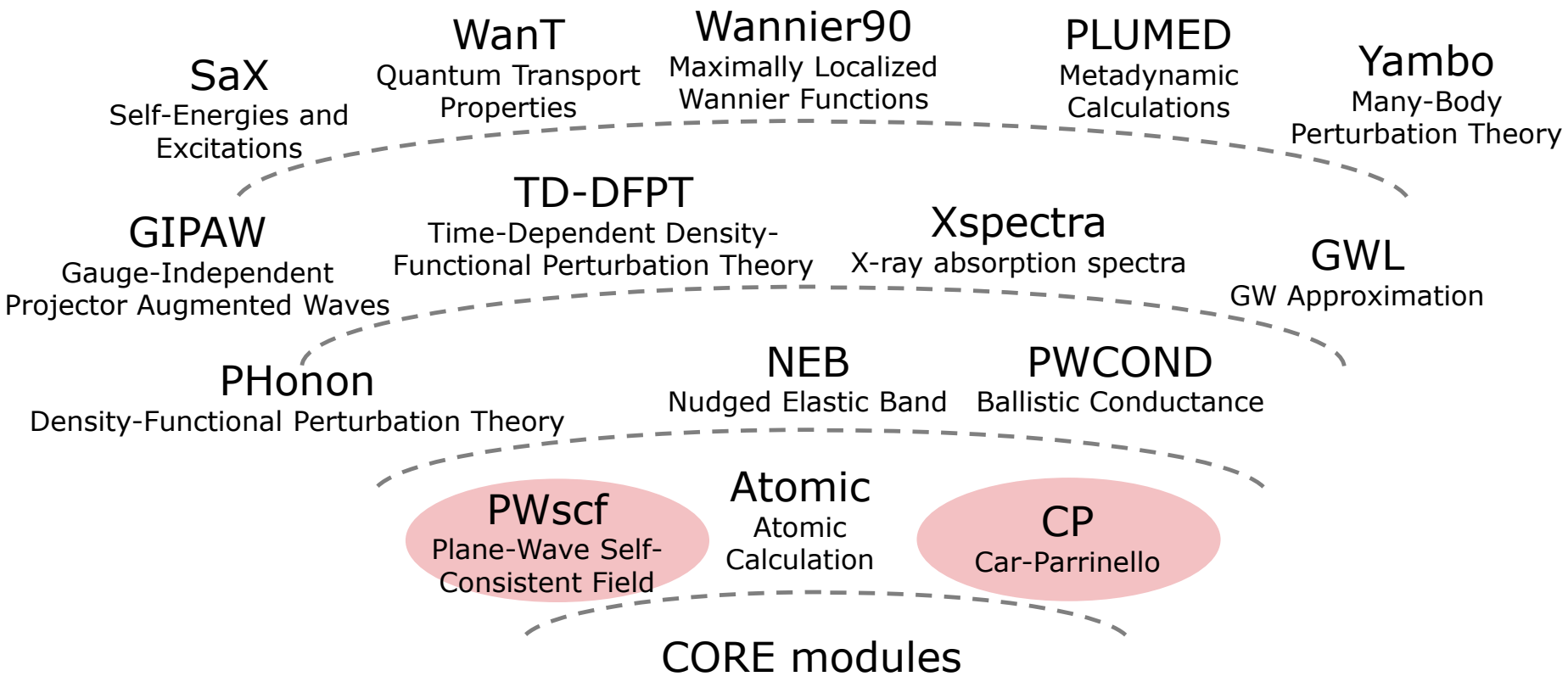


http://www.quantum-espresso.org/

Packages and Plug-ins

QE consists of various packages and plug-ins

- PWscf and CP are key packages



This presentation focuses on PWscf

Programing language and parallelization

Language

- Fortran90 + some Fortran2003 features & C

Parallelization

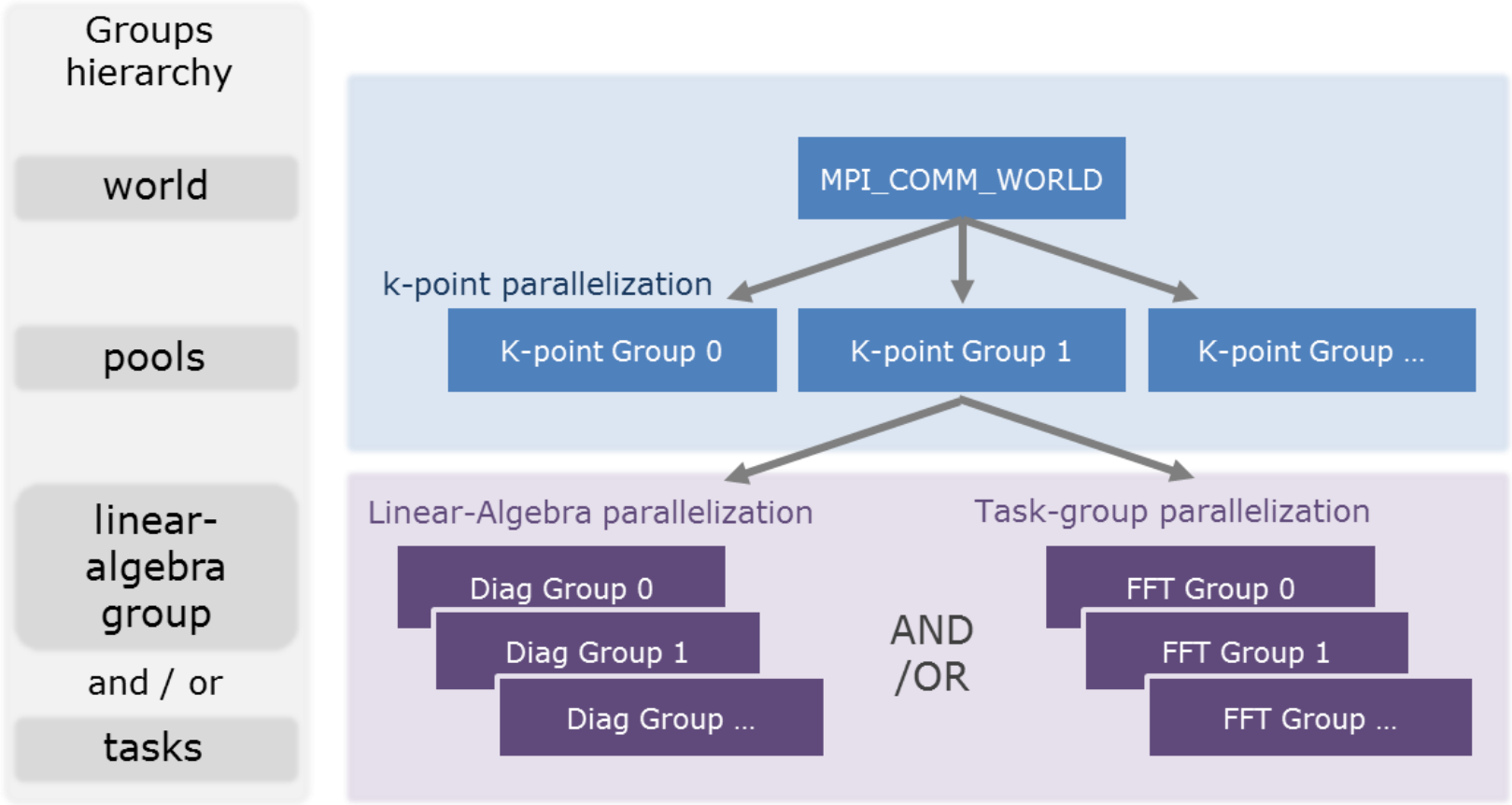
- MPI, OpenMP
  - In this evaluation, we use pure MPI for the parallelization

