



東北大学



東北大学新スーパーコンピュータシステムの紹介と 高性能計算に関する研究開発活動

Hiroaki Kobayashi

Director and Professor
Cyberscience Center
Tohoku University
koba@cc.tohoku.ac.jp

PC Cluster Workshop in Sendai
Feb. 19, 2016

Missions of Cyberscience Center As a National Supercomputer Center



★ High-Performance Computing Center founded in 1969

● Offering leading-edge high-performance computing environments to academic users nationwide in Japan

- 👁 24/7 operations of large-scale vector-parallel and scalar-parallel systems
- 👁 1500 users registered in AY 2014



1969



1982

● User supports

- 👁 Benchmarking, analyzing, and tuning users' programs
- 👁 Holding seminars and lectures



SX-1 in 1985



SX-2 in 1989

● Supercomputing R&D, collaborating work with NEC

- 👁 Designing next-generation high-performance computing systems and their applications for highly-productive supercomputing



SX-3 in 1994



SX-4 in 1998

- 👁 57-year history of collaboration between Tohoku University and NEC on High Performance Computing

● Education

- 👁 Teaching and supervising BS, MS and Ph.D. Students as a cooperative laboratory of graduate school of information sciences, Tohoku university

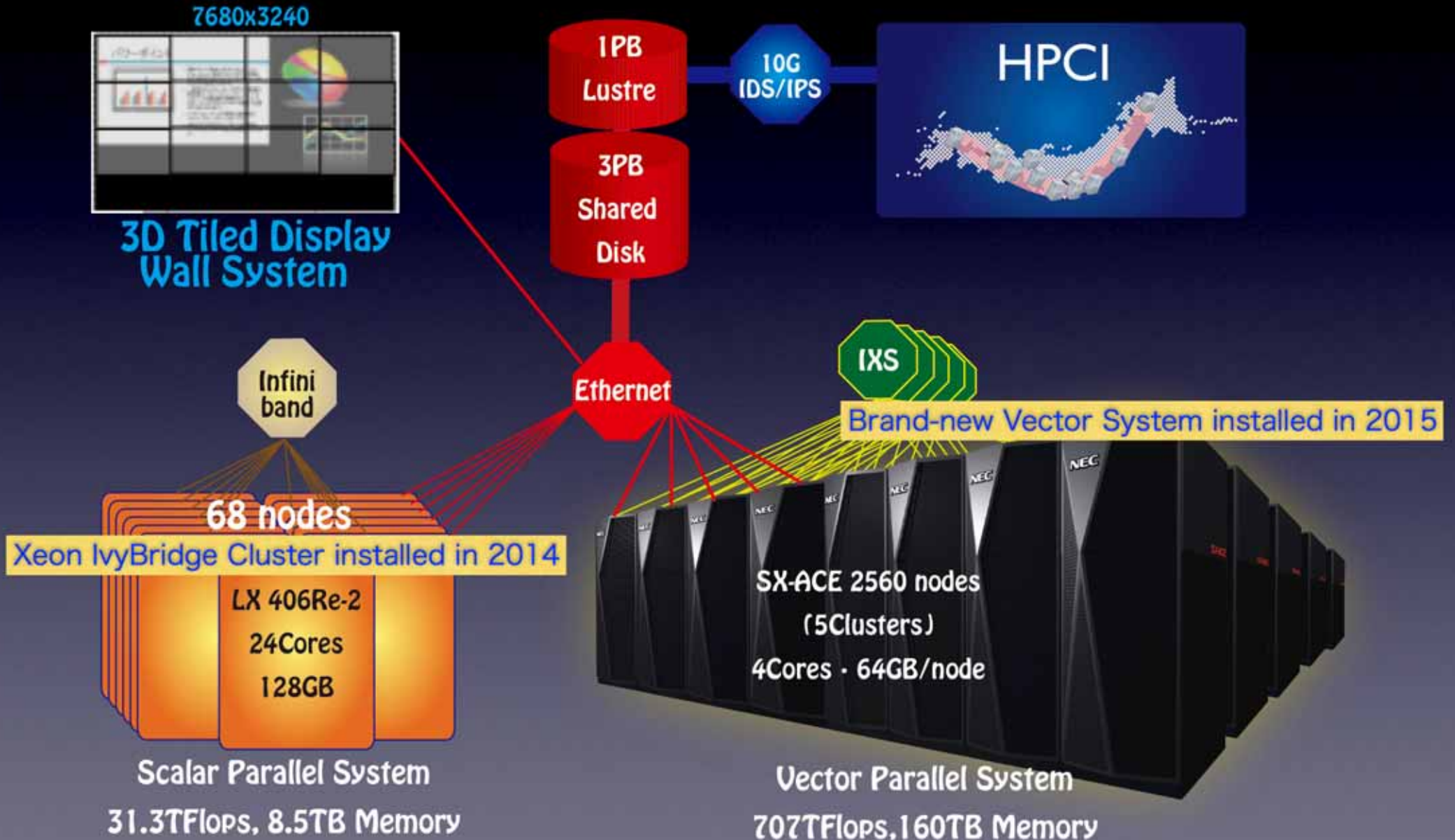


SX-7 in 2003



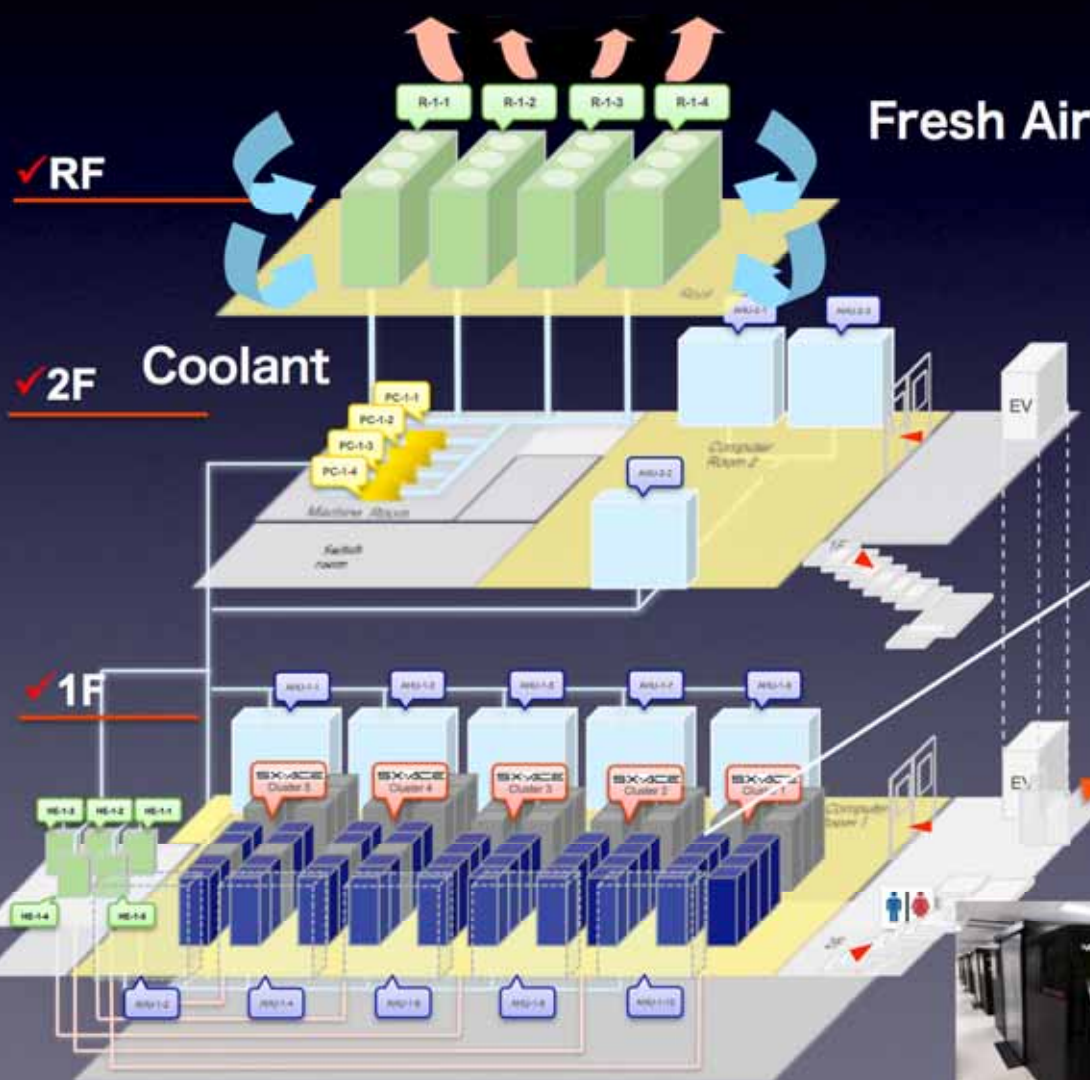
SX-9 in 2008

Tohoku Univ.'s New Supercomputer System (2015.2.20~)



Cooling Facility of HPC Building

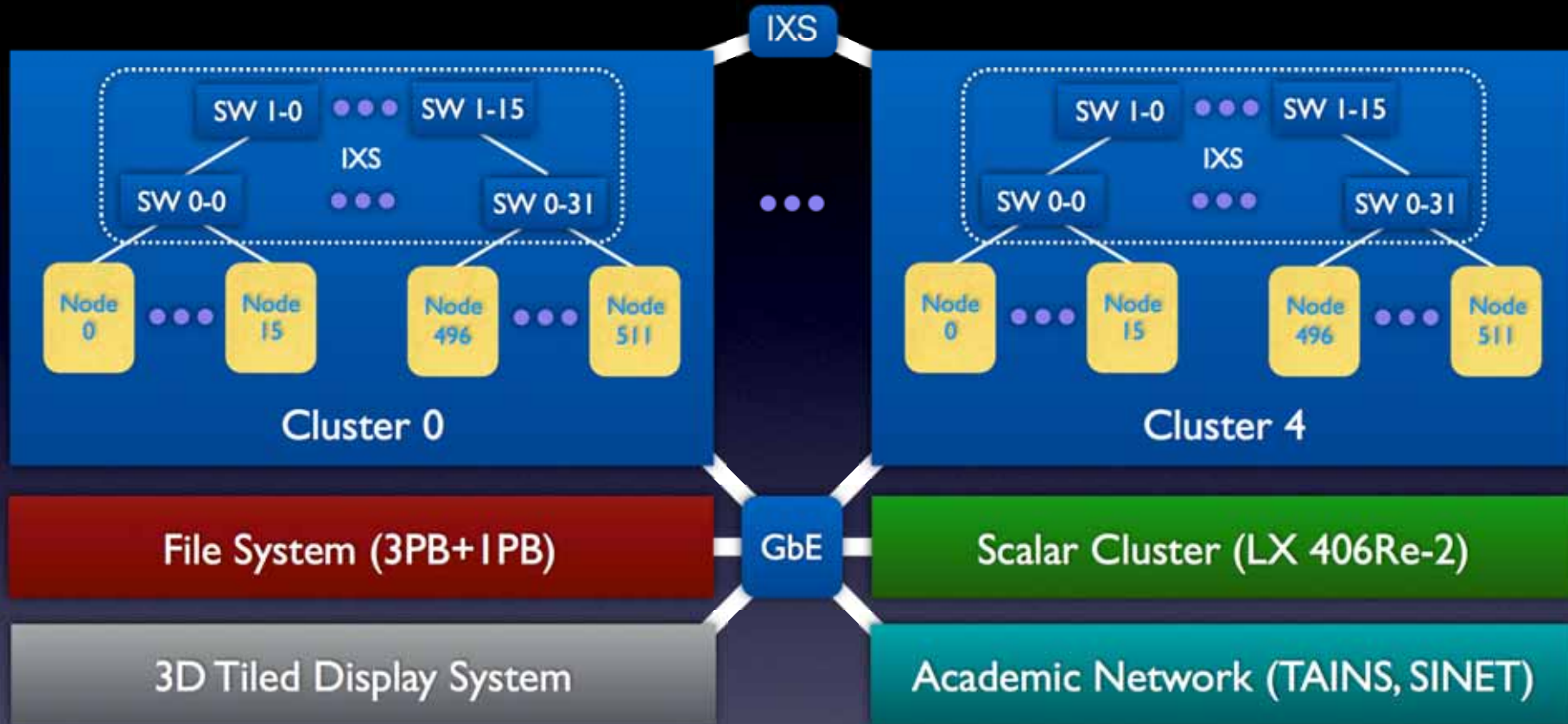
Evaporation



- SX-ACE
- Air-conditioning Equipment
- Heat-exchange equipment
- coolant pump



Organization of Tohoku Univ. SX-ACE System



	Core	CPU(Socket)	Node	Cluster	Total System
Size	1	4 Cores	1CPU	512 Nodes	5 Clusters
Performance (VPU+SPU)	69GFlop/s (68GF+1GF)	276GFlop/s (272GF+4GF)		141Tflop/s (139TF+ 2TF)	707Tflop/s (697TF+10TF)
Mem. BW	256GB/s			131TB/s	655TB/s
Memory Cap.	64GB			32TB	160TB
IXS Node BW	-		4GB/s x2		-

Features of the SX-ACE Vector Processor

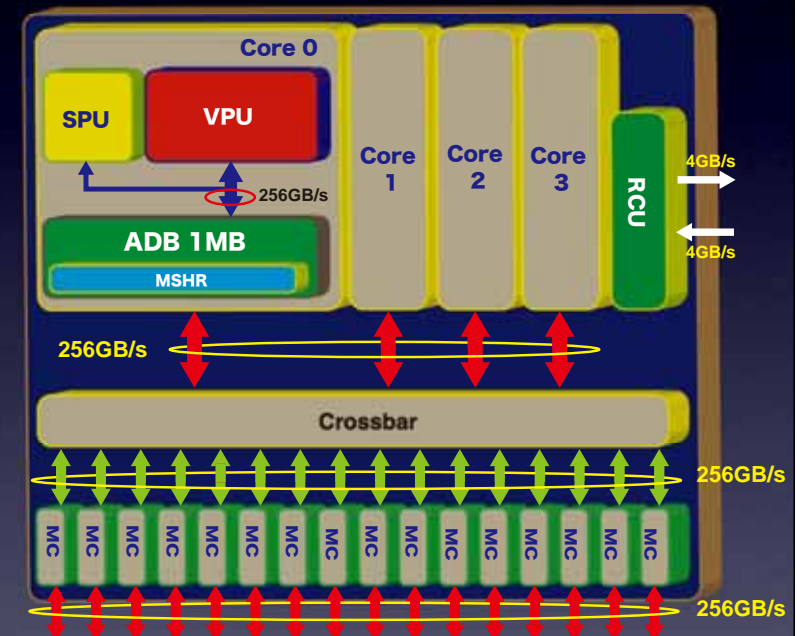
4 Core Configuration, each with High-Performance Vector-Processing Unit and Scalar Processing Unit

