The next-generation supercomputer and NWP system of JMA

Masami NARITA, Keiichi KATAYAMA

Numerical Prediction Division,
Japan Meteorological Agency

Contents

- JMA supercomputer systems
  - Current system (Mar 2006 ~ Feb 2012)
  - Next system (Mar 2012 ~)
  - Current, next and future NWP models

- Japanese Petascale Computing Projects
  - The “K Computer” (RIKEN)
  - TSUBAME 2.0 (Tokyo-tech)

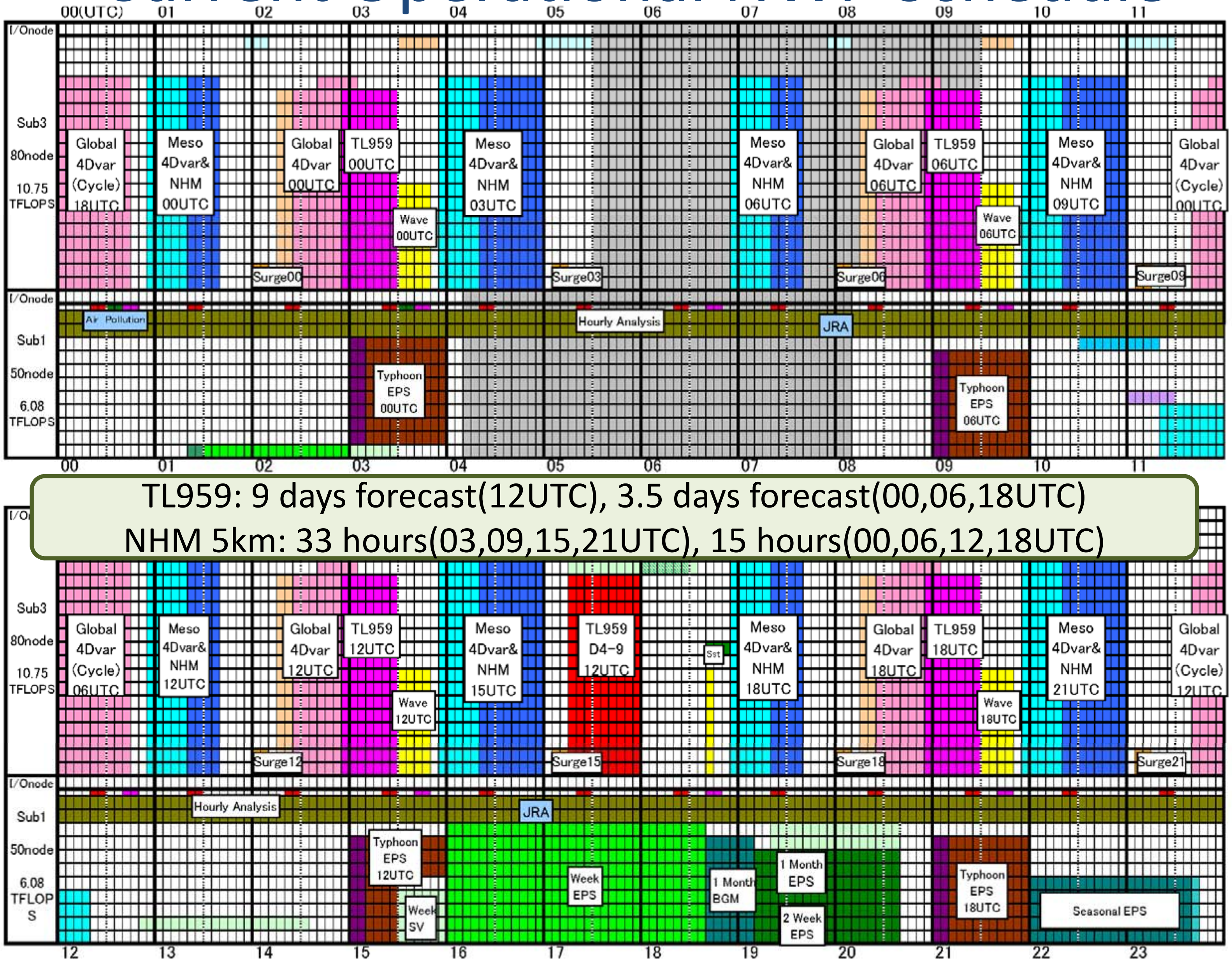
JMA SUPERCOMPUTER SYSTEMS

This picture is a view from Meteorological Satellite Center, Kiyose City, Tokyo (JMA HPC site)