Numerical Library

- BLAS, LAPACK (ScaLAPACK), FFT
  - QE supports several vendor's numerical libraries

Parallelization levels in QE

- Several MPI parallelization levels are implemented in Quantum ESPRESSO.
- Both calculations and data structures are distributed across processors in the MPI parallelization levels.



SX-ACE Overview

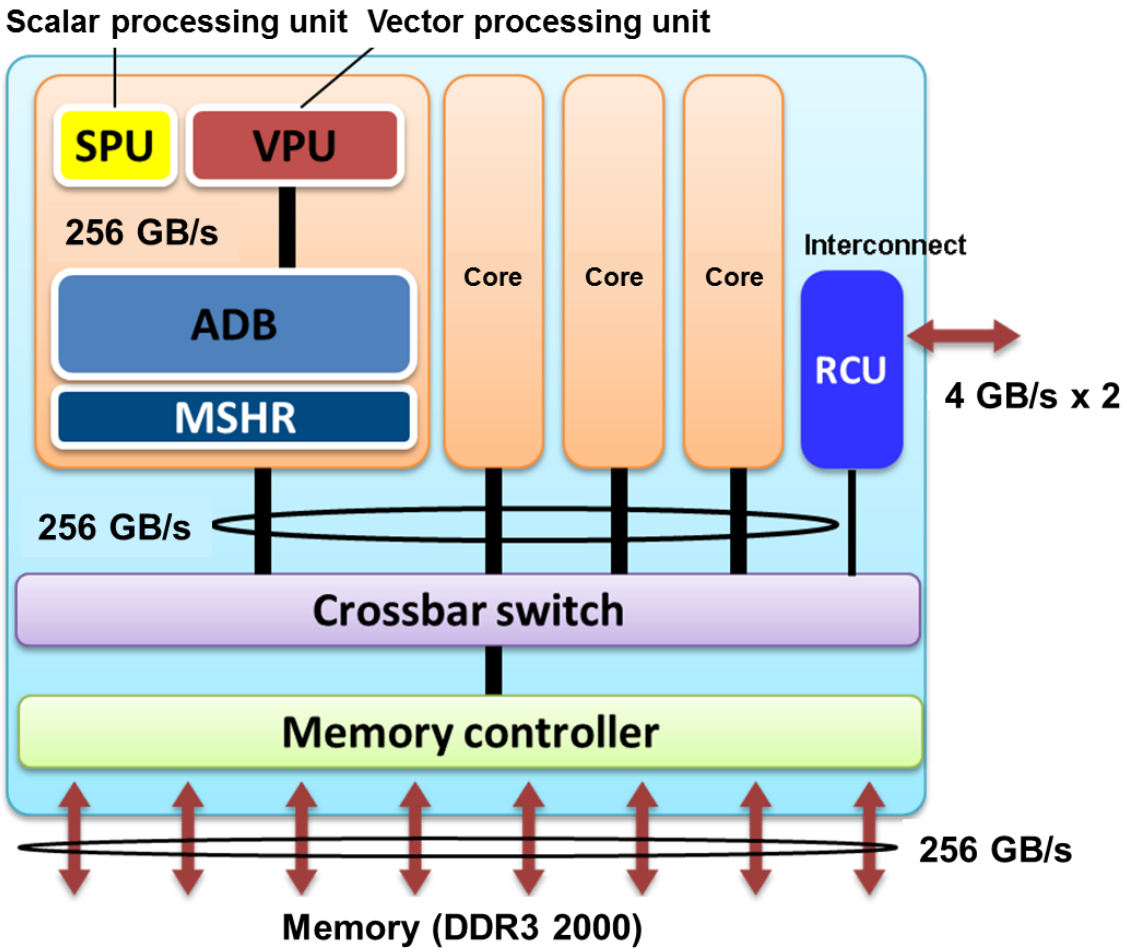
- SX-ACE
- SX-ACE processor

SX-ACE

SX-ACE

System	
# of nodes	1,024
Theoretical perf.	262 Tflop/s
Memory capacity	64 TB
Interconnect	4 GB/s x 2 (bidirectional)/node

SX-ACE processor



Core	
Theoretical perf.	64 Gflop/s
ADB capacity	1 MB
ADB bandwidth	256 GB/s
CPU	
# of cores	4
Theoretical perf.	256 Gflop/s
Memory bandwidth	256 GB/s
Node	
# of CPU	1
Memory capacity	64 GB

Performance Evaluation

- Dataset
- Performance of original PWscf(per CPU)
- Distribution of main numerical computations
- Performance improvement
- Parallelization levels
- Scalability
- Comparison with Intel Xeon server

Dataset

Using open benchmark dataset and research dataset

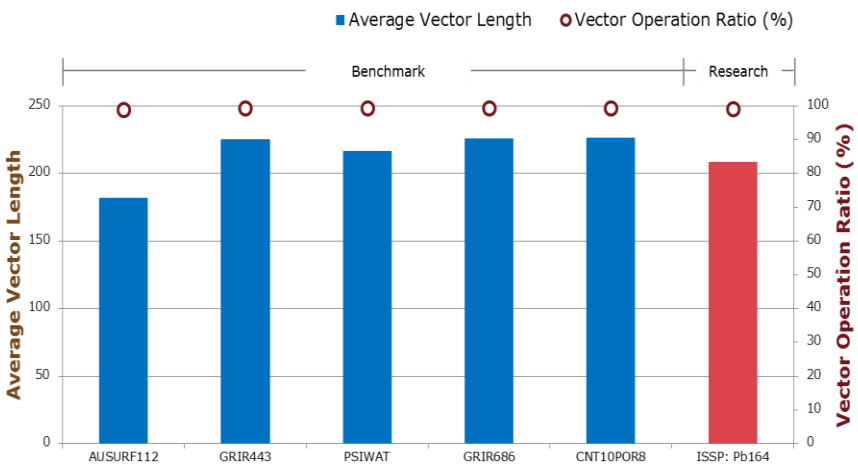
- Benchmark datasets obtained from QE official site
  (http://www.quantum-espresso.org/benchmarks)
- Research dataset provided from The Institute for Solid State Physics (ISSP), the University of Tokyo