- 272Gflop/s of VPU + 4Gflop/s of SPU per socket
 - 68Gflop/s + 1Gflop/s per core
- 1MB private ADB per core (4MB per socket)
 - Software-controlled on-chip memory for vector load/store
 - 4x compared with SX-9
 - 4-way set-associative
 - MSHR with 512 entries (address+data)
 - 256GB/s to/from Vec. Reg.
 - 4B/F for Multiply-Add operations
- 256 GB/s memory bandwidth, Shared with 4 cores
 - 1B/F in 4-core Multiply-Add operations
~ 4B/F in 1-core Multiply-Add operations
 - 128 memory banks per socket

Other improvement and new mechanisms to enhance vector processing capability, especially for efficient handling of short vectors operations and indirect memory accesses

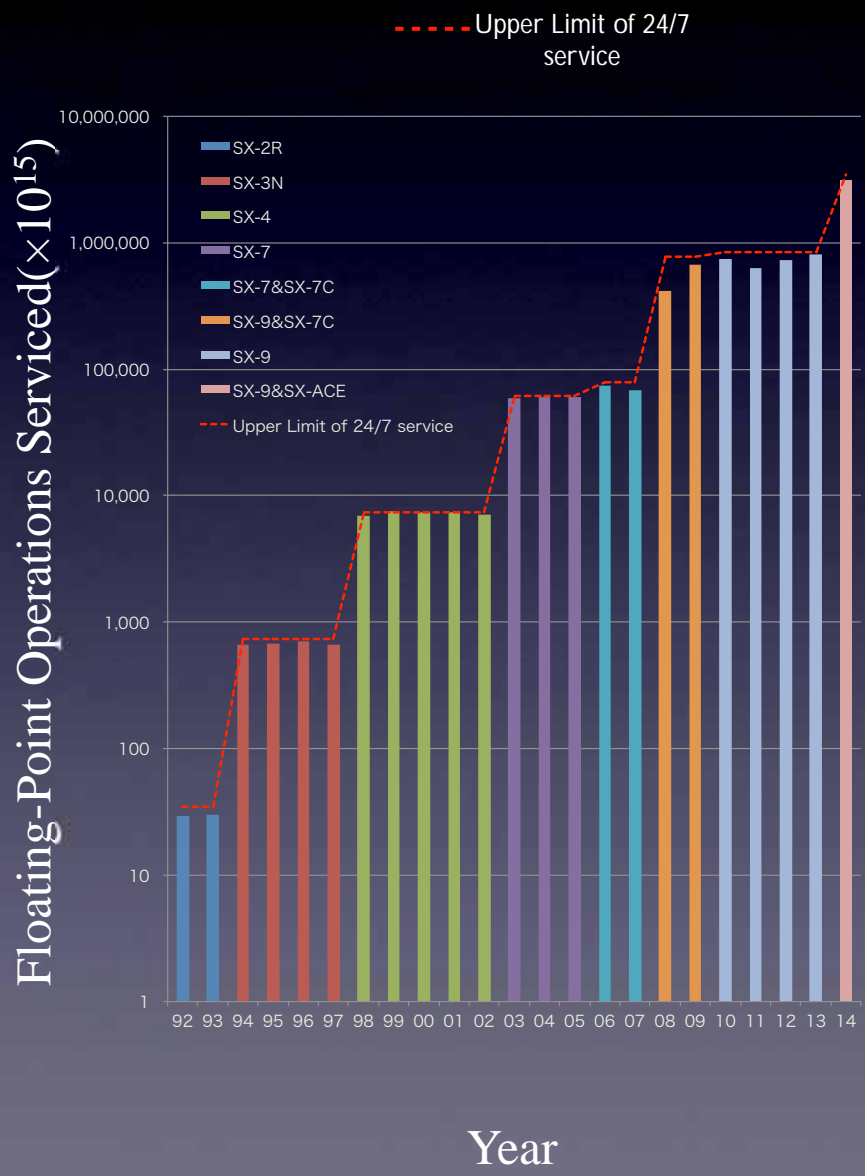
- Out of Order execution for vector load/store operations
- Advanced data forwarding in vector pipes chaining
- Shorter memory latency than SX-9

SX-ACE Processor Architecture

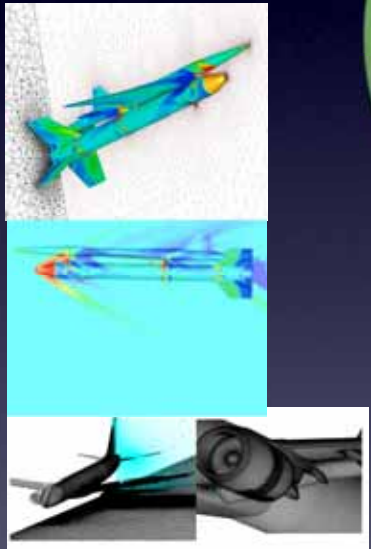


Source: NEC

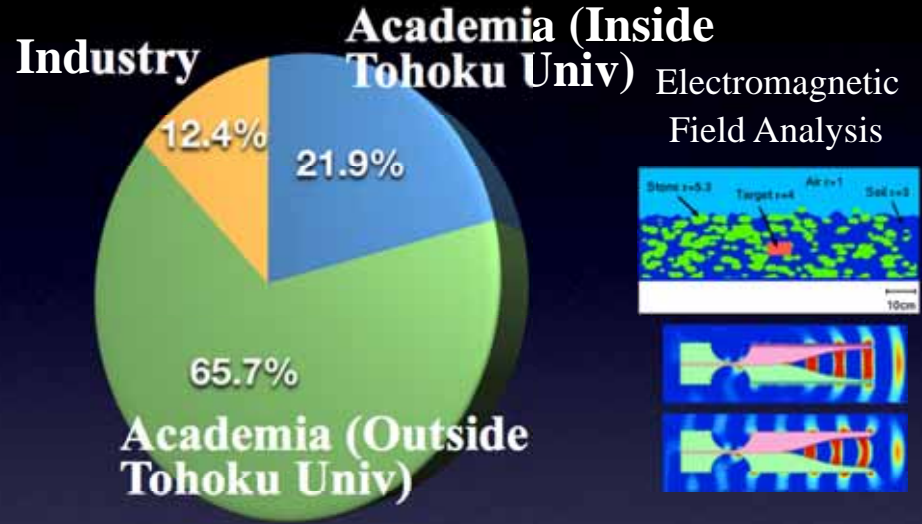
High Demands for Vector Systems in Memory-Intensive, Science and Engineering Applications



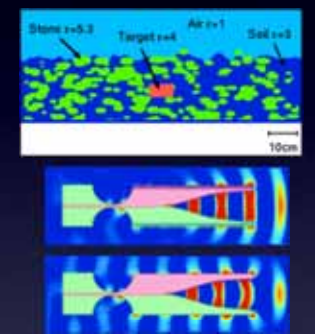
Advanced CFD model for digital flight



Turbine Design



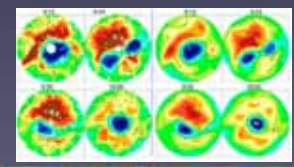
Electromagnetic Field Analysis



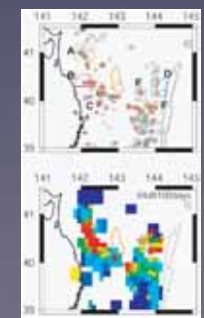
Catalyzer Design



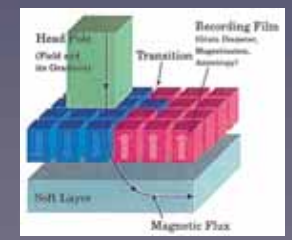
Atmosphere Analysis & Weather Forecasting



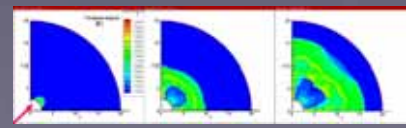
Earthquake analysis



Magnetic recording device Design



Combustion





東北大学



Performance Evaluations of SX-ACE

Specifications of Modern High End Systems

System	No. of Sockets/ Node	Perf./ Socket (Gflop/s)	No. of Cores	Perf. /core (Gflop/s)	Mem. BW GB/sec	On-chip mem	NW BW (GB/sec)	Sys. B/F
SX-ACE	1	256	4	64	256	1MB ADB /core	2 x 4 IXS	1.0
SX-9	16	102.4	1	102.4	256	256KB ADB/core	2 x 128 IXS	2.5
ES2	8	102.4	1	102.4	256	256KB ADB/core	2 x 64IXS	2.5
LX 406 (Ivy Bridge)	2	230.4	12	19.2	59.7	256KB L2/core 30MB Shared L3	5 IB	0.26
FX10 (SPARK64IX)	1	236.5	16	14.78	85	12MB shared L2	5 - 50 Tofu NW	0.36
K (SPARK64VIII)	1	128	8	16	64	6MB Shared L2	5 - 50 Tofu NW	0.5
SR16K M1 (Power7)	4	245.1	8	30.6	128	256KB L2/core 32MB shared L3	2 x 24 - 96 custom NW	0.52

Remarks: Listed performances are obtained based on total Multiply-Add performances of individual systems

Applications Used for Evaluation

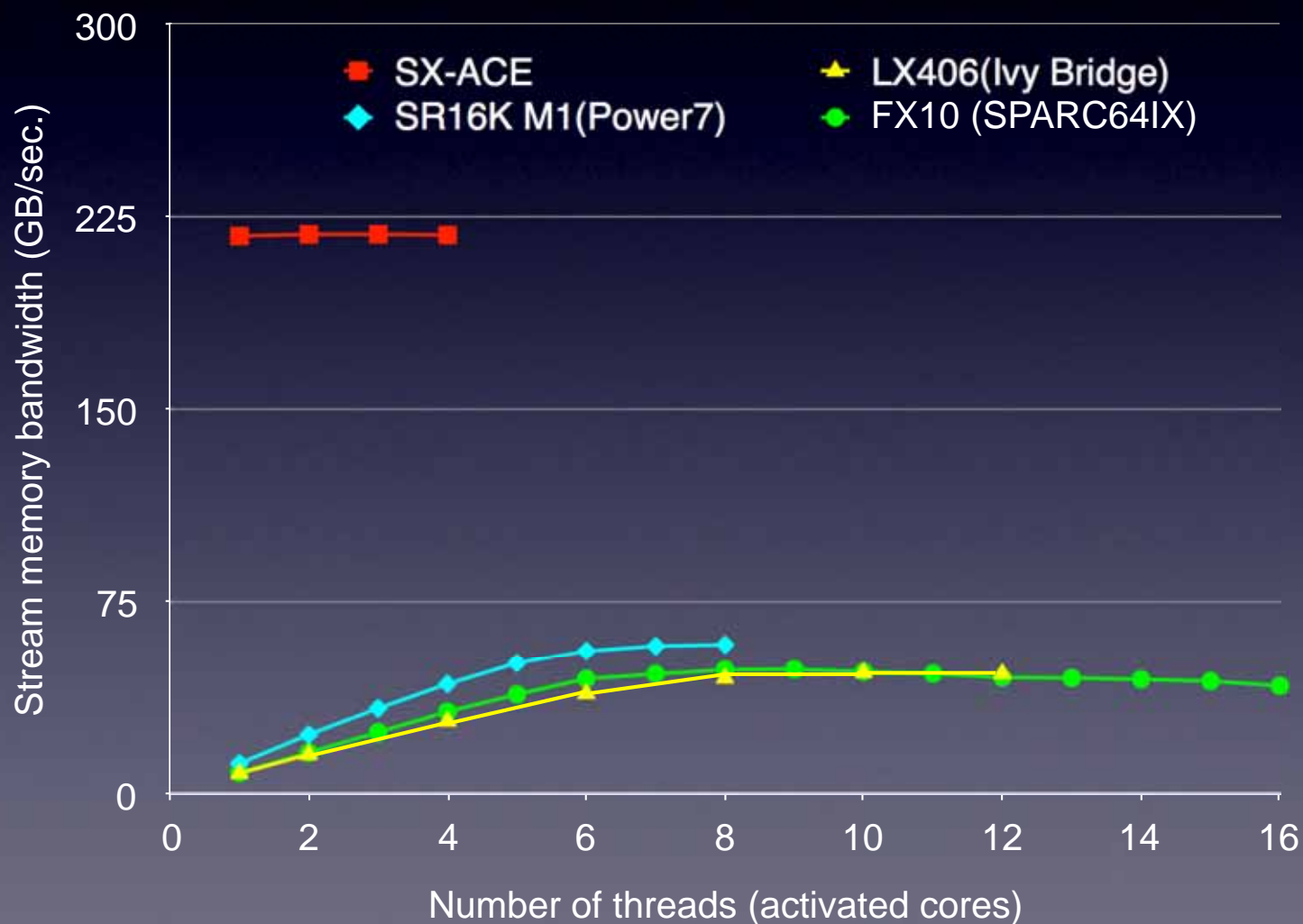
Applications	Fields	Methods	Mem Access Characteristics	Mesh Size	Code B/F	Actual B/F on ACE
QSFDM GLOBE	Seismology	Spherical 2.5D FDM	Stencil with sequential memory accesses	4.3×10^7 grids	2.16	0.78
Barotropic ocean	OGCM (Ocean General Circulation Model)	Shallow water model	Stencil with sequential memory accesses	4322×216	1.97	1.11
MHD (FDM)	MHD	Finite Difference Method	Stencil with sequential memory accesses	$200 \times 1920 \times 32$	3.04	1.41
Seism 3D	Seismology	Finite Difference Model	Stencil with sequential memory accesses	$1024 \times 512 \times 512$ † $4096 \times 2048 \times 2048$ ‡	2.15	1.68
MHD (Spectral)	MHD	Pseudo spectral Method	Stride memory access	$900 \times 768 \times 96$ † $3600 \times 3072 \times 2048$ ‡	2.21	2.18
TURBINE	CFD	DNS	Indirect memory access with short vectors	$91 \times 91 \times 91 \times 13$	1.78	5.47
BCM	CFD	Navier Stokes Equation	Stencil and Indirect memory access	$(128 \times 128 \times 128 \text{ cells}) \times 64 \text{ Cubes}$	7.01	5.86