Current JMA Supercomputer (Mar 2006 – Feb 2012)

- **HITACHI SR11000 (POWER5+)**
  - 3 subsystems, Total peak = **27.5TFLOPS**
    - **Sub1**:SR11000/J1 **50-nodes 6.08TFLOPS**
    - **Sub2**:SR11000/K1 **80-nodes 10.75TFLOPS**
    - **Sub3**:SR10000/K1 **80-nodes 10.75TFLOPS**
  - Memory: 13.1TB
  - Storage: 36.2TB
  - Tape Library: 2PB

Current Operational NWP Schedule



TL959: 9 days forecast(12UTC), 3.5 days forecast(00,06,18UTC)
NHM 5km: 33 hours(03,09,15,21UTC), 15 hours(00,06,12,18UTC)

System Procurement in 2010

- Next HPC System Procurement **～Jun 2010**
  - Government procurement : Comprehensive evaluation
  - Demand **8.2x faster** sustained computational speed
  - Become **operational in Mar 2012**
  - Benchmark Tests (Execution time)
    - Global: TL959, EPS TL479 & TL319, 4DVAR
    - Meso: JMA-NHM 5 km & 2 km, 4DVAR, 3DVAR, ASUCA
  - **HITACHI** won the procurement again!!
    - HITACHI has been supplying JMA HPC systems **for 50 years**.

Benchmark Tests

- Benchmark rules
  - Execute time by wall clock (UNIX “date” command)
  - Accurate calculation result
  - Permit code modifications such as
    - Loop unrolling, splitting, fusion…
    - Changing the processing order
    - Array splitting, fusion…
  - Permit inserting compiler directives
  - Permit inserting OpenMP directives

Next JMA Supercomputer

- HITACHI next SR series
  - Peak Performance : **829.4TFLOPS**
    - 2 Subsystems : 2x 414.7 TFLOPS
      - Operation + Backup(Model Development)
  - Total Memory : 108 TB
  - High-speed Storage : 348 TB
  - Data Storage : 3.7 PB + Tape Library
  - Benchmark result of **TL959L100**
    - 9 days run with 40 nodes (1280 cores)

= 35 minutes → 6～8 % of peak performance

Logical Computational Node

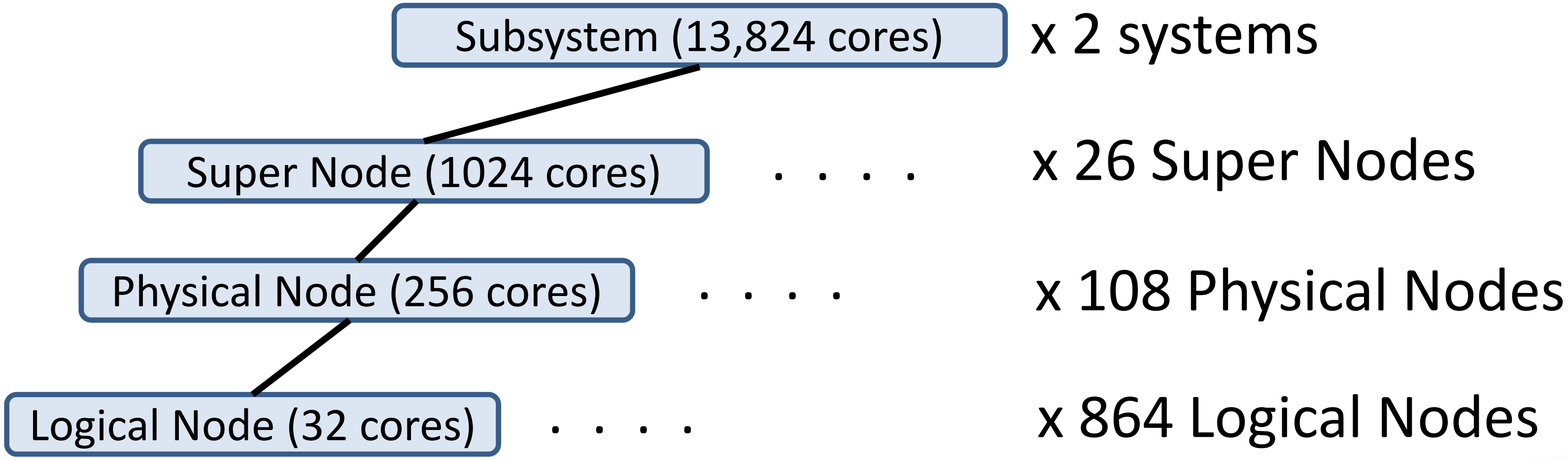
- Multi-Chip Module (MCM) = 1 Logical Node
  - **POWER7 3.75GHz** x 4
    - 8 cores x 4 sockets = 32 cores / node
  - 960 GFLOPS Peak performance / node
  - 128 GB Memory (SMP)
  - Water cooling system