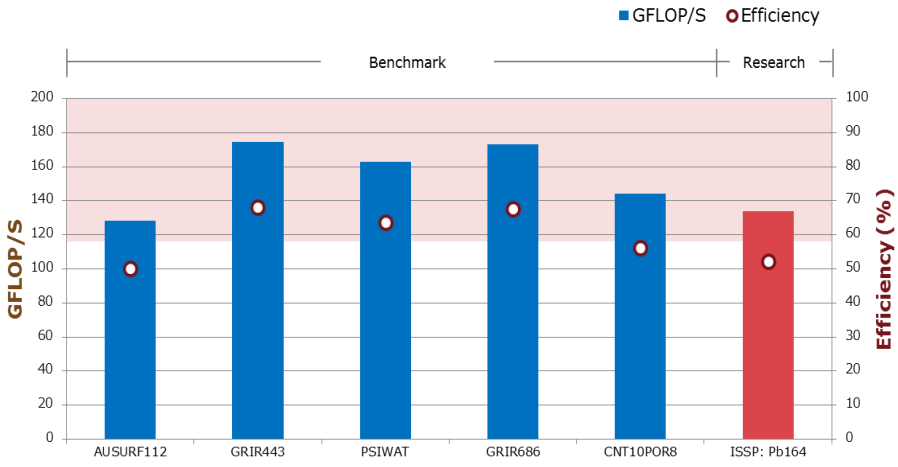
Data set	Test case	Detail	# atoms	G-vecs	k-points	FFT dimensions	# electrons	# bands	kinetic-energy cutoff	charge density cutoff
Benckmark	AUSURF112	a 112-atom gold surface	112	dense: 2,158,381 smooth: 763,307	2	dense: (180,90,288) smooth: (125,64,200)	1232	800	25 Ry	200 Ry
	GRIR443	a carbon-iridium complex (C200Ir243)	443	2,233,063	4	(180,180,192)	2987	1793	30 Ry	120 Ry
	PSIWAT	Thiol-covered gold surface and waetr	586	dense: 2,195,369 smooth: 776,119	4	dense: (90,180,288) smooth: (64,125,200)	2552	1531	25 Ry	200 Ry
	GRIR686	a carbon-iridium complex (C200Ir486)	686	2,469,147	4	(180,180,216)	5174	3104	30 Ry	120 Ry
	CNT10POR8	one hydrogen-saturated carbon nanotube with four porphirn rings chemically linked to the CNT surface (C116416N32H320)	1532	dense: 55,274,481 smooth: 19,543,423	1 (Γ point)	dense: (540,540,540) smooth: (375,375,375)	5232	2616	25 Ry	200 Ry
Research	Pb164	Slabmodel consisting of Pb(111) surface and probe (Pb)	164	dense: 5,844,523 smooth: 2,066,193	4	dense: (450,180,180) smooth: (320,125,125)	2296	1378	50 Ry	400 Ry

: Factors affecting computation amount

Performance of original PWscf on SX-ACE (per CPU)

- Performance on SX-ACE is very high in all datasets
  - Vector operation ratios are over 98.9%
  - Average vector lengths are over 180
  - Efficiencies are over 50%

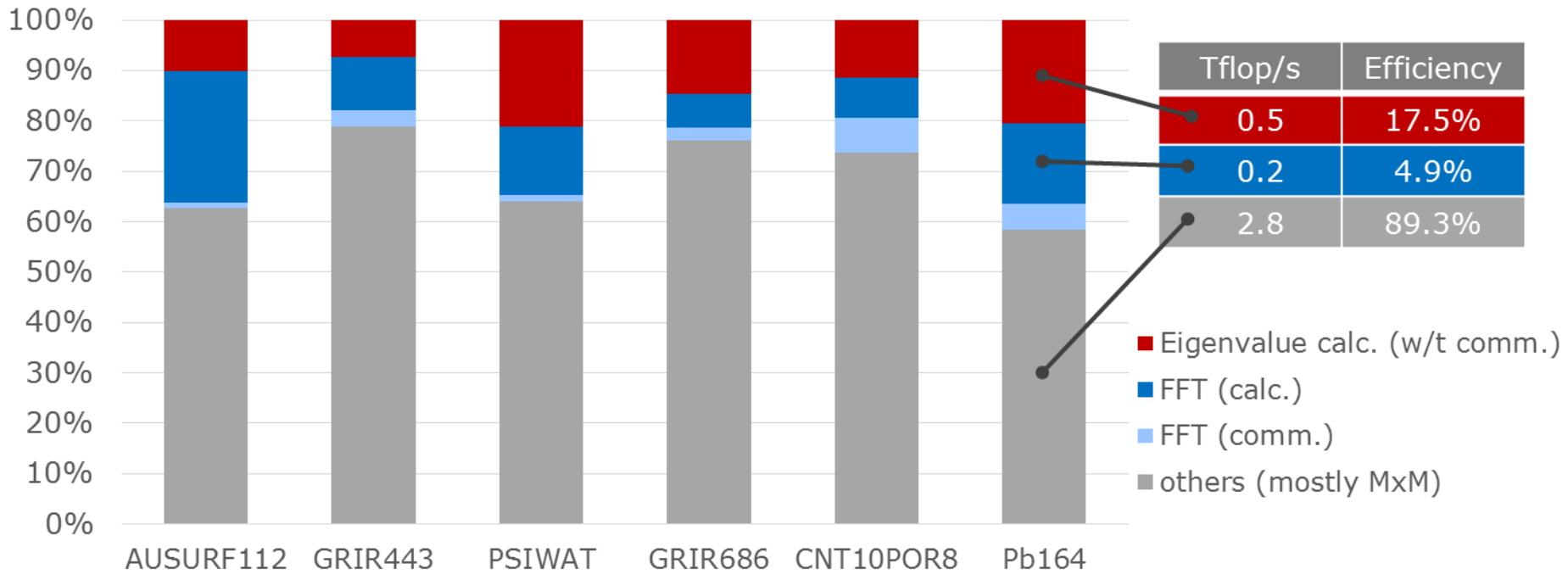




PWscf is suitable for SX-ACE

Distribution of main numerical computations

- Matrix multiplication execution time is dominant
  - This extracts very high efficiency (close to 90%)
- Need to improvement in eigenvalue calculation and FFT for further speed-up



Improvement in eigenvalue calculation

- No room for applying any optimization to ScaLAPACK routine PZHEEVD from a user side directly.

Any other appropriate numerical libraries ?