† for single-node evaluation

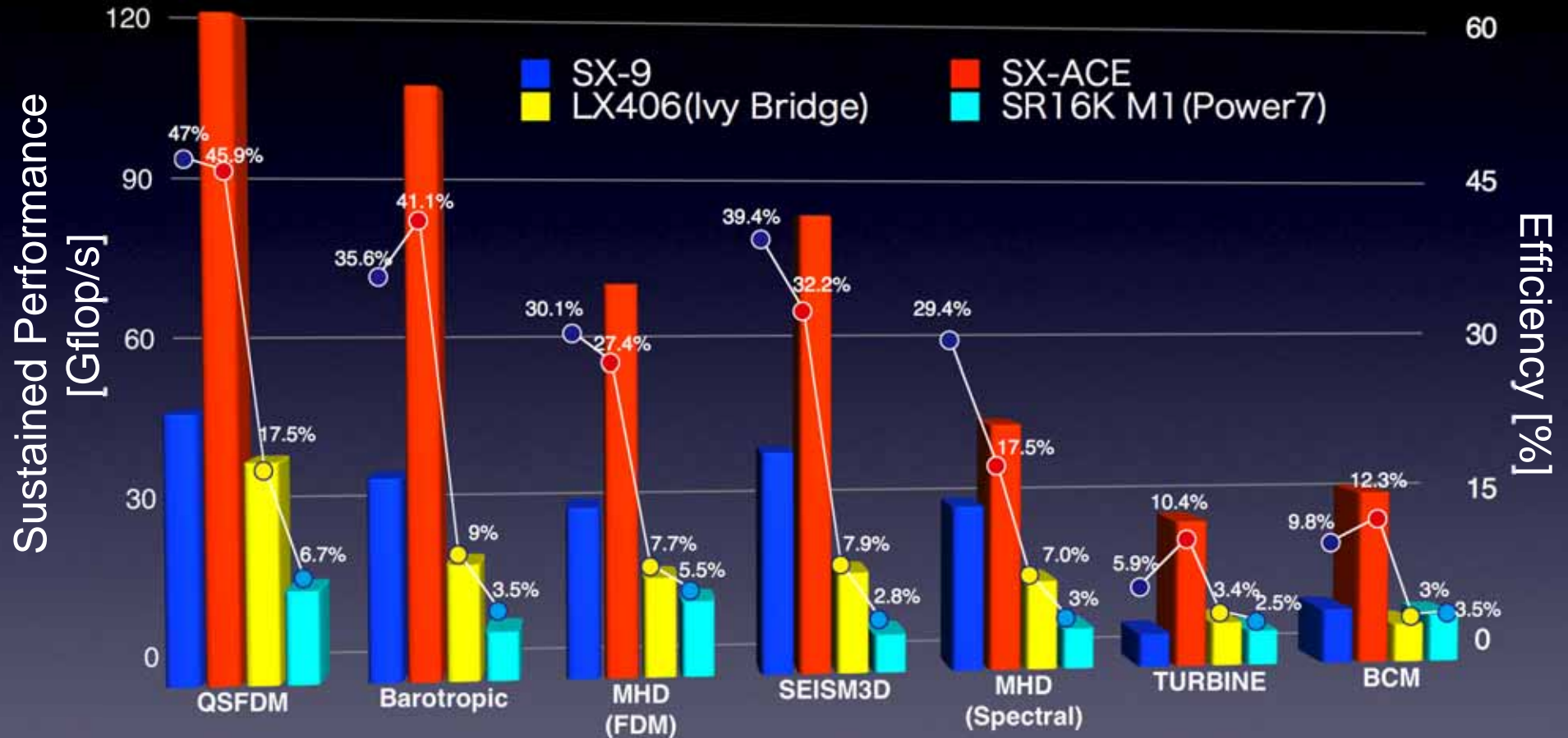
‡ for multi-node evaluation

Sustained Memory Bandwidth

- STREAM (TRIAD)

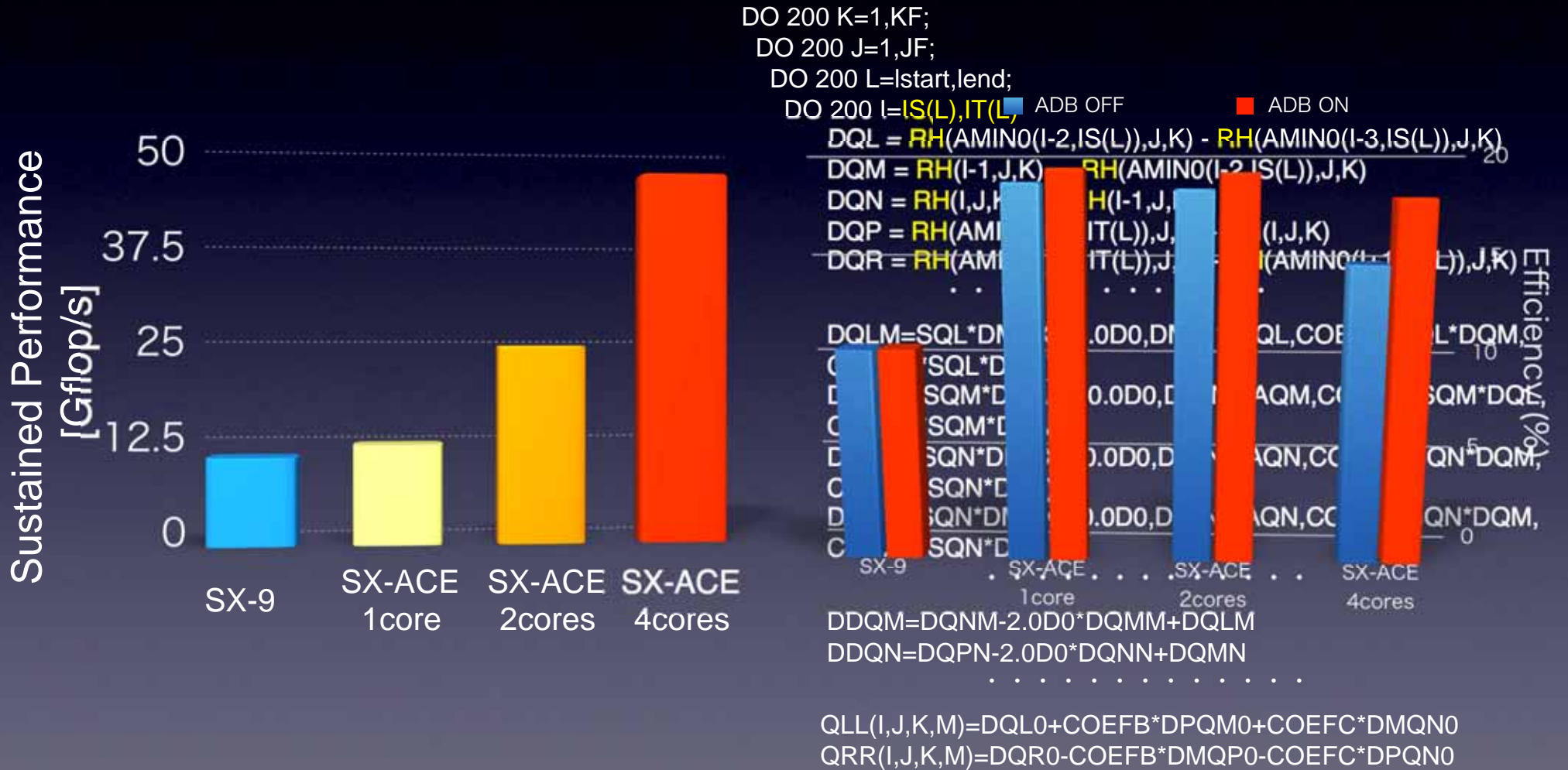


Sustained Single CPU Performance



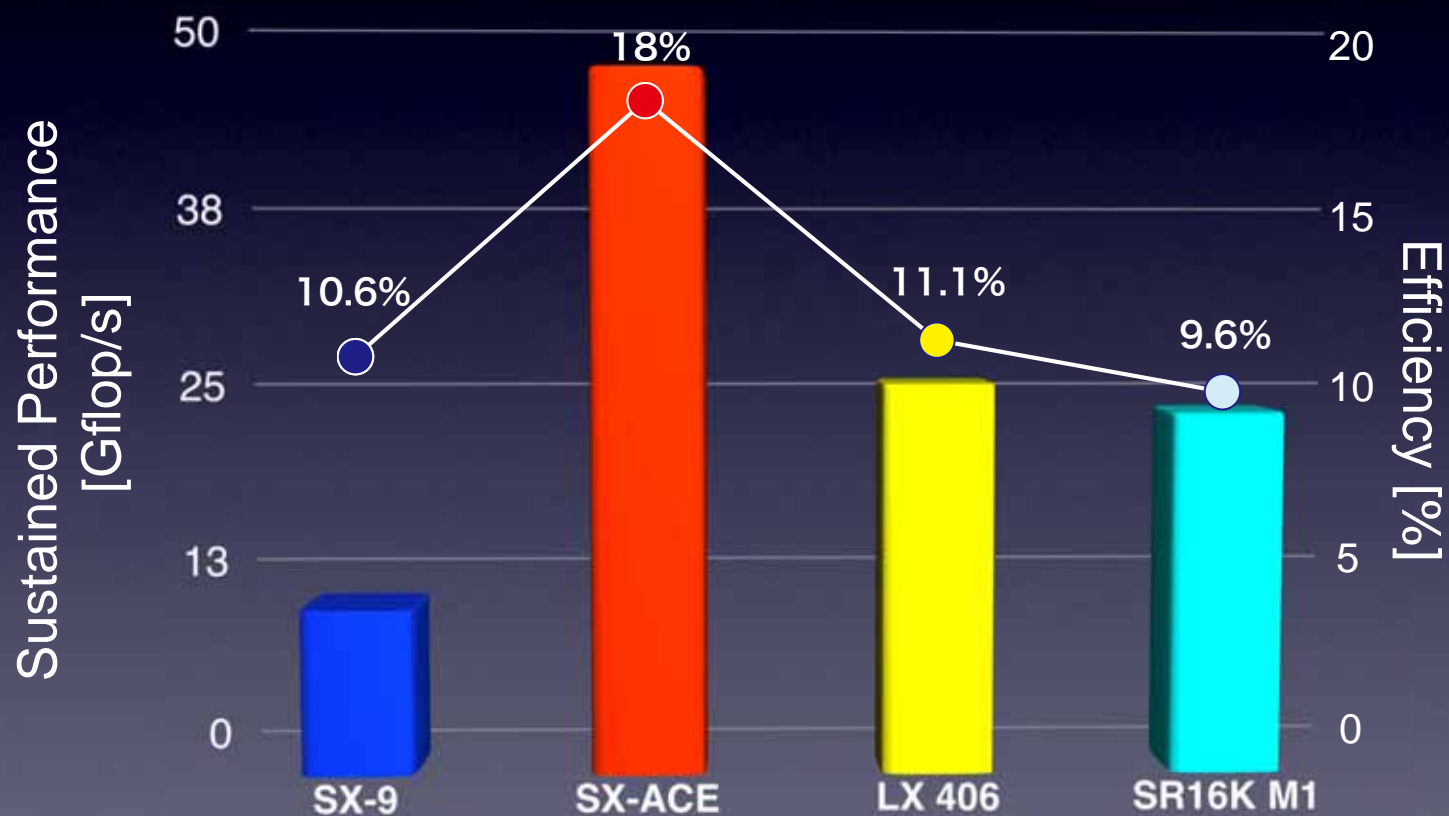
	QSFDM	Barotropic	MHD (FDM)	Seism3D	MHD (Spectral)	TURBINE	BCM
Code B/F							
Memory Intensity	2.16	1.97	3.04	2.15	2.21	1.78	7.01

Performance of Indirect Memory Accesses in TURBINE

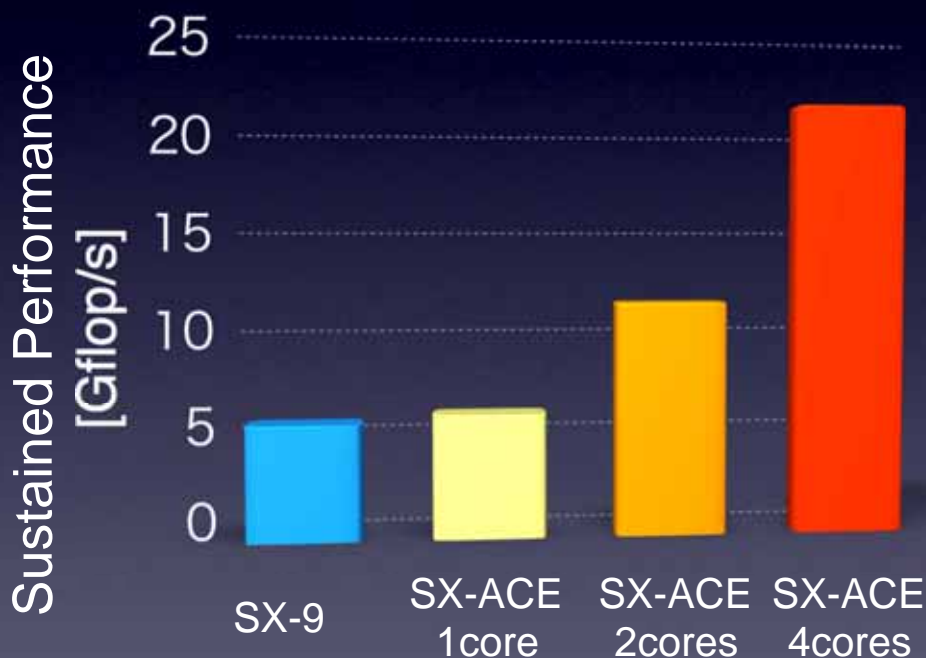


200 CONTINUE

Performance of Indirect Memory Accesses in TURBINE on Modern HPC Processors



Performance of Short-Vector Processing in TURBINE (1/2)