Current system : 134.4 GFLOPS / node

→ Next system : 960 GFLOPS / node (**7.1x / node**)

Node Hierarchies

- **1 Physical Node** contains 8 Logical Nodes(**7.7TF**)
- **1 Super Node** contains 4 Physical Nodes(**30.7TF**)
  - 2 Super Nodes / Rack (Total : 14 computer racks)
- **One subsystem** contains 13 Super Nodes



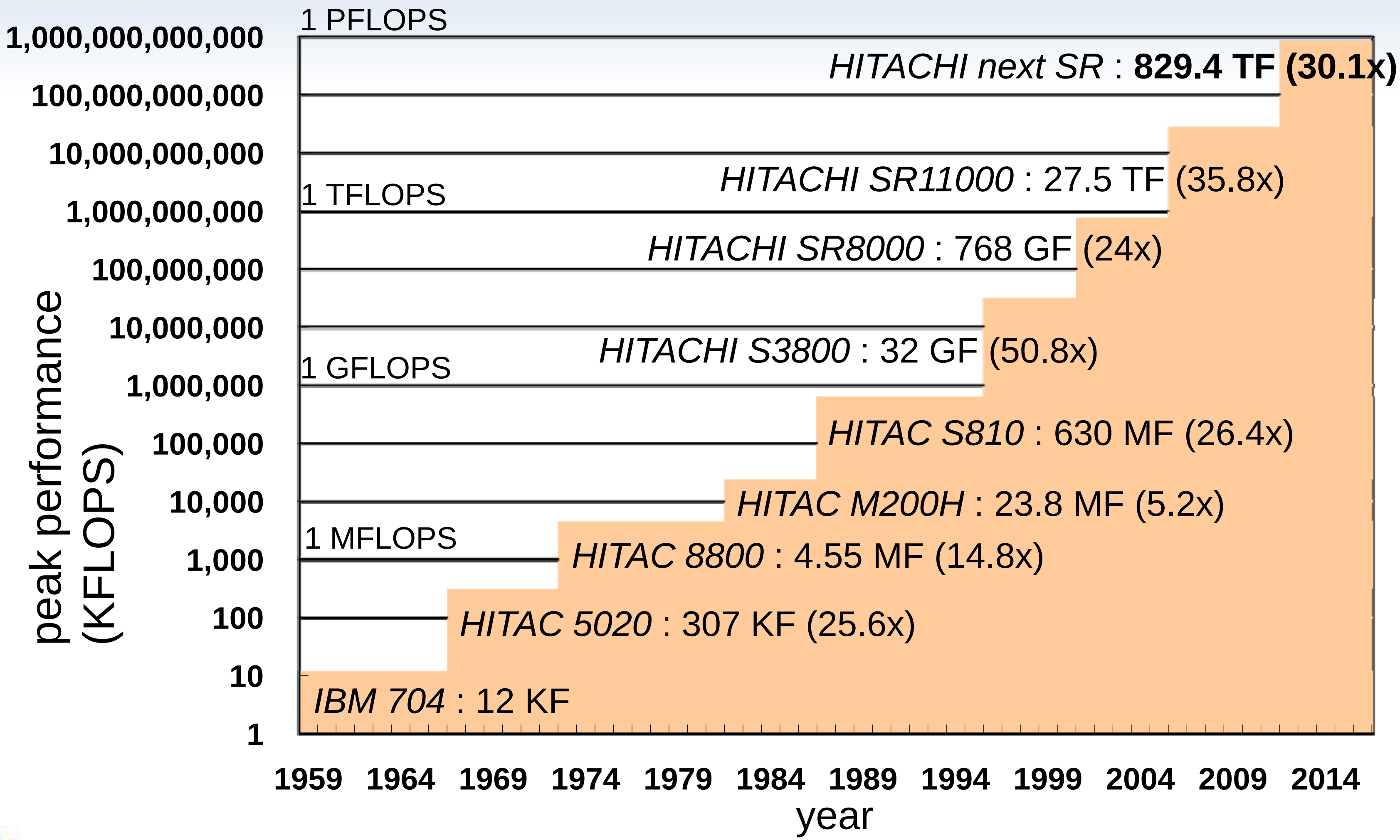
Subsystem (13,824 cores) x 2 systems

Super Node (1024 cores) . . . . x 26 Super Nodes

Physical Node (256 cores) . . . . x 108 Physical Nodes

Logical Node (32 cores) . . . . x 864 Logical Nodes

History of HPC Performance



HITAC : **HI**tachi **T**ransistor **A**utomatic **C**omputer

History of HPC systems

IBM

HITACHI

		Num of Nodes	Num of Cores
IBM 704	1959-1967	1	1
HITAC 5020	1967-1973	1	1
HITAC 8800	1973-1982	1	1
HITAC M-200H	1982-1987	1	1
HITAC S-810	1987-1996	1	1
HITACHI S-3800	1996-2001	4	4
HITACHI SR8000	2001-2006	80	640
HITACHI SR11000	2006-2012	210	3,360
HITACHI next SR	2012-	864	27,648

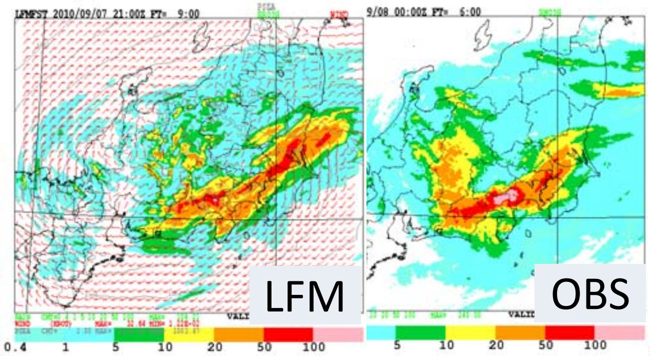
Global NWP Models

	Current (～2012) Supercomputer	Next (2012～) Supercomputer
High Resolution Deterministic	TL959L60, 4Dvar (20 km, 9 days forecast)	TL959**L100**, 4Dvar (20 km, 9 days forecast)
EPS for 1 week	TL319L60M51, SV (60 km, 51 members)	**TL479L100**M51, SV (40 km, 51 members)
EPS for Typhoon	TL319L60M11, SV (60 km, 11 members)	**TL479L100M25**, SV (40 km, 25 members)