- Utilize a highly parallel computing library in place of ScaLAPACK

ELPA (Eigenvalue Solver for Petaflop-Applications)

- A high-performance computational library for the massively parallel solution of symmetric or Hermitian, standard or generalized eigenvalue problems
- The aim to develop and implement an efficient eigenvalue solver for petaflop applications
- OSS (http://elpa.mpcdf.mpg.de/)

Improvement in FFT computations

One-dimensional (1-d) complex-to-complex FFT

ZZFFT (isign, n, scale, x, y, table, work, isys)

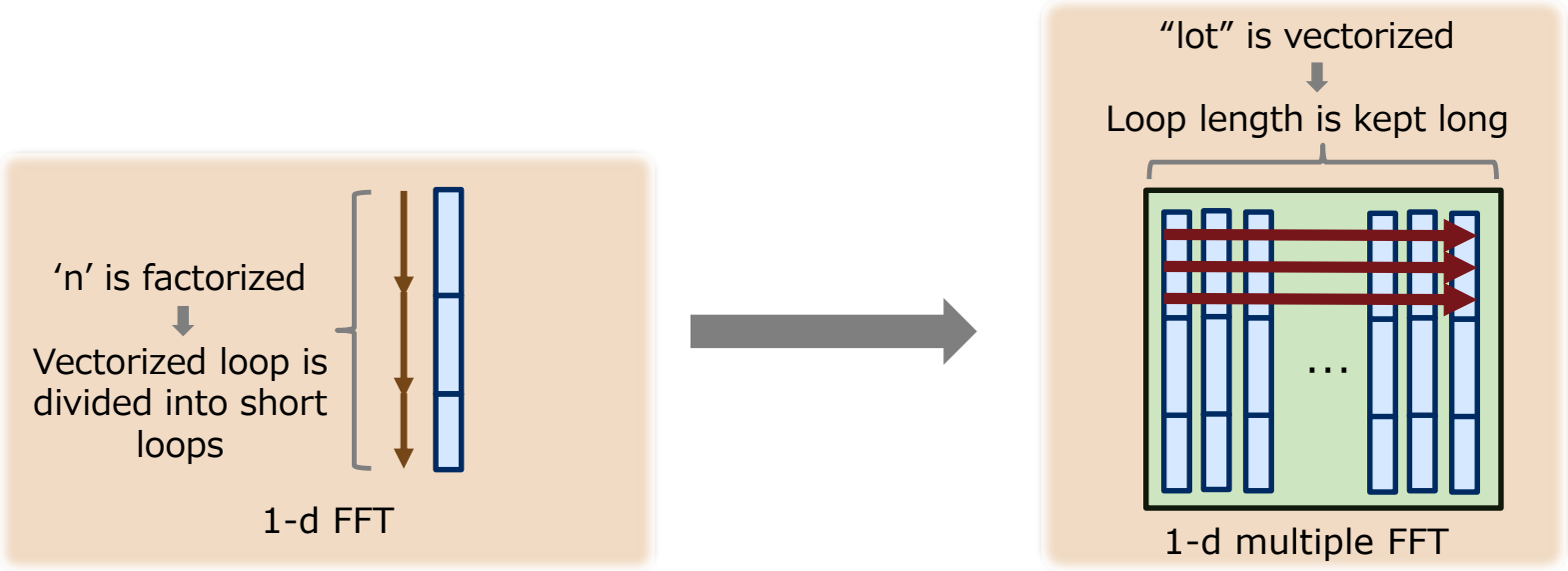
n: Number of data points (Vectorized)

One-dimensional (1-d) multiple complex-to-complex FFT

ZZFFTM (isign, n, lot, scale, x, ldx, y, ldy, table, work, isys)

n: Number of data points in a given data set

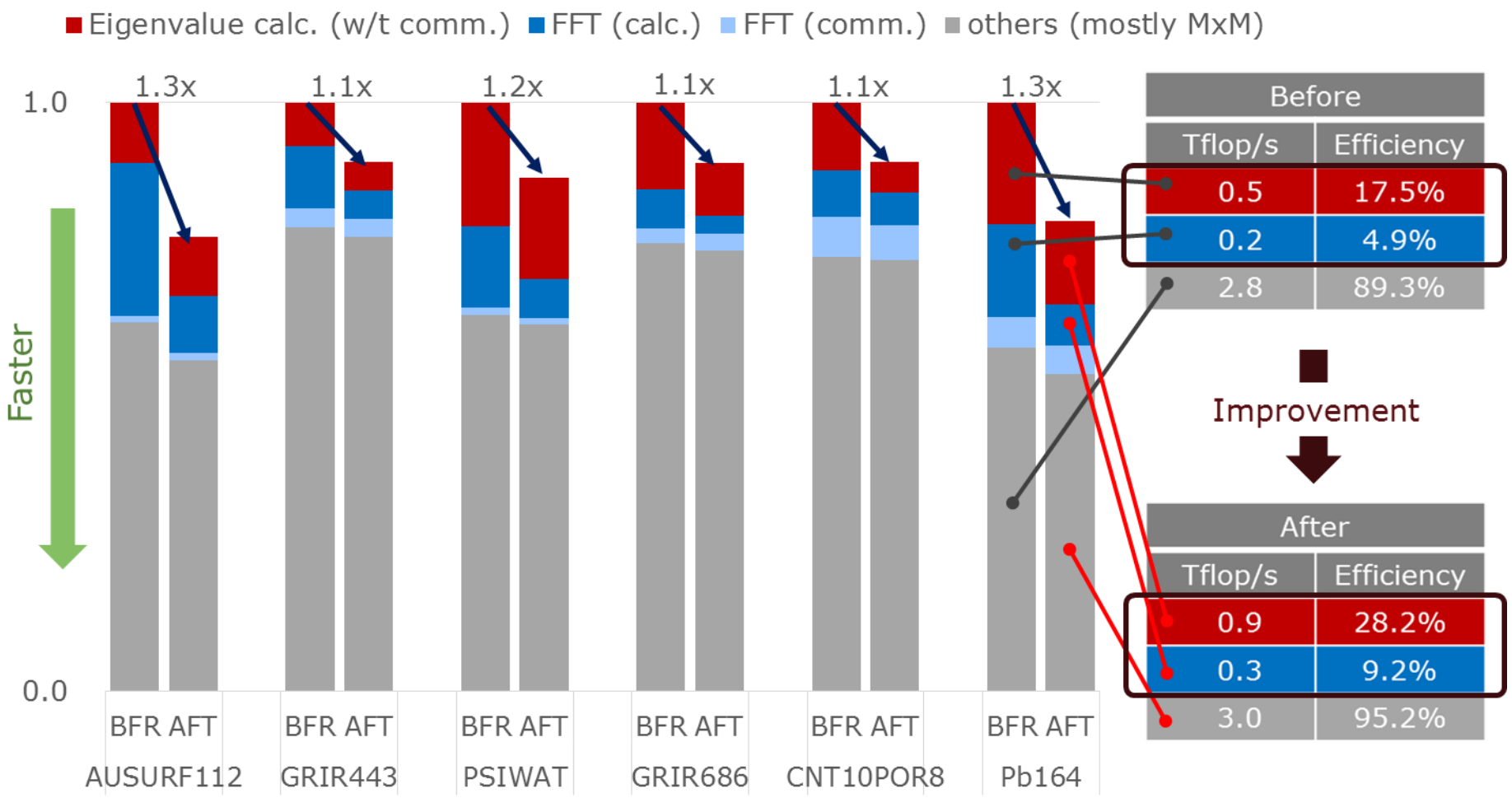
lot: Number of data sets to transform (Vectorized)



(*) SX MathKeisan FFT I/F is based on HP VECLIB and CRAY LIBSCI 3.1.