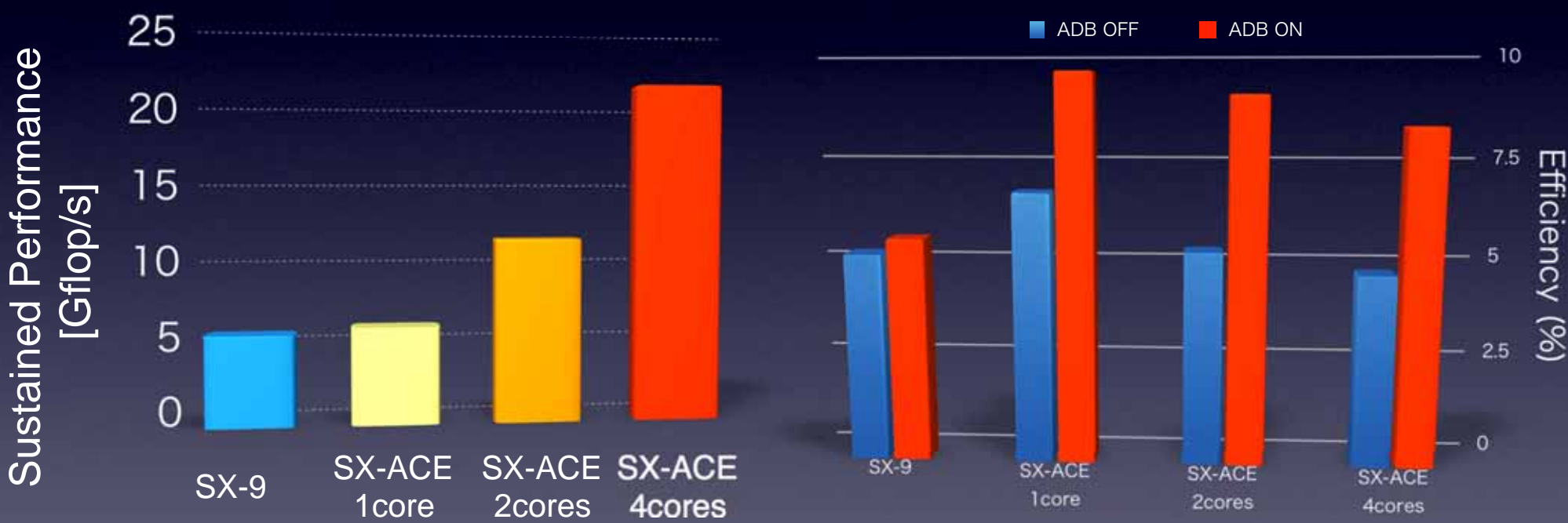


```

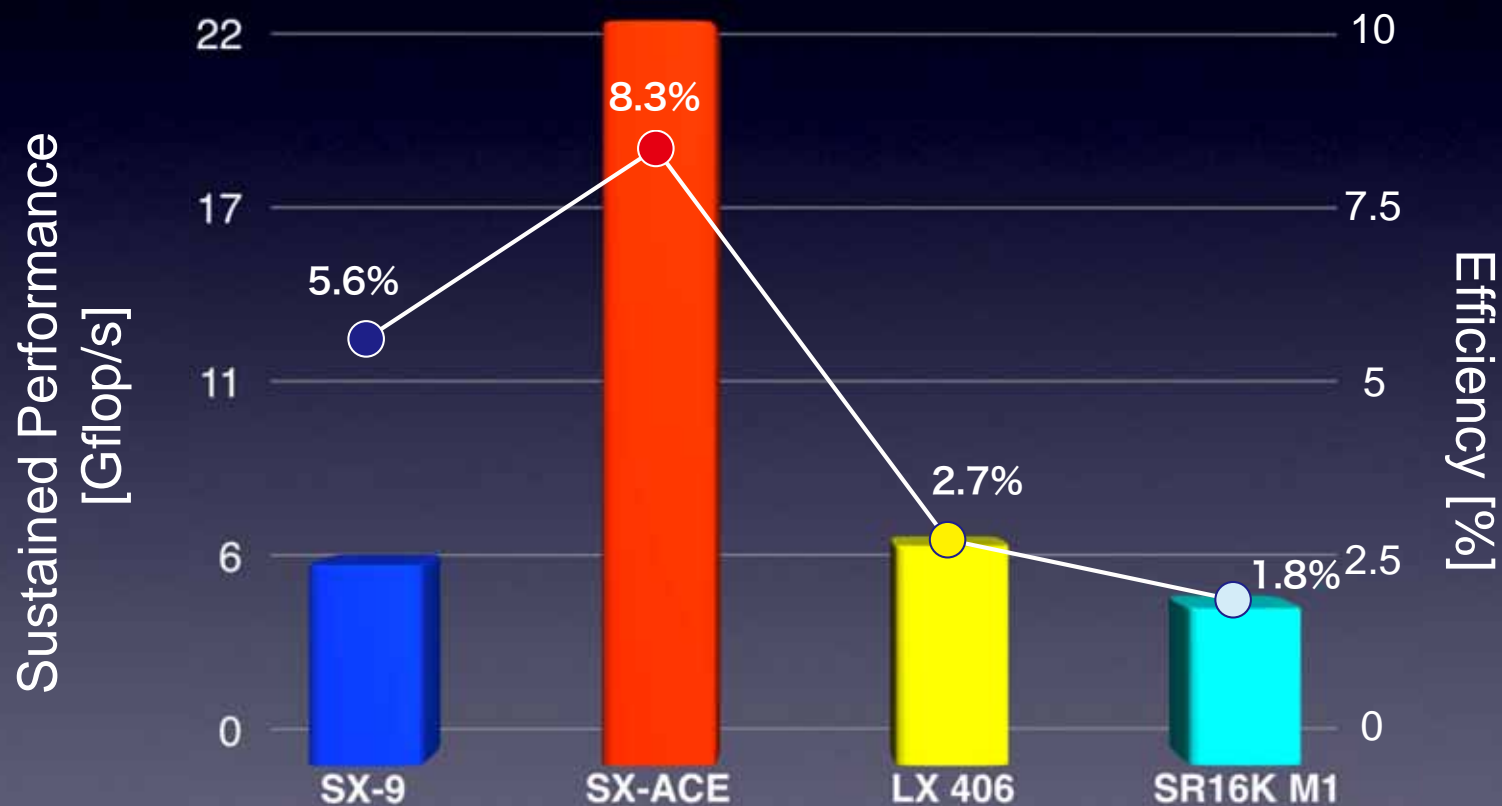
DO 10 M=MMIN,MMAX;
DO 10 K=KSTA,KEND;
DO 10 I=ISTA,IEND
  UC11=U(I-1,J,K,1)*XIX(I-1,J,K,1)
  &   +U(I-1,J,K,2)*XIX(I-1,J,K,4)
  &   +U(I-1,J,K,3)*XIX(I-1,J,K,7)
  UC22=U(I,J,K,1)*XIX(I,J,K,1)
  &   +U(I,J,K,2)*XIX(I,J,K,4)
  &   +U(I,J,K,3)*XIX(I,J,K,7)
  . . . . .
  AJXIX1=(AJR(I-1,J,k)*XIX(I-1,J,k,1)
  &       +AJR(I,J,k)*XIX(I,J,k,KL))*0.5D0
  AJXIX2=(AJR(II,JJ,kk)*XIX(II,JJ,kk,KL+3)
  &       +AJR(I,J,k)*XIX(I,J,k,KL+3))*0.5D0
  . . . . .
10 continue
  
```

Vector length 46

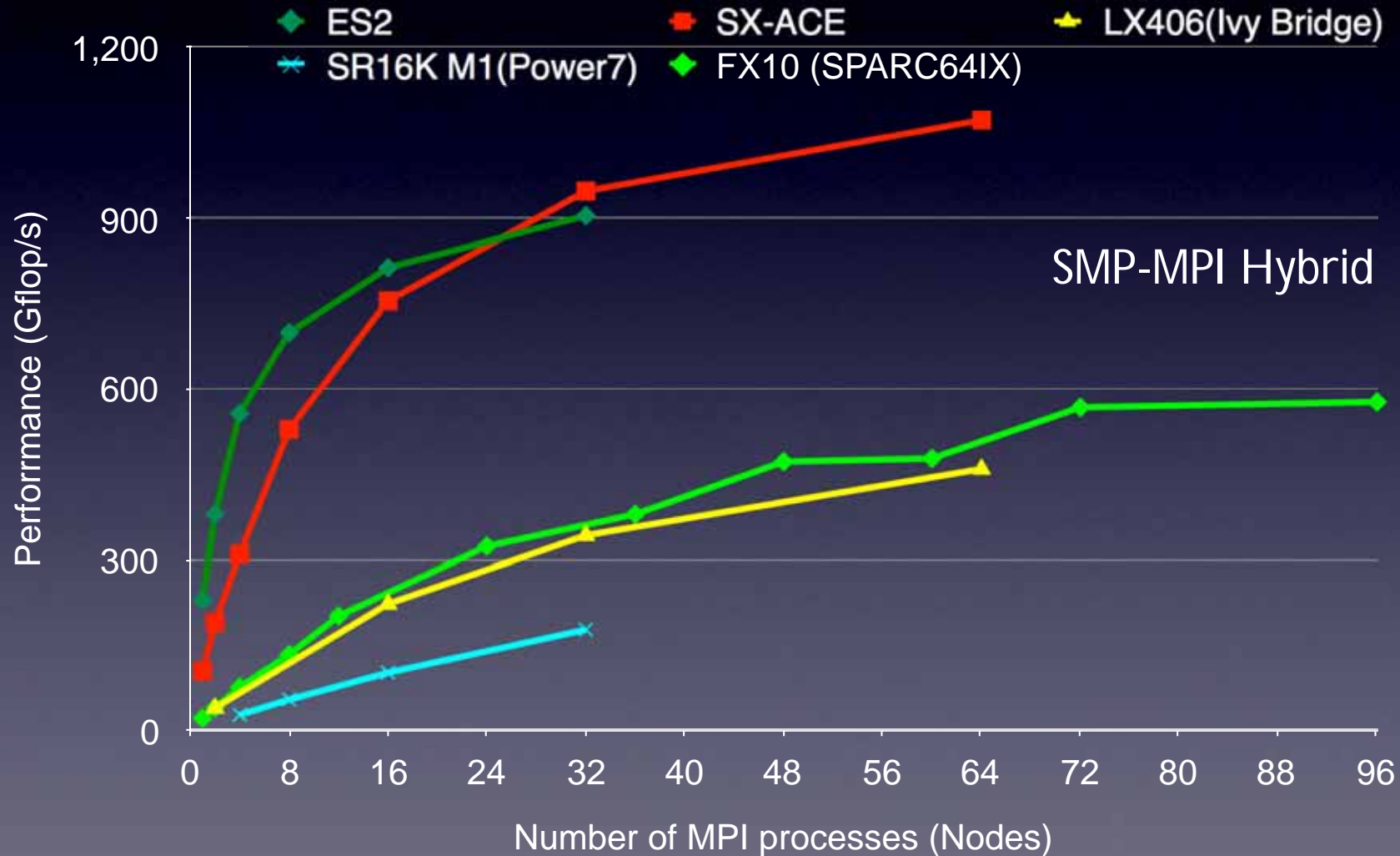
Performance of Short-Vector Processing in TURBINE (2/2)



Performance of Short-Vector Processing in TURBINE on Modern HPC Processors



Sustained Performance of Barotropic Ocean Model on Multi-Node Systems



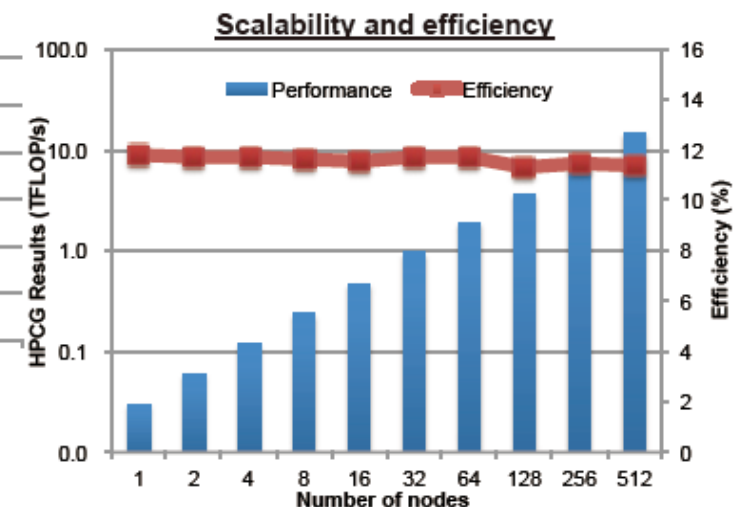
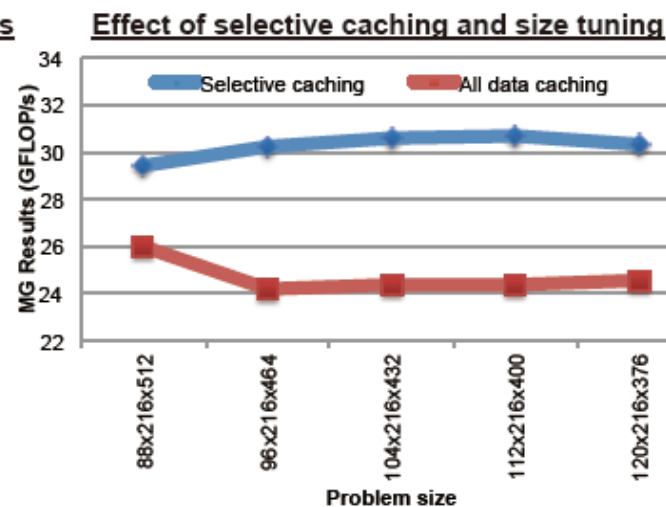
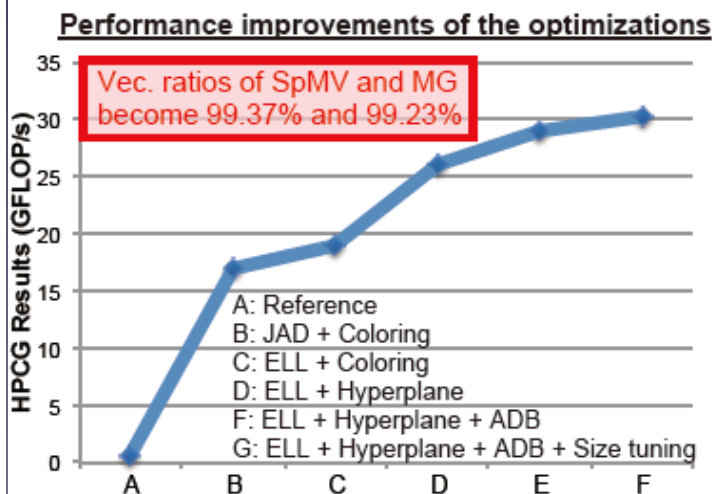
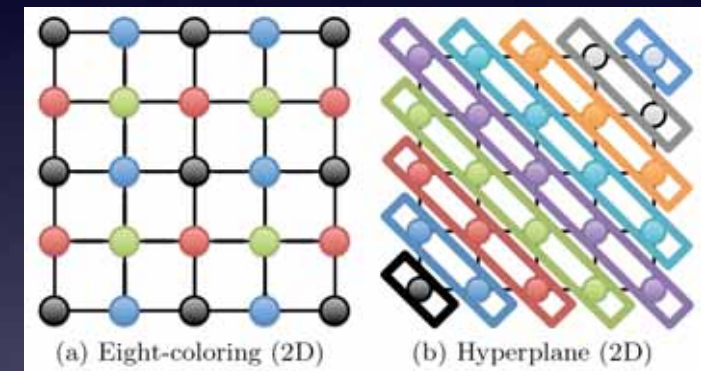
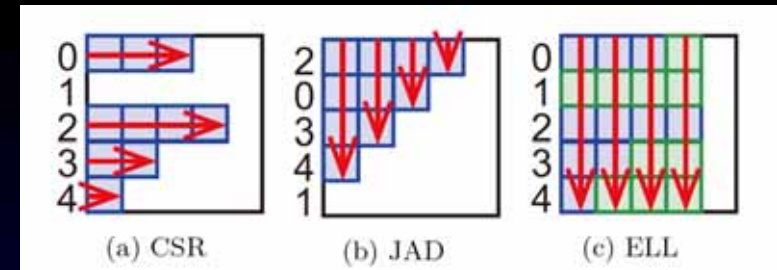
Performance Evaluation of SX-ACE by using the HPCG Benchmark