EPS for 1 month	TL159L60M50, BGM (120 km, 50 members)	**TL319L100**M50, BGM (60 km, 50 members)
EPS for Seasonal	TL95L40M51, SV (180 km, 51 members)	**TL159L60**M51, BGM (120 km, 51 members)

Regional NWP Models

	Current (～2012) Supercomputer	Next (2012～) Supercomputer
Meso-Scale Model (MSM)	JMA-NHM 5 km L50 (33 hours, 8 times/day)	JMA-NHM 5 km **L75** (**36 hours**, 8 times/day)
EPS for Meso-Scale		JMA-NHM **10 km L60 M41**? **SV or LEKF**? (**39 hours, 4 times/day**)
Local Forecast Model (LFM)	(Test run)	JMA-NHM **2 km L60** (**9 hours, 24 times/day**)

New LFM products will support aviation weather forecasts and severe weather warnings.



Future Modeling Plans

	Current & Next (～1 PFLOPS)	**Future 2017～ (～100 PFLOPS)**
Global	Spectral, Hydrostatic TL959 (20 km)	***New Dynamical Core ?*** **→** **Yin-Yang** or **Geodesic?** **Non-hydrostatic, 10～15 km?**
Meso-scale	JMA-NHM 2～5 km	***ASUCA (NHM for MPP machine)*** (～1 km?)
EPS	TL319～479 (40～60 km), 51 members	***TL959 (20 km)?*** **> 100 members?**

Development of "ASUCA"

- The Japan Meteorological Agency (JMA) is operating a non-hydrostatic regional model (JMA-NHM) .
- JMA-NHM has been used since 1990's.
  - It is well tested and checked but ...
    - The dynamical core of JMANHM is almost retained while a lot of physics processes are developed.
    - It is extended for many years … model codes are not simple.
- The recent rapid increase in market share of scalar multi-core architecture machines is noticeable.