Performance improvement

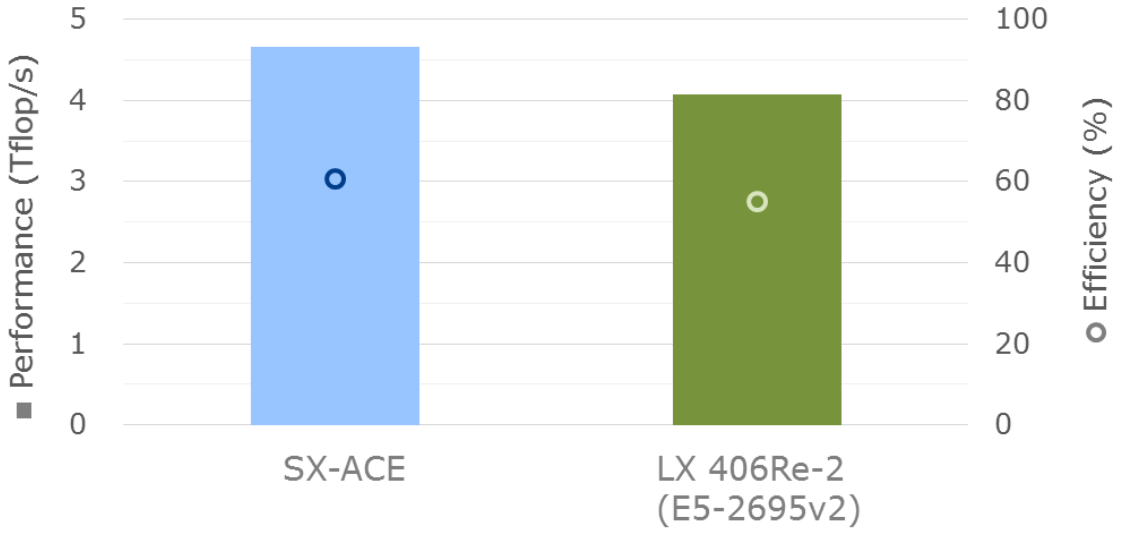
Performance improvement due to speed-up of eigenvalue calculation and FFT



Comparison with Intel Xeon server (LX 406Re-2)

Performance comparison on almost same theoretical peak performance (7.4～7.7 Tflop/s)

- The efficiencies on both SX-ACE and LX are over 50% of the peak performance
- The execution time on SX-ACE is 1.15 times faster than that on LX 406Re-2.



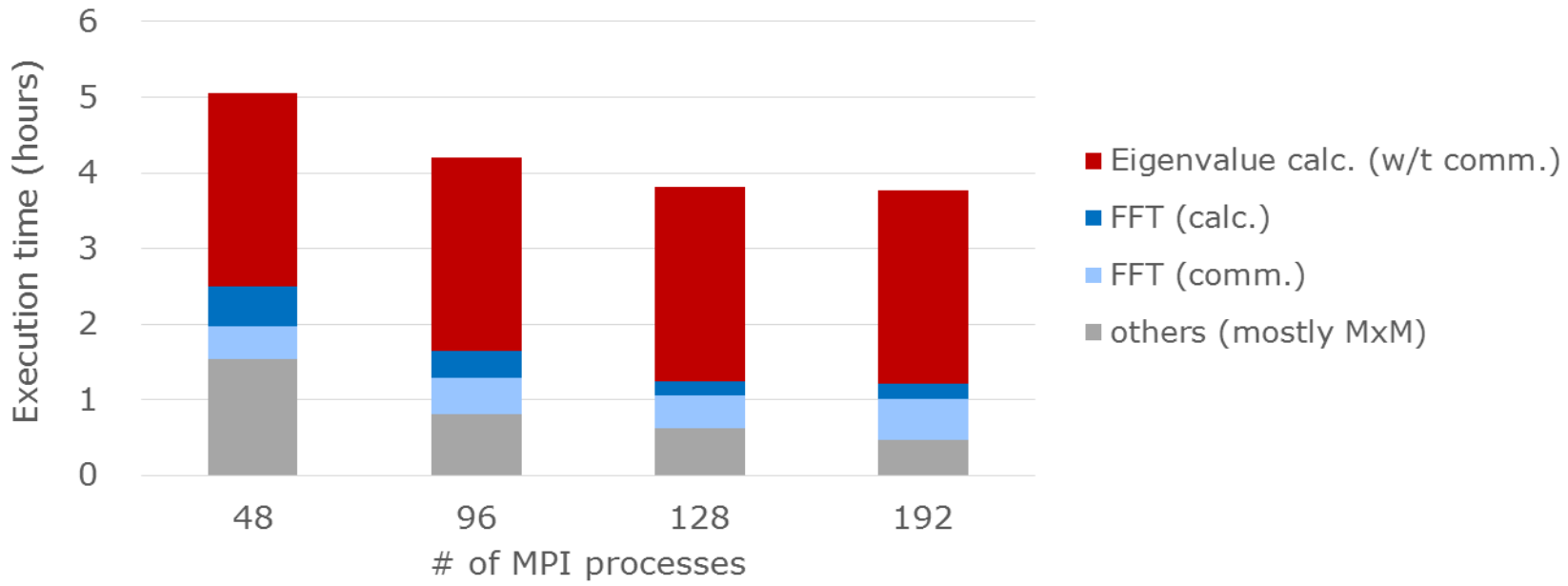
	SX-ACE	LX 406Re-2 (E5-2695v2)
# of MPI processes	120	384
# of CPUs	30	32
Total peak perf.	7.7 Tflop/s	7.4 Tflop/s

PWscf is suitable for a vector architecture!