- ★ HPCG (High Performance Conjugate Gradients) is designed to exercise computational and data access patterns that more closely match a broad set of important applications,
 - ✓ HPL for top500 is increasingly unreliable as a true measure of system performance for a growing collection of important science and engineering applications.
- ★ HPCG is a complete, stand-alone code that measures the performance of basic operations in a unified code:
 - ✓ Sparse matrix-vector multiplication.
 - ✓ Sparse triangular solve.
 - ✓ Vector updates.
 - ✓ Global dot products.
 - ✓ Local symmetric Gauss-Seidel smoother.
 - ✓ Driven by multigrid preconditioned conjugate gradient algorithm that exercises the key kernels on a nested set of coarse grids.
 - ✓ Reference implementation is written in C++ with MPI and OpenMP support.

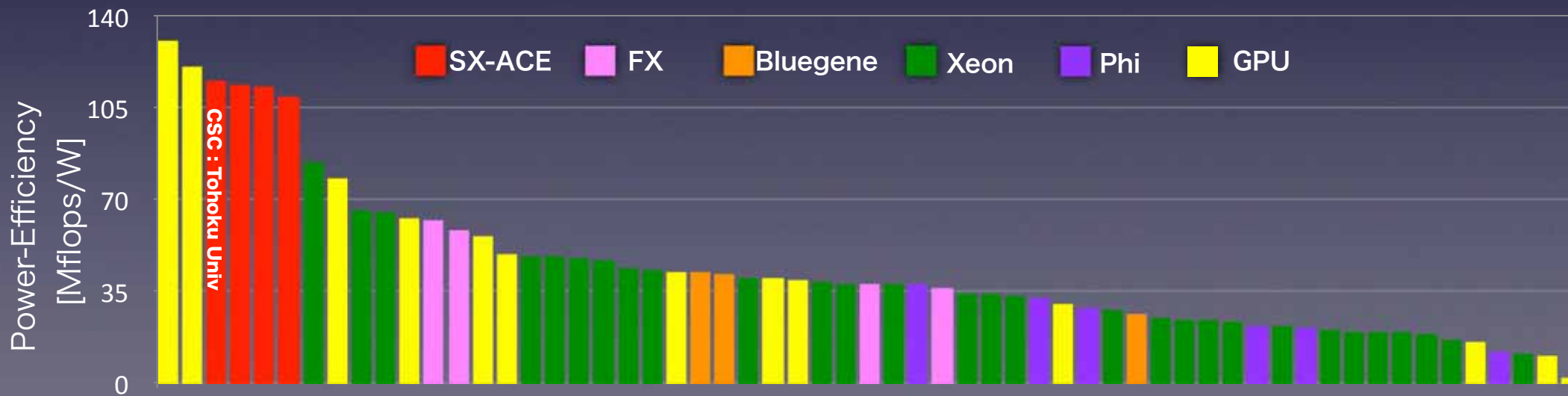
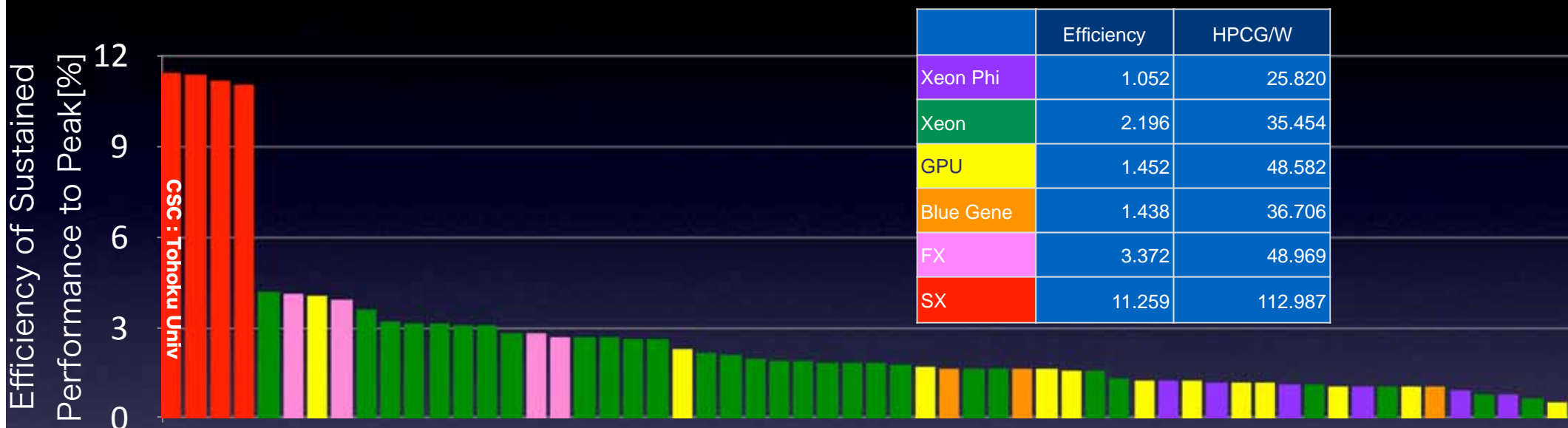
Optimizations of the the HPCG Benchmark for SX-ACE

*Komatsu et al.@SC15

- ★ Data packing for vector-friendly matrix memory allocation of sparse matrices
- ★ Parallelization by using coloring and hyperplane methods
- ★ Selective data caching and blocking for effective use of ADB

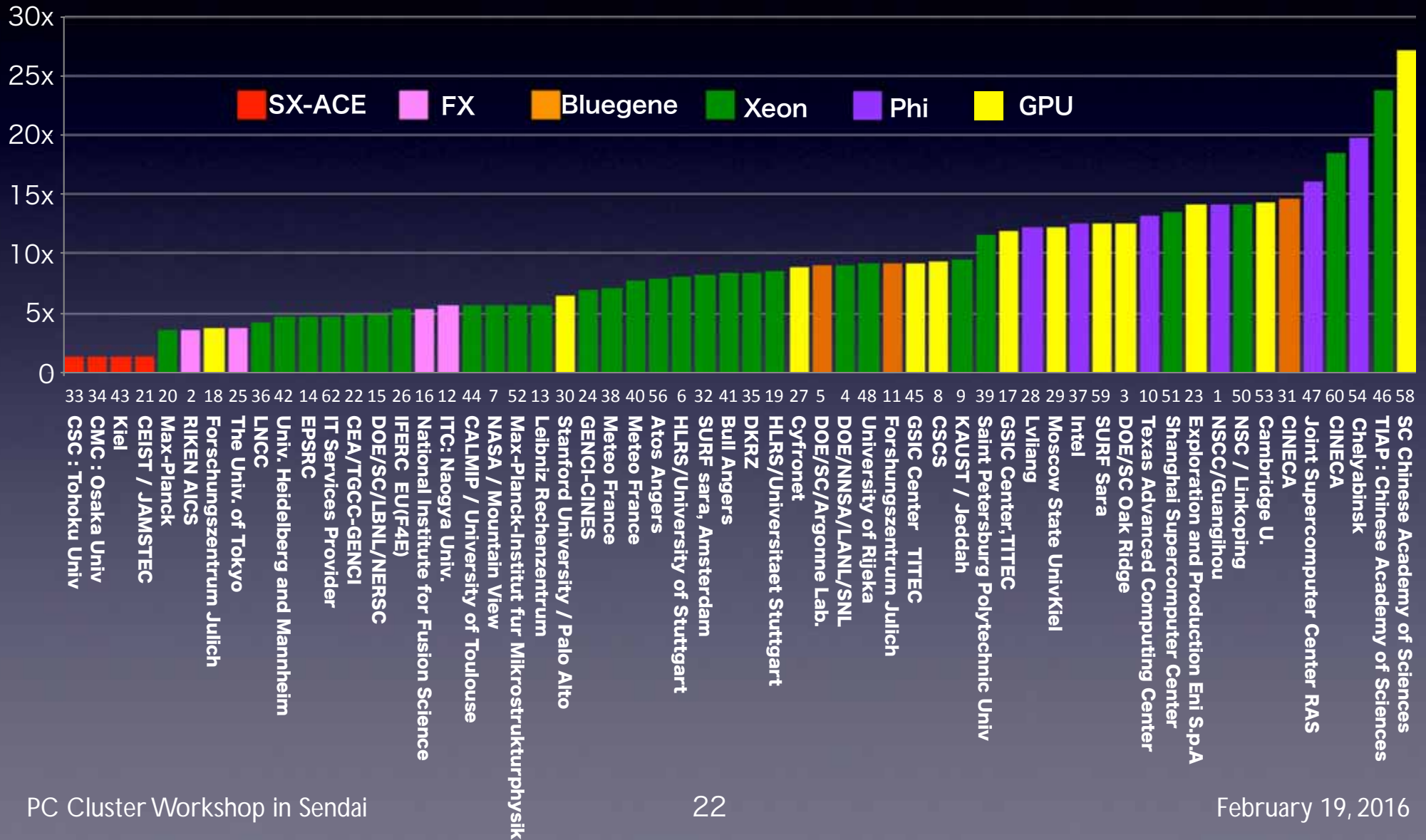


Efficiency Comparison in the HPCG Performance (1/2)



Efficiency Comparison in the HPCG Performance (2/2)

Peak Performance DOES NOT Track Observed Performance!





東北大学



Cyberscience
Center

R&D of A Real-Time Tsunami Inundation Forecasting System on SX-ACE



IRIDeS
International Research Institute
for Disaster Science
災害科学国際研究所



AOB **NEC**



KOKUSAI KOGYO CO., LTD.

Background: 2011 East-Japan Great Earthquake

- Main shock at 2:46pm, March 11, & huge Tsunami 30 min later...
- Magnitude 9.0, the Largest in Japan and the 5th largest in the world
- Around 20,000 victims (dead or missing), mainly due to Tsunami, 100,000 people evacuated to shelters in the first several months.
- A huge of debris of houses, cars and buildings remained in the coastal area of Tohoku over one year
- Important infrastructures such as gas/water/electricity/train/road are destroyed and/or stopped their services for one month in sendai city

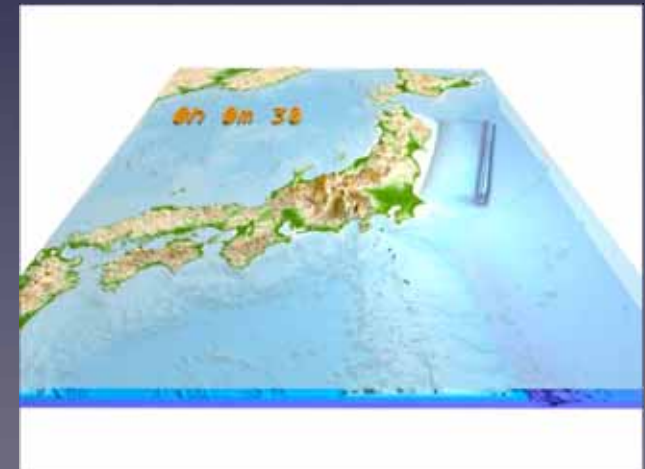
Sendai, Miyagi
(Hometown of Tohoku Univ)