… these have motivated us to renovate the model

Comparison of dynamical core between ASUCA and JMANHM

	ASUCA	JMANHM
Governing equations	Flux form Fully compressible equations	Quasi flux form Fully compressible equations
Prognostic variables	ρu, ρv, ρw, $\rho\theta$, ρ	ρu, ρv, ρw, θ, p
Spatial discretization	Finite volume method	Finite difference Method
Time integration	Runge-Kutta 3rd (long) Runge-Kutta 2nd (short)	Laepflog with time filter (long) Forward backward (short)
Treatment of sound	Split explicit	Split explicit
Advection	Flux limiter function by Koren	4th (hor.) and 2nd(ver.) order with advection correction
Treatment of raindrop	Time-split	Box-Lagrangian
Coordinate	Generalized coordinate	Conformal mapping (hor.) Hybrid – Z (ver.)
Grid	Arakawa-C (hor.) Lorentz (ver.)	Arakawa-C (hor.) Lorentz (ver.)

Accurate mass conservation

Higher accuracy

Computational efficiency

Computational stability

JAPANESE PETASCALE COMPUTING PROJECTS

Japanese Petascale Computing

- **K-Computer** ~10 PFLOPS in 2012
  - Initiative by MEXT (the Ministry of Education, Culture, Sports, Science and Technology)
- University of Tokyo 1 PFLOPS in 2011
- Tokyo Tech ~2.4 PFLOPS in 2010
  - **TSUBAME2.0** developed by Global Scientific Information and Computing Center, Tokyo Institute of Technology

K-computer

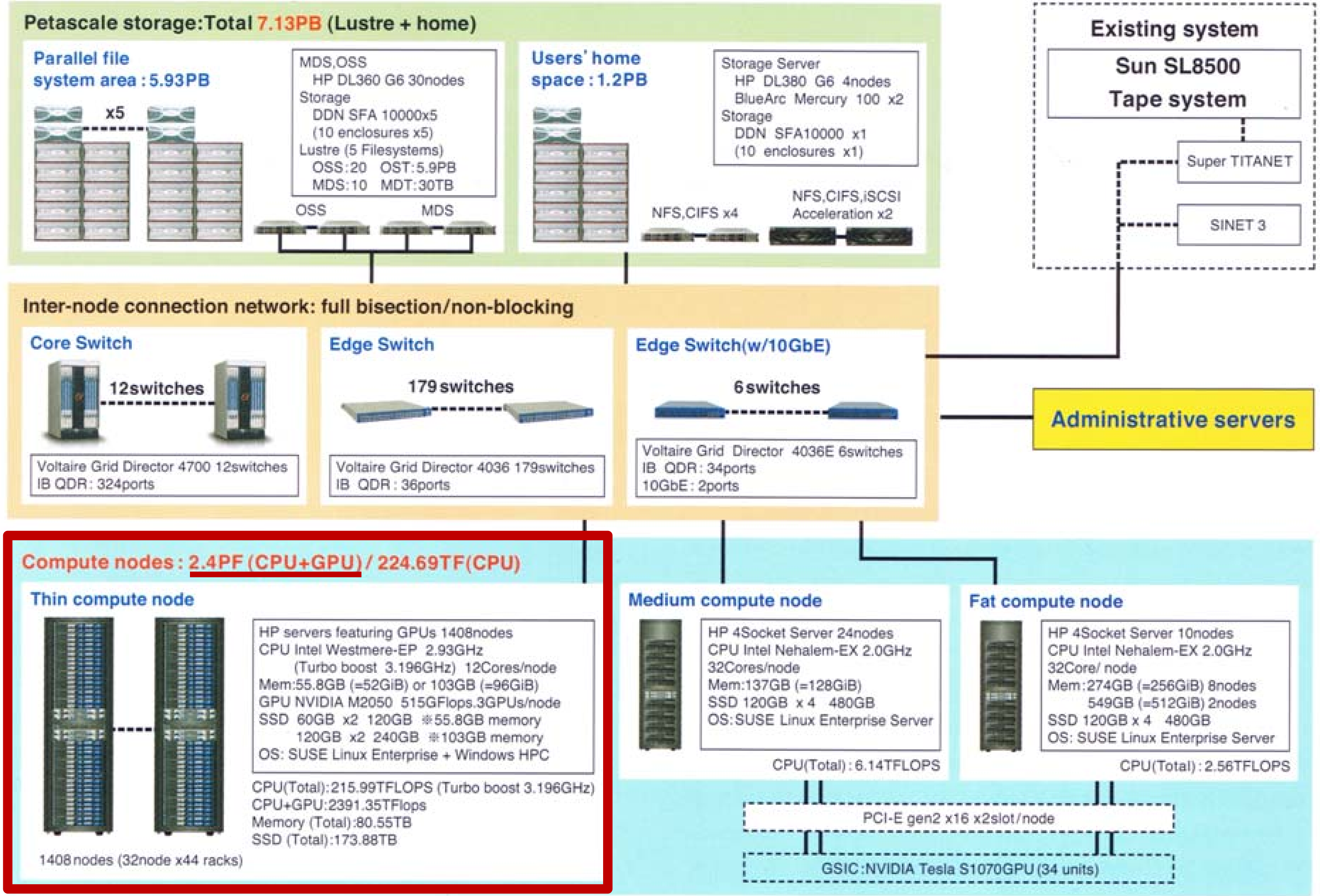
- **10 PFLOPS** Peak Performance in **2012**
  – The Japanese word “**K**eisoku” means **10 petaflops**.
- National Leadership (Initiative by MEXT)
- Next-Generation Supercomputer project
  – Carried out by **RIKEN**
  – Fujitsu **SPARC64** VIIIfx **80,000 CPUs**
- 112 billion yen ($1.3 billion)
- The site is being built in **Kobe**.