Parallelization levels : all levels = 1

Scalability in the case of all parallelization levels:1

(# of MPI processes: variable, pools=1, linear-algebra group=1, tasks=1)

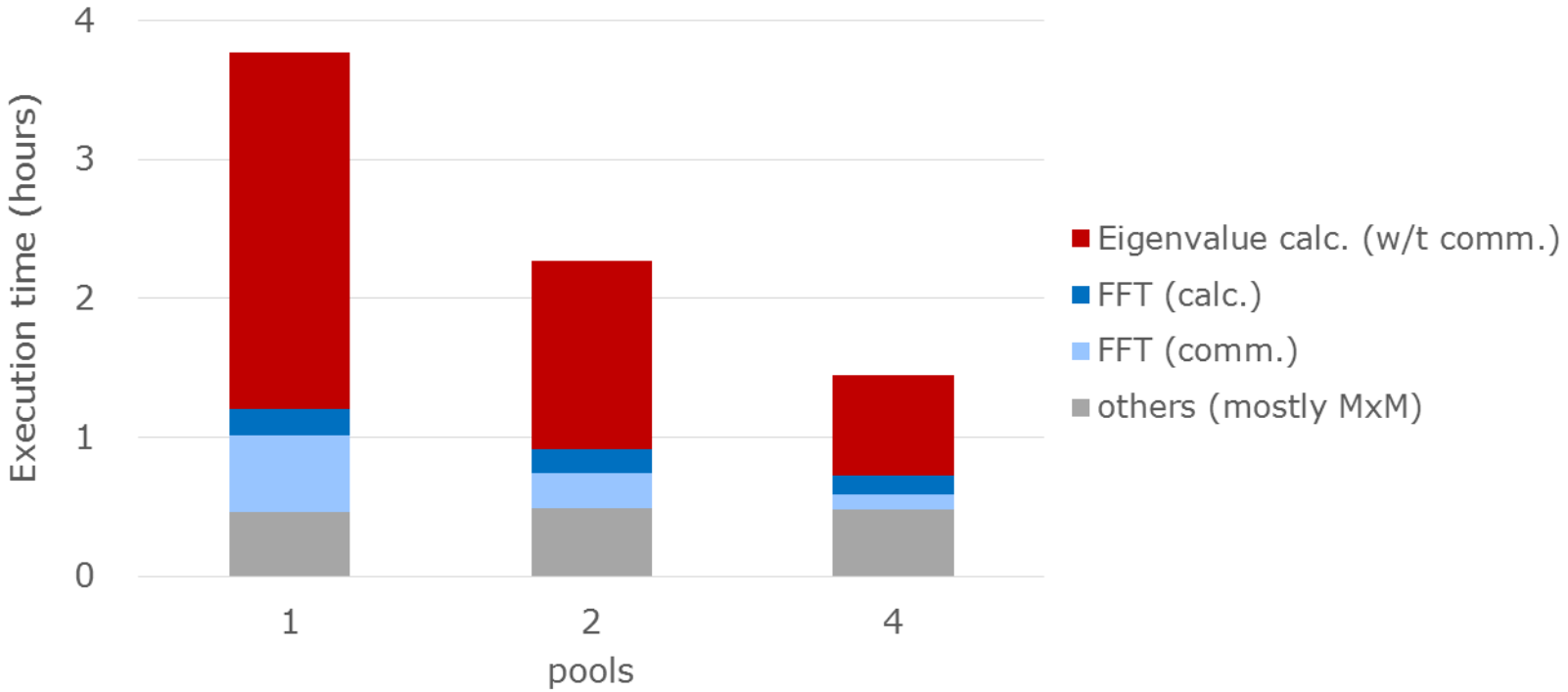


- Matrix multiplication executes in parallel.
- Eigenvalue calculation executes in serial.
- FFT communication (alltoall) time increases with increase in number of MPI processes.

Parallelization levels : k-point parallelization

Scalability in the case of distributed k-points

(# of MPI processes: 192mpi, pools=1~4, linear-algebra group=1, tasks=1)

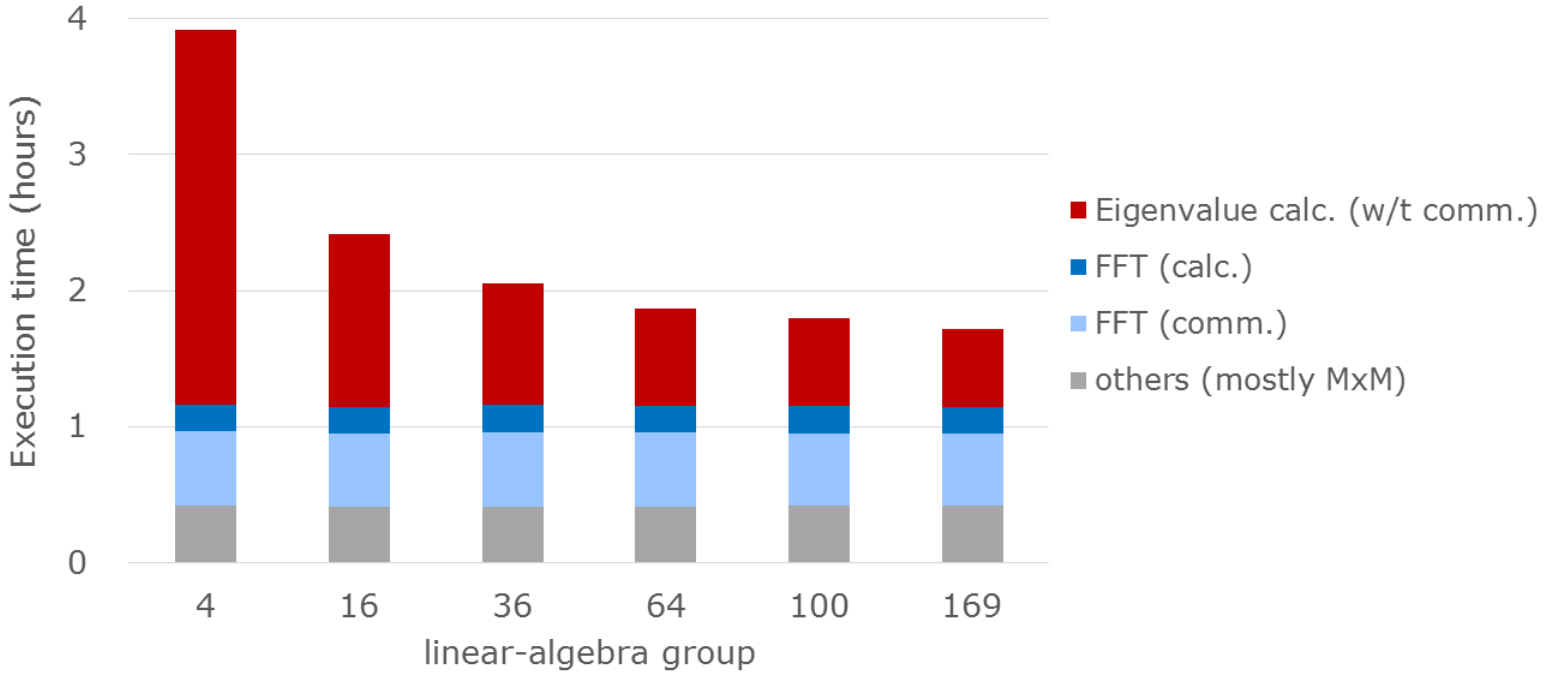


- Matrix multiplication is not affected.
- Eigenvalue calculation executes in parallel.
- FFT communication (alltoall) time decreases with increase of pools.