Courtesy of Prof. Furumura,
U. of Tokyo



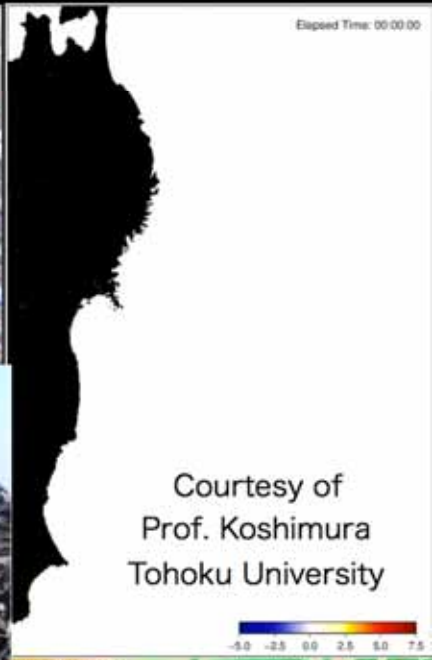
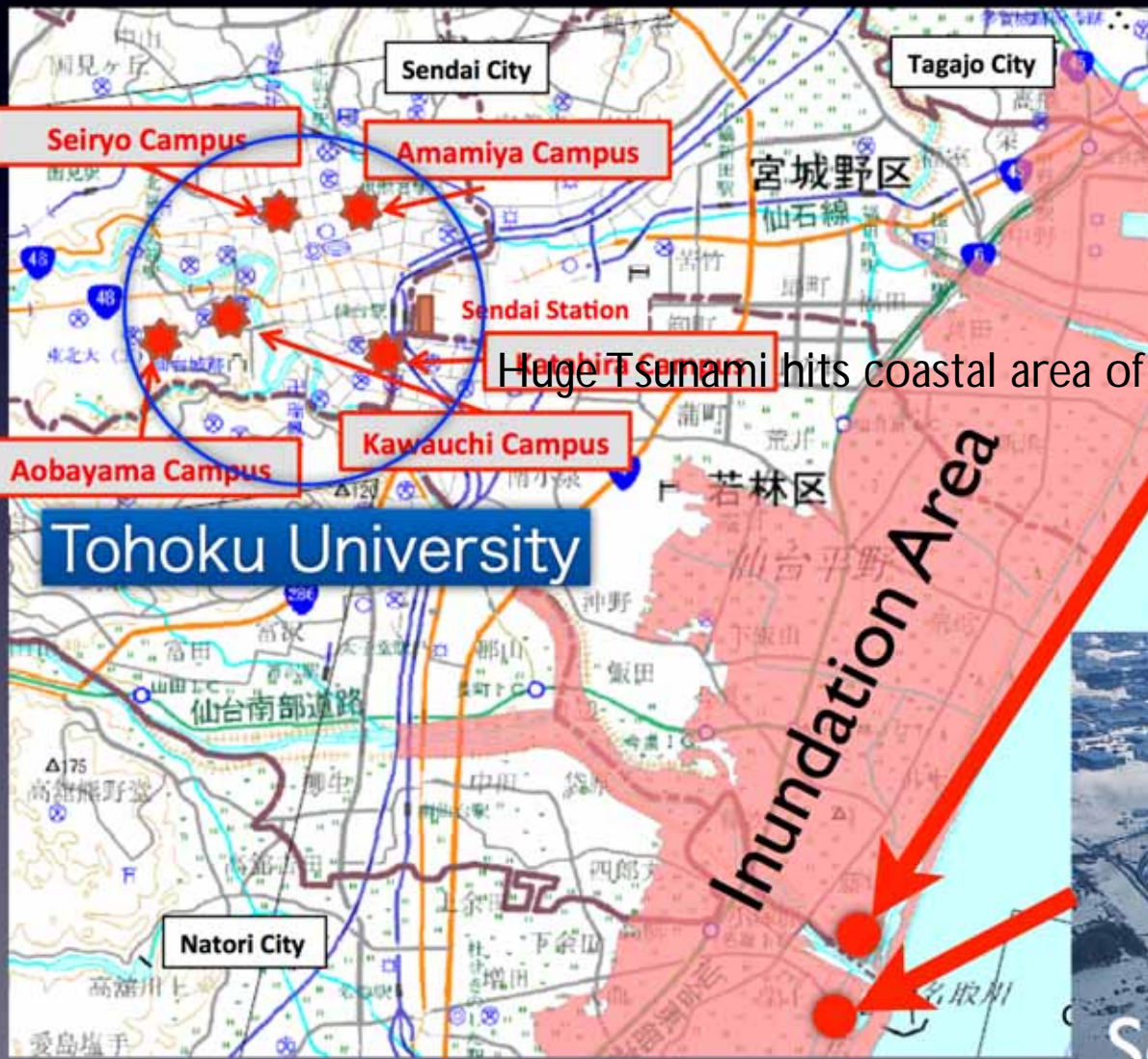
Sendai Station





Motivation:

Serious Damage to Sendai Area Due to 2011 Tsunami Inundation



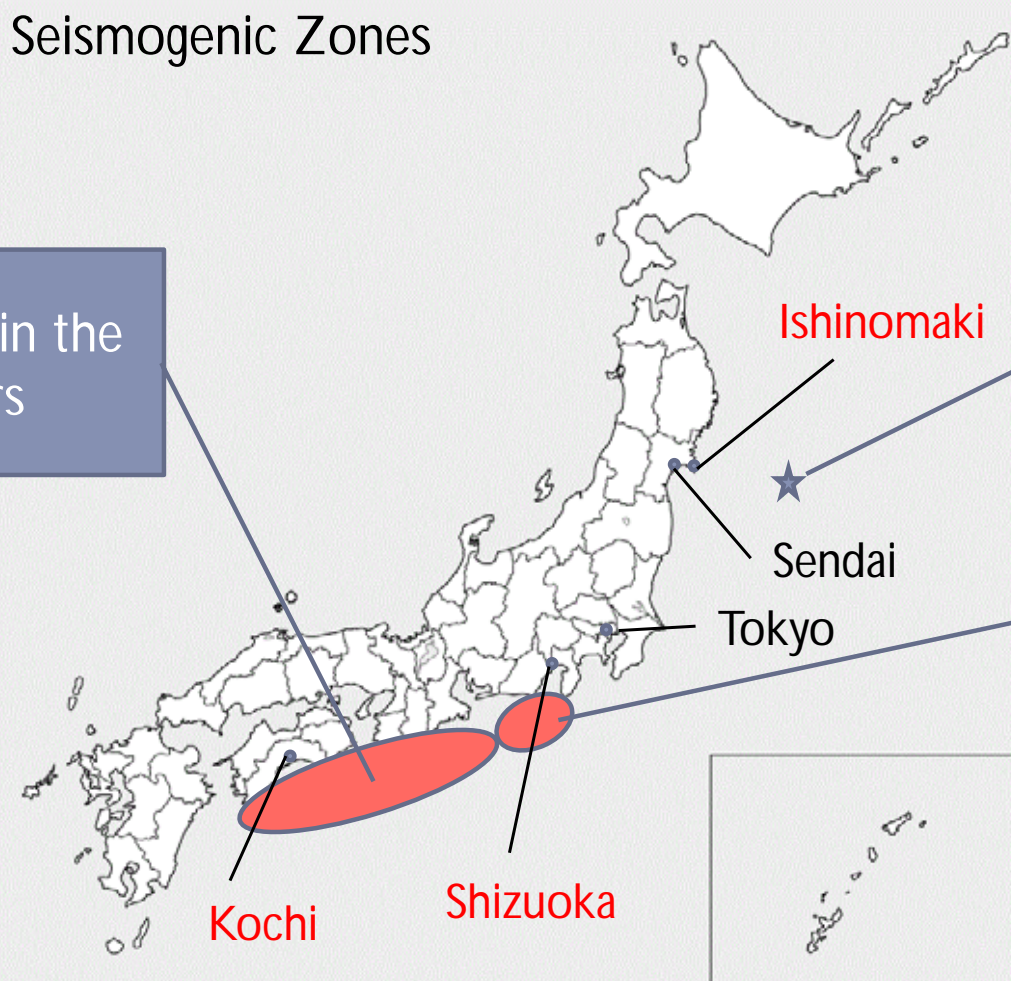
Huge Tsunami hits coastal area of Tohoku region

Courtesy of Prof. Koshimura Tohoku University

Sendai Airport

It's not End: High Probability of Big Earthquakes in Japan

- Japan may be hit by severe earthquakes and large tsunamis in the next 30 years



70 % probability in the next 30 years

The 2011 Great Tohoku Earthquake

88 % probability in the next 30 years

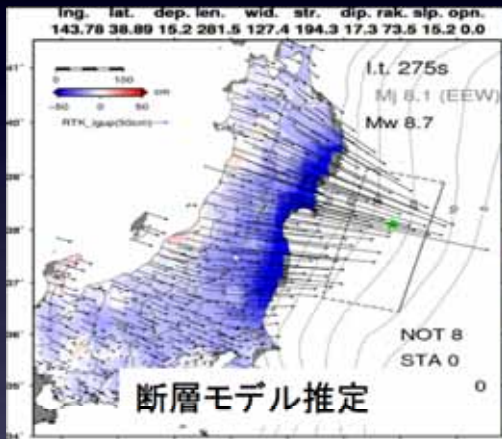
Objective of Our Work

Make HPC Available as a Social Infrastructure
for Homeland Safety in Japan!

- ★ **Prompt responses** to disaster to reduce damages such as warning evacuation from dangerous zones and rescuing survivors as soon as possible.
- ★ **Detailed and highly accurate analysis and forecasting** of Tsunami Inundation soon after the Big Earthquake is mandatory.
- ★ **Enhancement of social resiliency** against natural disasters by precise simulation using HPC to satisfy these demands

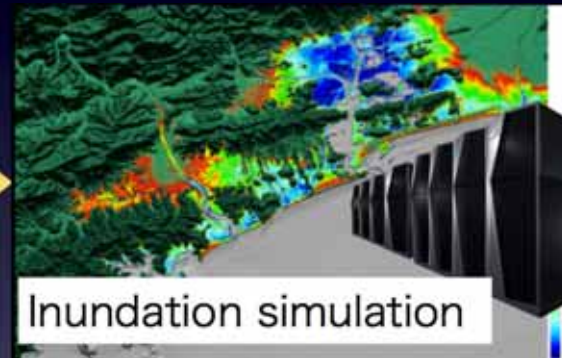
Design and Development of A Real-Time Tsunami Inundation Forecasting System