Parallelization levels : Linear-Algebra parallelization

Scalability in the case of distributed subspace hamiltonians and constrains matrices

(# of MPI processes: 192mpi, pools=1, linear algebra group=1~169, tasks=1)

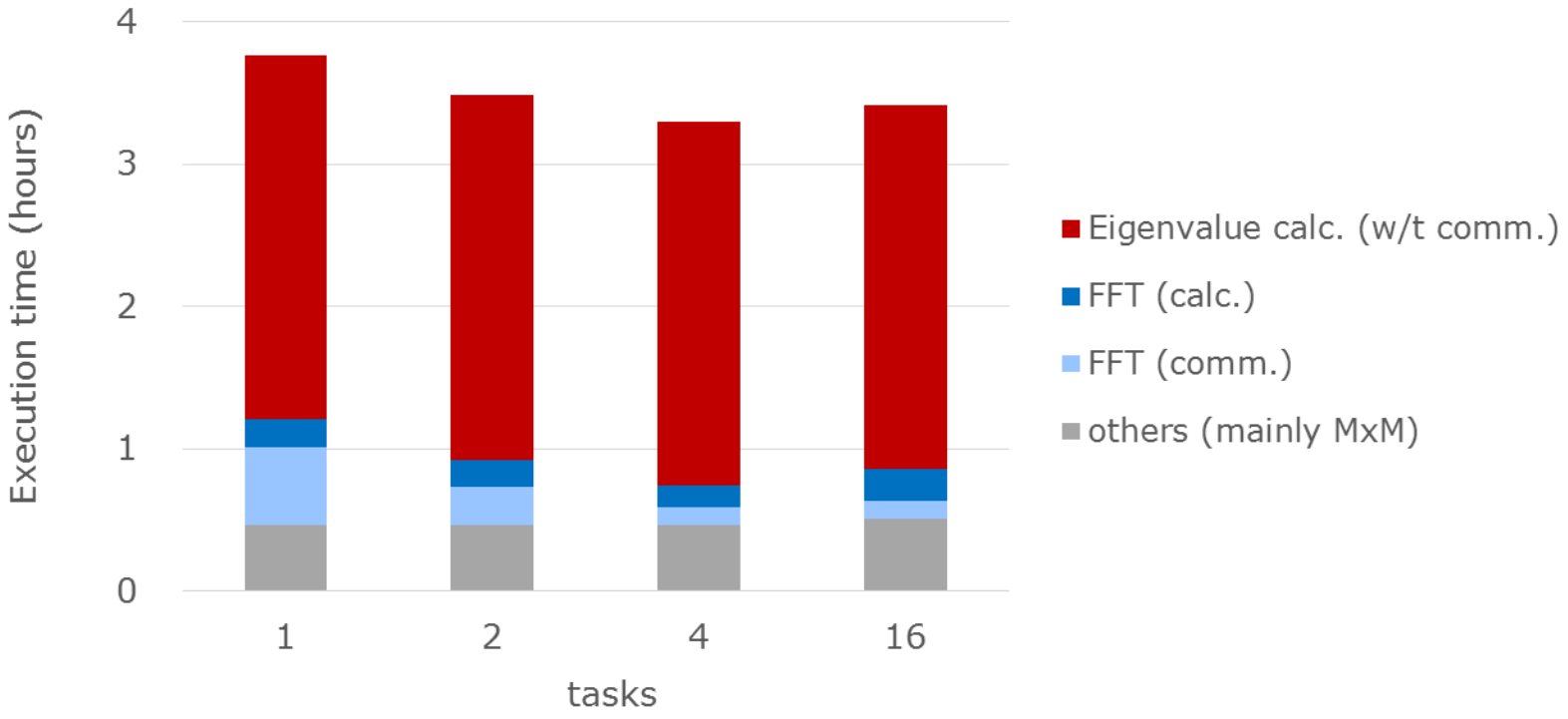


- Matrix multiplication and FFT are not affected.
- Eigenvalue execution time decreases with increase of linear algebra group.
  - The effect is good for lesser number of linear algebra group (~16).
  - In larger number, the effect is small.

Parallelization levels : Task-group parallelization

Scalability in the case of distributed FFT on electron states

(# of MPI processes: 192mpi, pools=1, linear-algebra group=1, tasks=1~32)