GPS-Observation



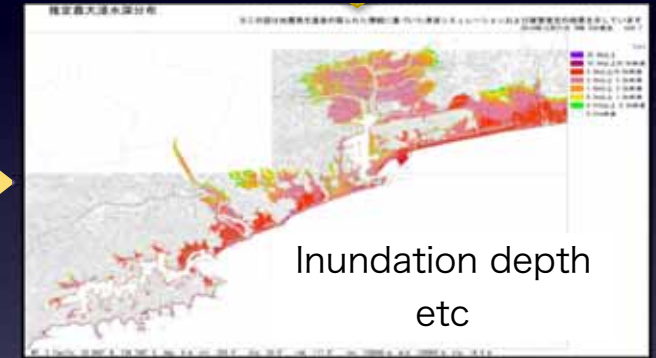
Fault estimation based on GPS data

Simulation on SX-ACE



10-m mesh models of coastal cities

Information Delivery



Just-In-Time access of Visualized information by local governments

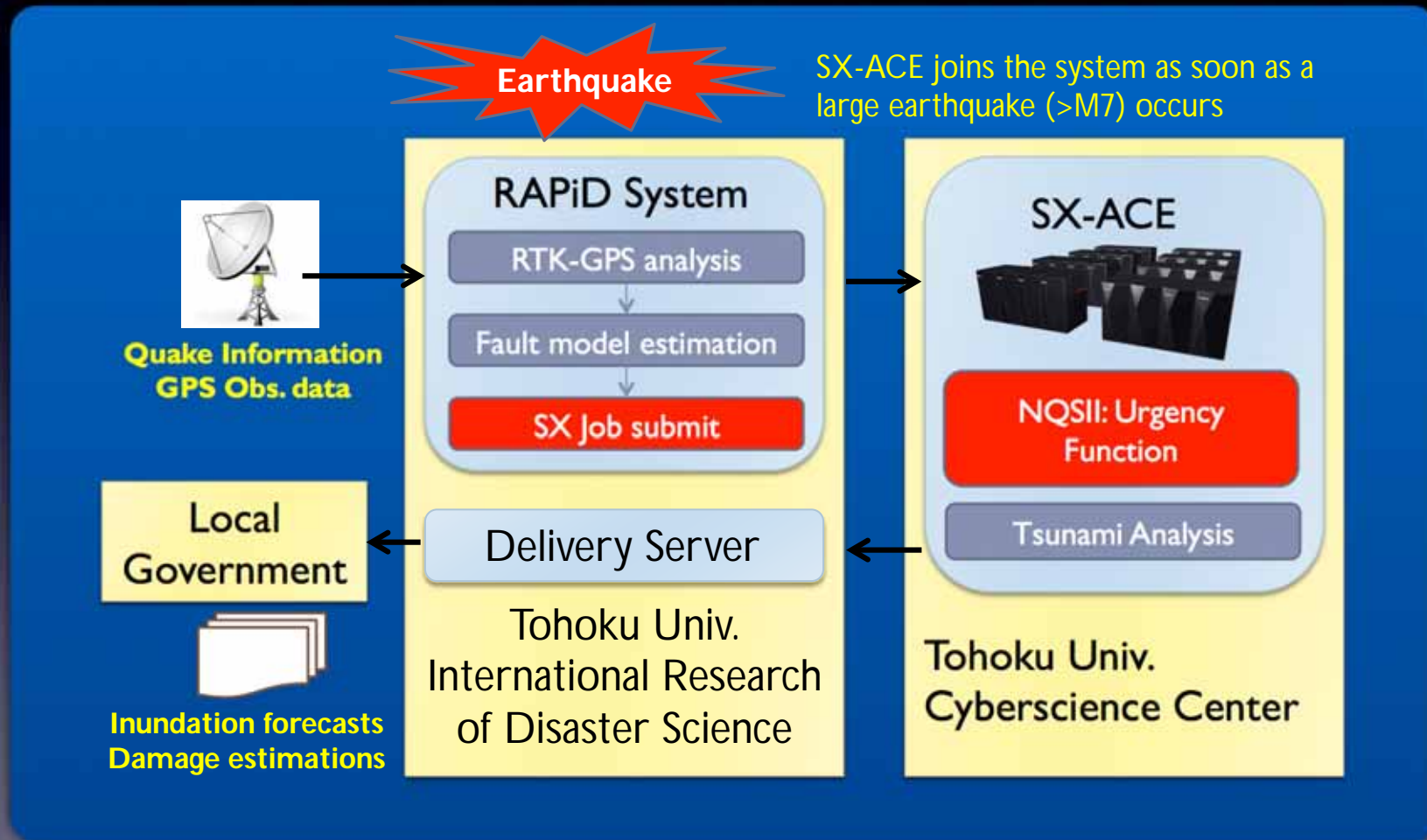
< 8 min

< 8 min

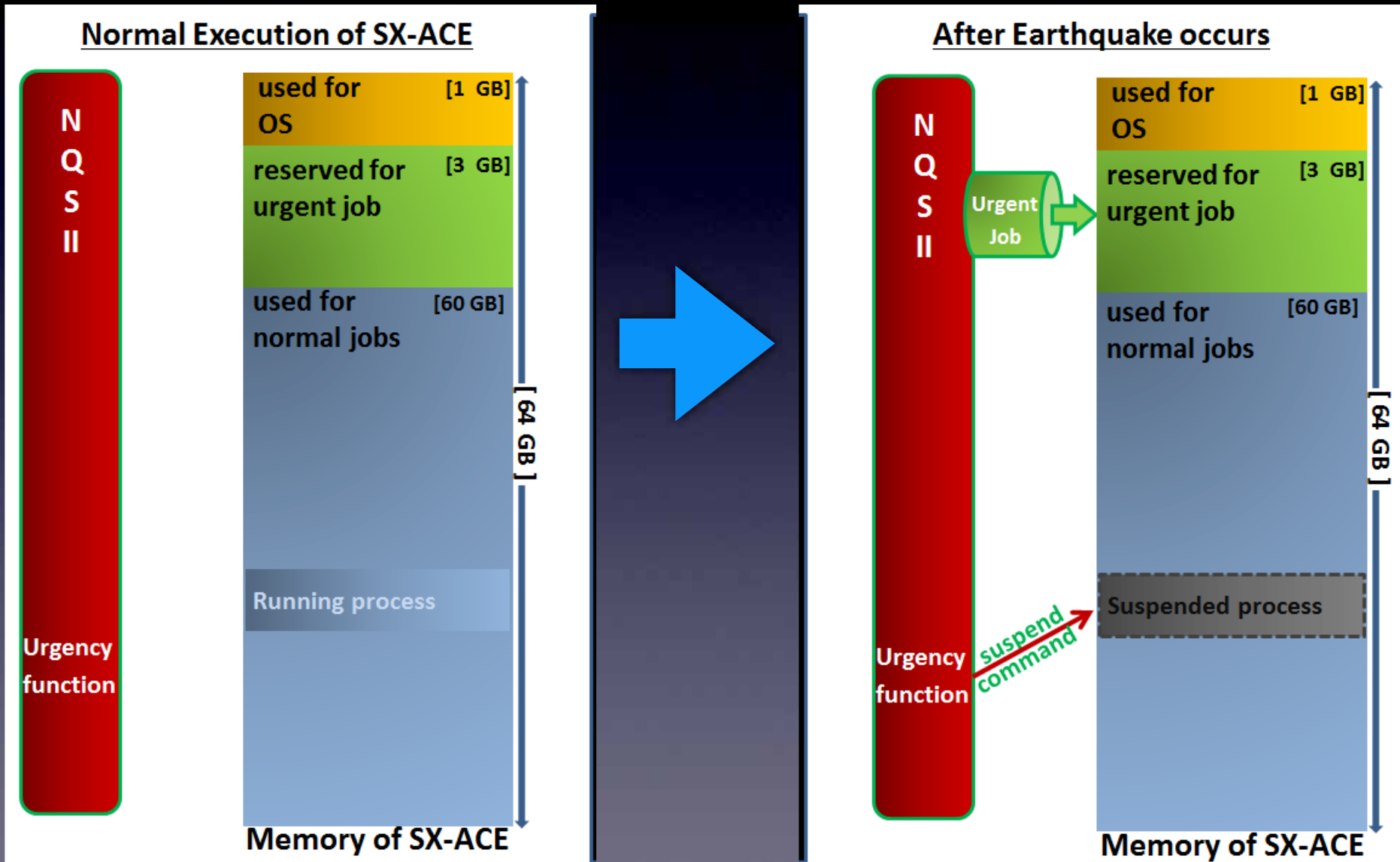
< 4 min

< 20 min

System Organization



Emergency Job Handling by NQSII of SX-ACE



Simulation: Target Code & Areas

★ Target Code TUNAMI: Tohoku University's Numerical Analysis Model for Investigating Tsunami

- Developed by Prof. Koshimura of Tohoku University
- Authorized by UNESCO and Japanese Government

★ Governing Equations

- Non-Linear Shallow Water Equations

★ Numerical Scheme

- Staggered Leap-Frog Finite Difference Method

★ Memory-intensive application

- B/F = 1.82 (single precision)

🌐 Target Areas: Miyagi, Shizuoka & Kochi

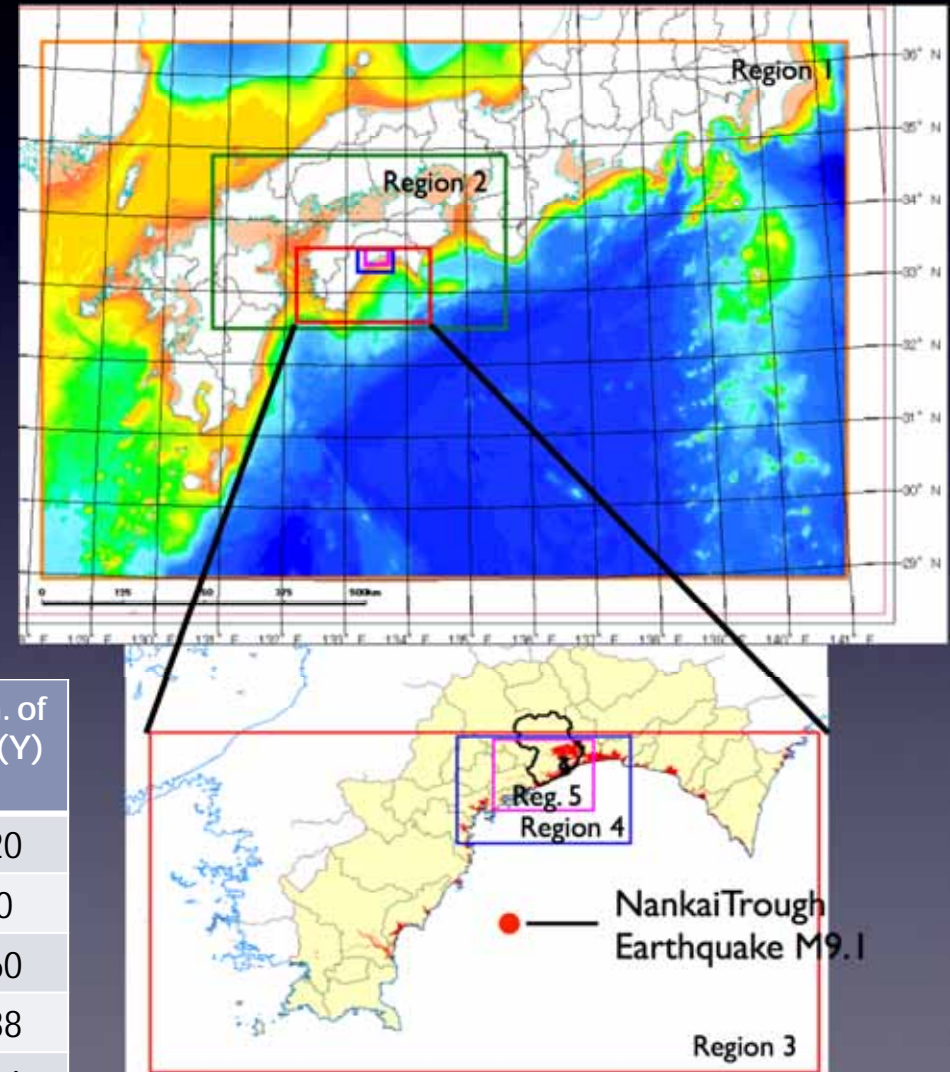


Computation Domain

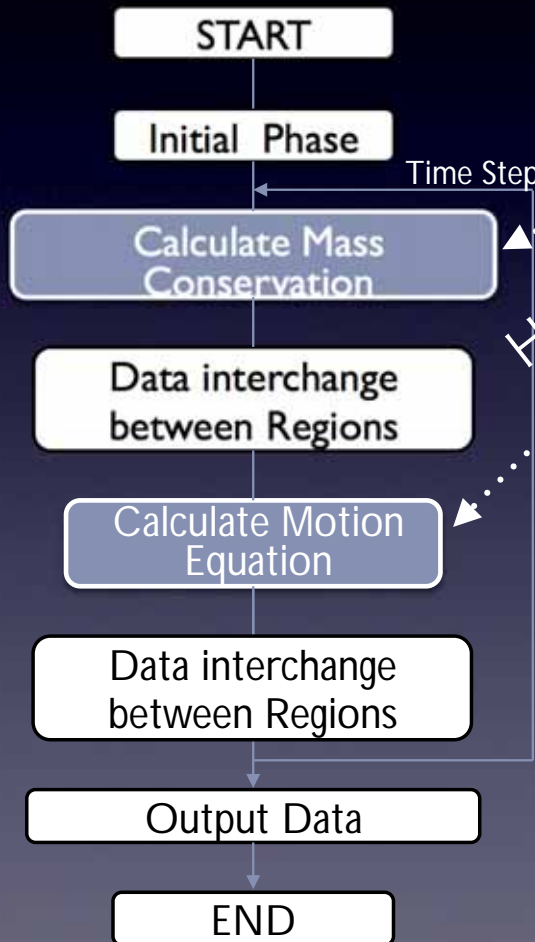
- ★ Hierarchical multi-level grid models
- ★ Computation Domain of Kochi City:
 - 1244km x 826km
 - 5 nested grids
 - 6 hours of Tsunami Inundation

✓ $\Delta t = 0.1$ sec.

Region	Grid Size (m)	Num. of Grid(X)	Num. of Grid(Y)
1	810	1536	1020
2	270	1680	990
3	90	2292	1260
4	30	1782	1188
5	10	3504	2364



Program Structure



▶ Doubly nested loops

```

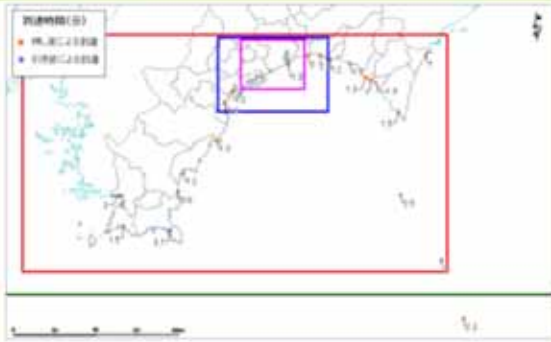

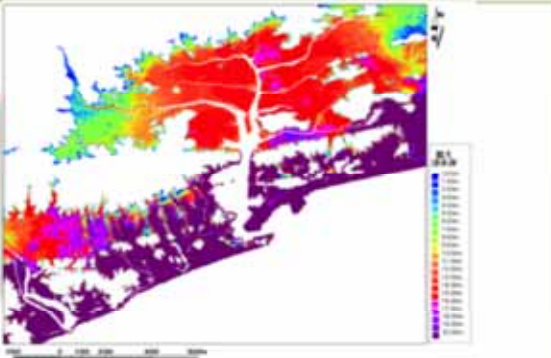
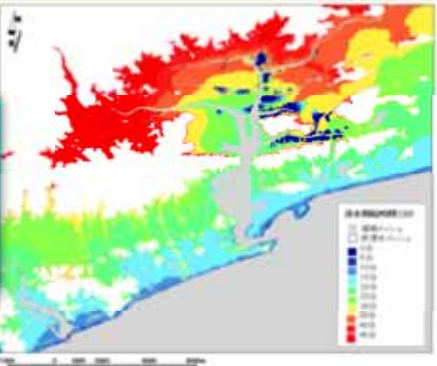
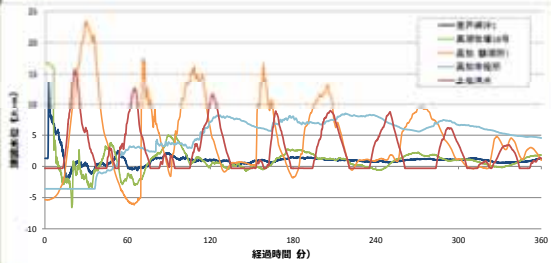
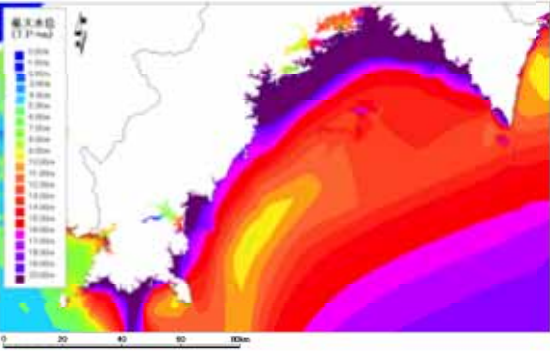
DO J=2, JF (Latitude) ← Parallelized
  DO I=2, IF (Longitude) ← Vectorized
    .....
    ZZ = Z(I,J,I) - RX*(M(I,J,I)-M(I-1,J,I))
           - RY*(N(I,J,I)-N(I,J-1,I))
    .....
  END DO
END DO
  
```

Stencil

▶ Tuning

- ▶ Inlining subroutines
- ▶ Optimization of I/O routines
- ▶ Vectorization & Parallelization
 - ▶ Vect. Ratio=99.6%, Vect. Length=235
- ▶ ADB Tuning of Stencil kernels

Visualization of Simulation Results by Delivery Server

優先順位	項目	イメージ	優先順位	項目	イメージ
1	Tsunami Arrival Time		4	Damage Estimation (damaged population, houses, buildings)	
2	Maximum Inundation Depth		5	Inundation Start Time	
3	Tsunami Level Change		6	Maximum Water Level	

The information is delivered to Local Governments through the Web



Real-Time Tsunami Inundation Forecasting

edit6.a-2.co.jp

TSUNAMI Simulator EEW Transfer

Rapid coseismic fault determination system for real-time tsunami inundation forecasting [This is a test based on expected events]

Lapse time from the earthquake: 420 Seconds

Display Full Range (Miyagi:RED Shizuoka:BLUE Kochi:GREEN)
 Miyagi(Ishinomaki and Higashi-Matsushima) Shizuoka Kochi

No.	Type	Epicenter & Mechanism	Select
1	EEW	North-eastern Japan, Mjma 7.0 / Interplate	<input type="checkbox"/>
2	RAPiD	North-eastern Japan, Mjma 9.2 / Interplate	<input type="checkbox"/>
3	EEW	North-eastern Japan, Mjma 8.5 / Outer-Rise	<input type="checkbox"/>
4	GRiD MT	North-eastern Japan, Mw 7.8 / Shallow Offshore	<input type="checkbox"/>
5	EEW	South-western Japan, Mjma 8.3 / Interplate	<input type="checkbox"/>
6	EEW	South-western Japan, Mjma 8.7 / Interplate	<input type="checkbox"/>

Server connection status: Connected

- GREEN CIRCLE denotes P-wave arrival area in the map.
- RED CIRCLE denotes S-wave arrival area in the map.
- P-wave & S-wave circle disappears after 120 seconds.
- Estimated optimum fault plane appears, in WHITE rectangle, after 420 seconds.
- Earthquake detail appears by clicking marker on the map.
- Once you click the "START" button, you can not stop the procedure.
- "CANCEL" button only stops updating EEW information.
- [日本語表記](#)

© 2015 A2 Corporation.

Demo: Visualization of Simulation Results

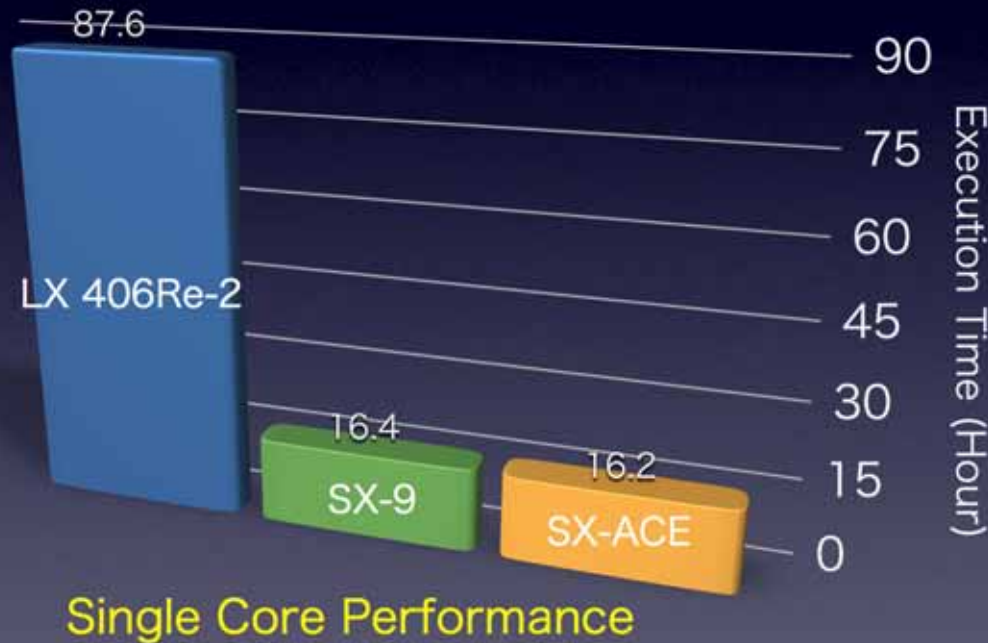
Simulation Results of Inundation of Kochi City Caused by Nankai Trough Earthquake

0 Hour 0 M 10 S



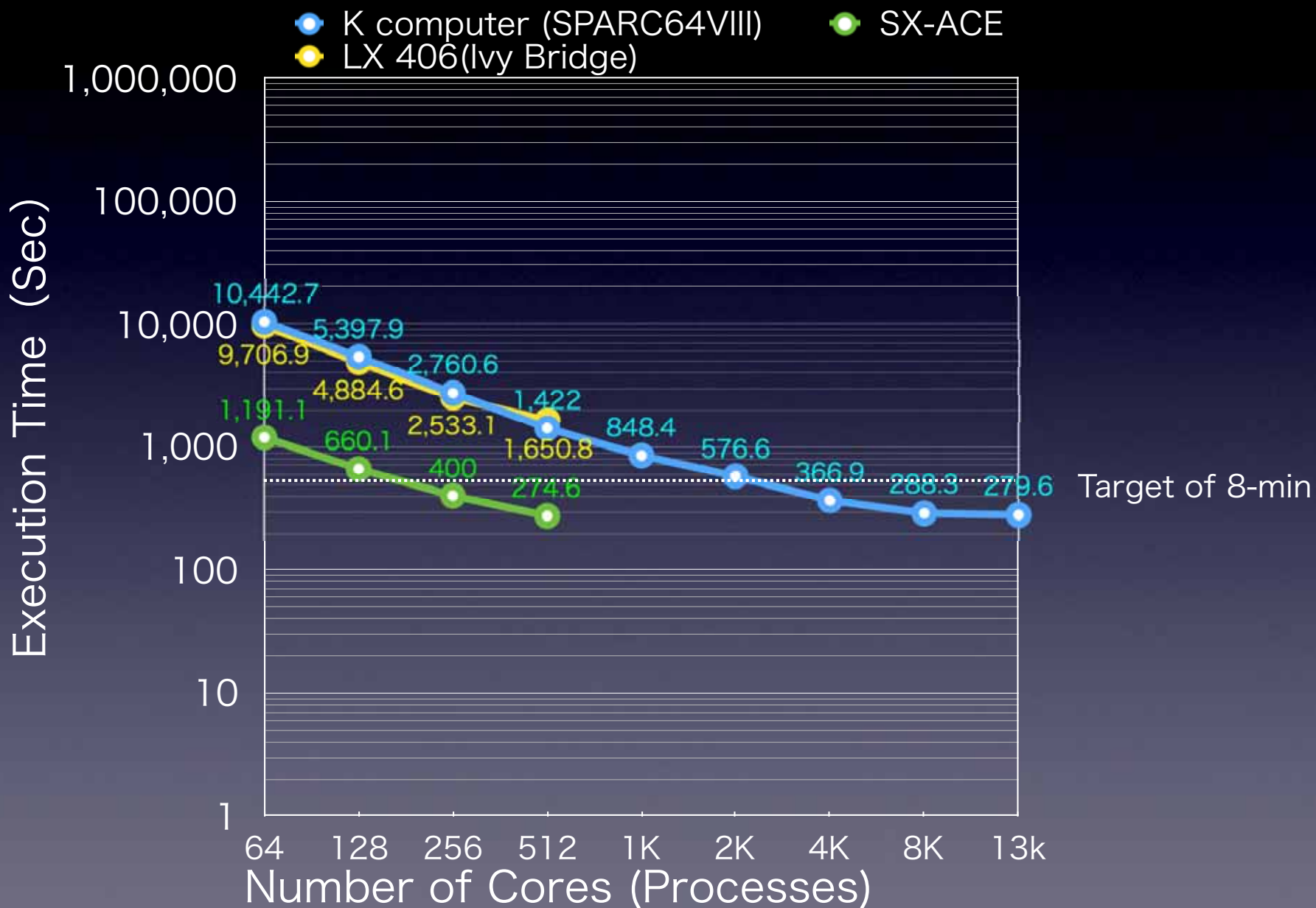
Performance of Tunami Code on SX-ACE

5.5x performance improvement against LX
(peak performance ratio is only 3x)

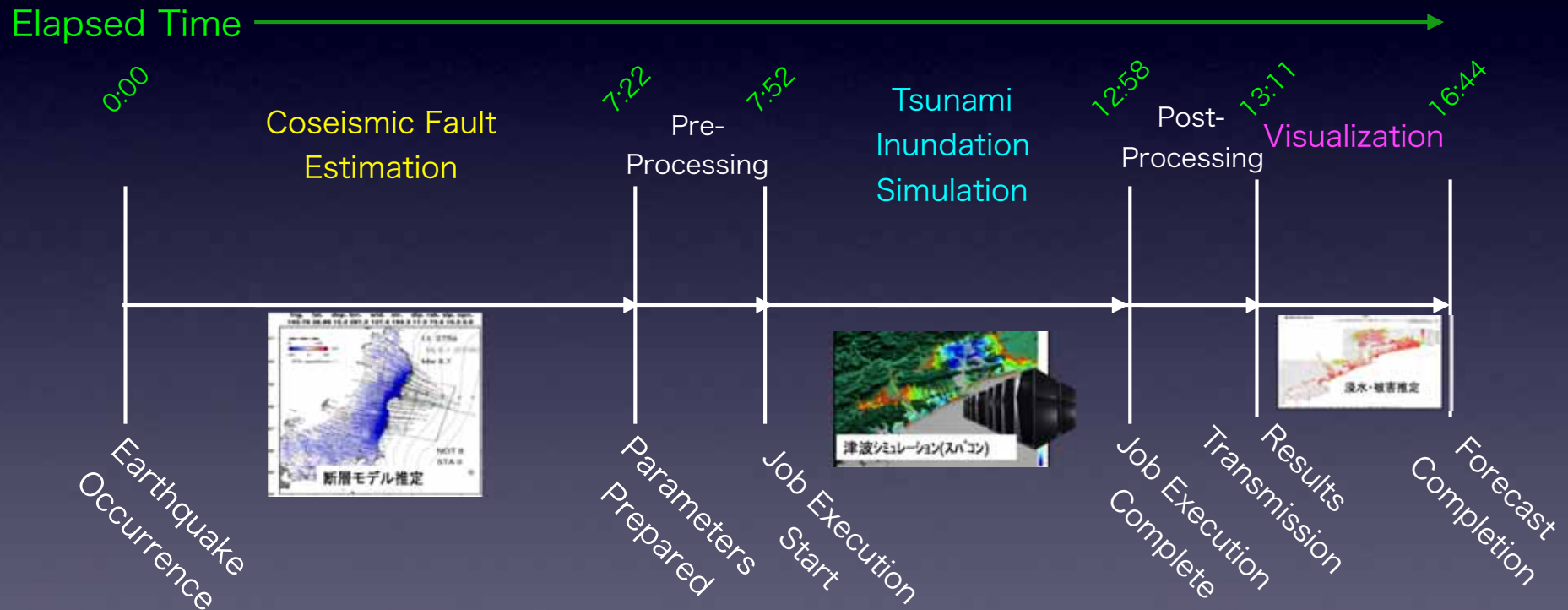


System	Perf. / Socket (Gflop/s)	No. of Cores	Perf. / Core (Gflop/s)	Mem. BW (GB/s)	Socket B/F	Core B/F
SX-ACE	256	4	64	256	1	4
SX-9	102.4	1	102.4	256	2.5	2.5
LX406Re-2	230.4	12	19.2	59.7	0.26	N/A

Scalability of Tunami Code



Total Execution Time of Tsunami Forecasting Workflow (Kochi-City Case)



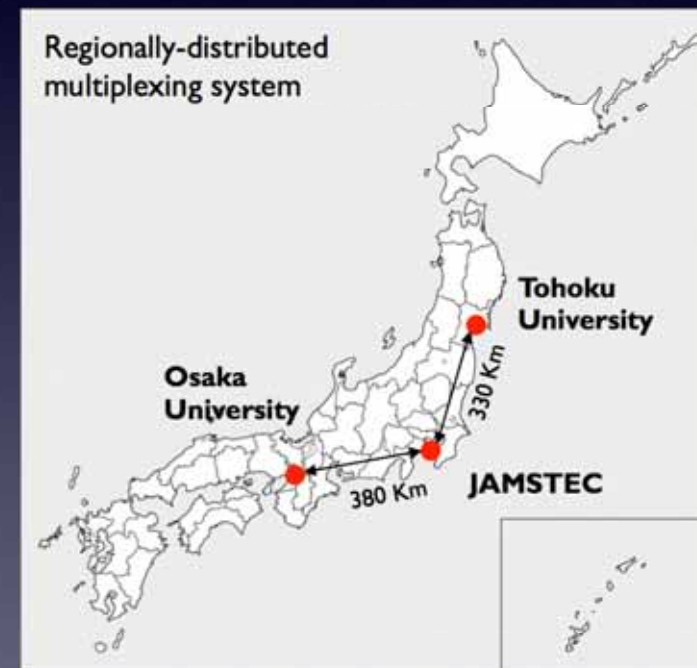
Summary

- ★ SX-ACE is involved as the social infrastructure for Tsunami Inundation forecasting, like a weather forecasting system, in addition to the research infrastructure for computational science and engineering in Japan

- ★ Current work



Target Area Extension:
Full coverage of Japan



System Extension:
Complemental operations
of multiple systems

Summary

- ★ **SX-ACE shows high sustained performance compared with SX-9, in particular a significant improvement in short-vector processing and indirect memory accesses**
 - ✓ achieved the same single core performance in practical applications even with 60% of peak performance
 - ✓ No1. computing-efficiency and power-efficiency in the HPCG Benchmark ranking
 - ✓ Pave the way to a new social infrastructure for homeland safety in Japan
- ★ **Well balanced HEC systems regarding memory performance is the key to success for realizing high productivity in science and engineering simulations**
 - ✓ Demands for Supercomputers for the rest of us, especially for 2020 and beyond!
 - ✓ Brute force to Smart Force in HPC design
 - ✓ Quality, not Quantity!

WSSP開催案内



21st WSSP in Sendai

- 23rd Workshop on Sustained Simulation Performance
 - Held on March 16-17, 2016 at Tohoku University, Sendai Japan
 - Organized by Tohoku University and HLRS, Stuttgart, JAMSTEC, NEC
 - International researchers and engineers get together to discuss and exchange ideas, experience and perspectives on current and future HPC technologies
 - Confirmed invited Speakers
 - Michael Resch (HLRS)
 - Sabine Roller (University of Siegen)
 - Vladimir Voevodin (Moscow State University)
 - Toshimitsu Yokobori (Tohoku Univ)
 - Mitsuo Yokokawa (Kobe Univ)
 - Akiko Matsuo (Keio Univ)
 - Ken-ichi Itakura (JAMSTEC)
 - and more!
 - <https://www.sc.cc.tohoku.ac.jp/wssp23/index.ja.html>