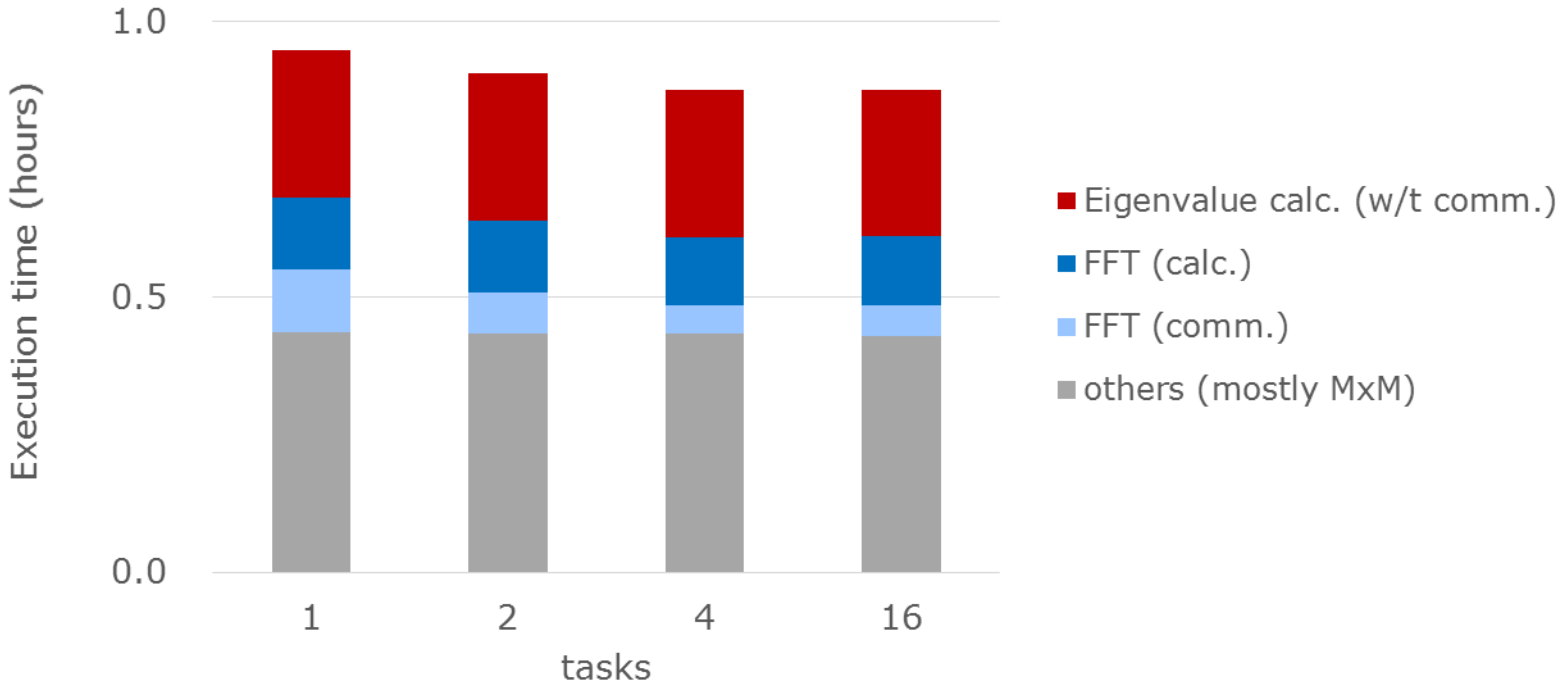


- Matrix multiplication and eigenvalue calculation are not affected.
- Effect on FFT calculation time is not much.
- FFT communication (alltoall) time decreases with increase of tasks up to 4.

Parallelization levels : Task-group parallelization

Scalability in the case of distributed FFT on electron states

(# of MPI processes: 192mpi, pools=4, linear-algebra group=36, tasks=1~32)

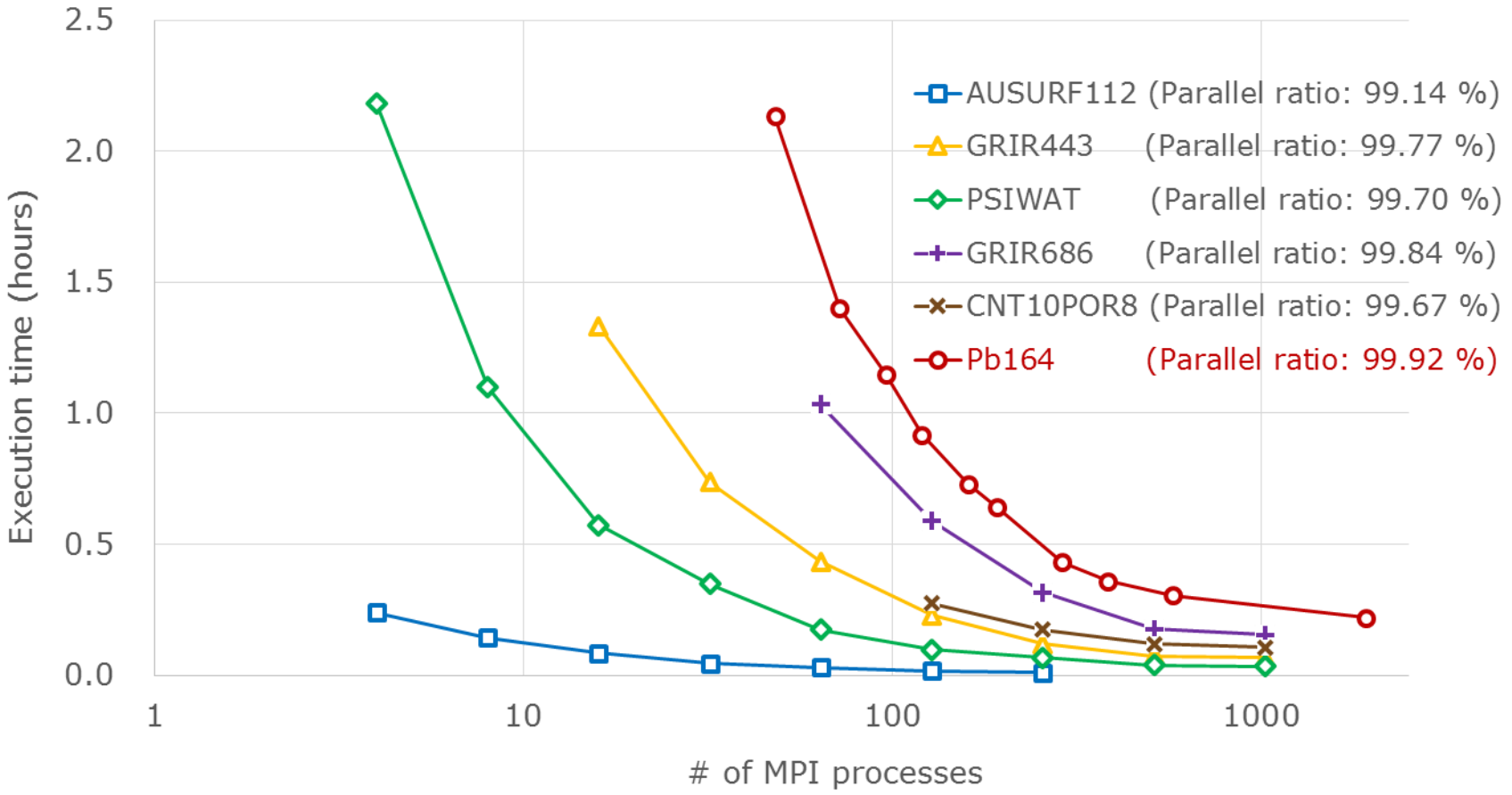


- When used with other parallelization levels, increase of tasks has an effect slightly on reduction of FFT execution time.
  - Especially, FFT communication (alltoall) time is reduced.

Scalability

After the optimization (library replacement), the parallelization ratio improves from 99.54% to 99.66% on avarage

- Both optimizations (eigenvalue calculation and FFT) shows good effect.



Conclusions

- Conclusion
- Future work

Conclusion

- PWscf can be executed very efficient on a vector architecture.
  - PWscf can achieve high performance on SX-ACE.
  - The vector function on Intel Xeon also works with the vectorization of PWscf.
- The performance of PWscf depends on numerical libraries and parallelization levels.
  - 1.3 times faster by applying more high-performance numerical libraries' routines.
  - The scalability improvements by using the optimum setting of parallelization levels.

Future work

- Performance evaluation of more packages in QE using latest version.

- Expansion of target OSS applications and computational science areas.
  - For example, meteorology and climatology, plasma/fusion science, thermal fluid, and so on.

Thank you for your kind attention!

Many thanks to

- Prof. Shinji Tsuneyuki (ISSP, The University of Tokyo)
- Dr. Mitsuaki Kawamura (ISSP, The University of Tokyo)
- Cyberscience Center, Tohoku University