Kobe
Tokyo
450km (280miles) west from Tokyo

TSUBAME

- **T**okyo-tech **S**upercomputer and **UB**iquitously **A**ccessible **M**ass-storage **E**nvironment
  - The Japanese word "**tsubame**" means a **swallow**.
- TSUBAME1.0 (Apr 2006)
  - AMD Opteron 10,480 Cores, 50.4 TFLOPS
- TSUBAME1.2 (Oct 2008)
  - NVIDIA Tesla S1070 170 nodes, 170 TFLOPS
- **TSUBAME2.0 (Nov 2010)**
  - HP servers (CPU+GPU) 1408 nodes, **2.4 PFLOPS**

TSUBAME 2.0 System Configuration



Petascale storage:Total 7.13PB (Lustre + home)

Parallel file system area : 5.93PB

x5

MDS,OSS
HP DL360 G6 30nodes
Storage
DDN SFA 10000x5
(10 enclosures x5)
Lustre (5 Filesystems)
OSS : 20 OST : 5.9PB
MDS : 10 MDT : 30TB

OSS MDS

Users' home space : 1.2PB

Storage Server
HP DL380 G6 4nodes
BlueArc Mercury 100 x2
Storage
DDN SFA10000 x1
(10 enclosures x1)

NFS,CIFS x4

NFS,CIFS,iSCSI Acceleration x2

Existing system

Sun SL8500 Tape system

Super TITANET

SINET 3

Inter-node connection network: full bisection/non-blocking

Core Switch

12switches

Voltaire Grid Director 4700 12switches
IB QDR : 324ports

Edge Switch

179 switches

Voltaire Grid Director 4036 179switches
IB QDR : 36ports

Edge Switch(w/10GbE)

6 switches

Voltaire Grid Director 4036E 6switches
IB QDR : 34ports
10GbE : 2ports

Administrative servers

Compute nodes : 2.4PF (CPU+GPU) / 224.69TF(CPU)

Thin compute node

HP servers featuring GPUs 1408nodes
CPU Intel Westmere-EP 2.93GHz
(Turbo boost 3.196GHz) 12Cores/node
Mem:55.8GB (=52GiB) or 103GB (=96GiB)
GPU NVIDIA M2050 515GFlops.3GPUs/node
SSD 60GB x2 120GB ※55.8GB memory
120GB x2 240GB ※103GB memory
OS: SUSE Linux Enterprise + Windows HPC

CPU(Total):215.99TFLOPS (Turbo boost 3.196GHz)
CPU+GPU:2391.35TFlops
Memory (Total):80.55TB
SSD (Total):173.88TB

1408 nodes (32node x44 racks)

Medium compute node

HP 4Socket Server 24nodes
CPU Intel Nehalem-EX 2.0GHz
32Cores/node
Mem:137GB (=128GiB)
SSD 120GB x 4 480GB
OS:SUSE Linux Enterprise Server

CPU(Total) : 6.14TFLOPS

Fat compute node

HP 4Socket Server 10nodes
CPU Intel Nehalem-EX 2.0GHz
32Core/ node
Mem : 274GB (=256GiB) 8nodes
549GB (=512GiB) 2nodes
SSD 120GB x 4 480GB
OS: SUSE Linux Enterprise Server

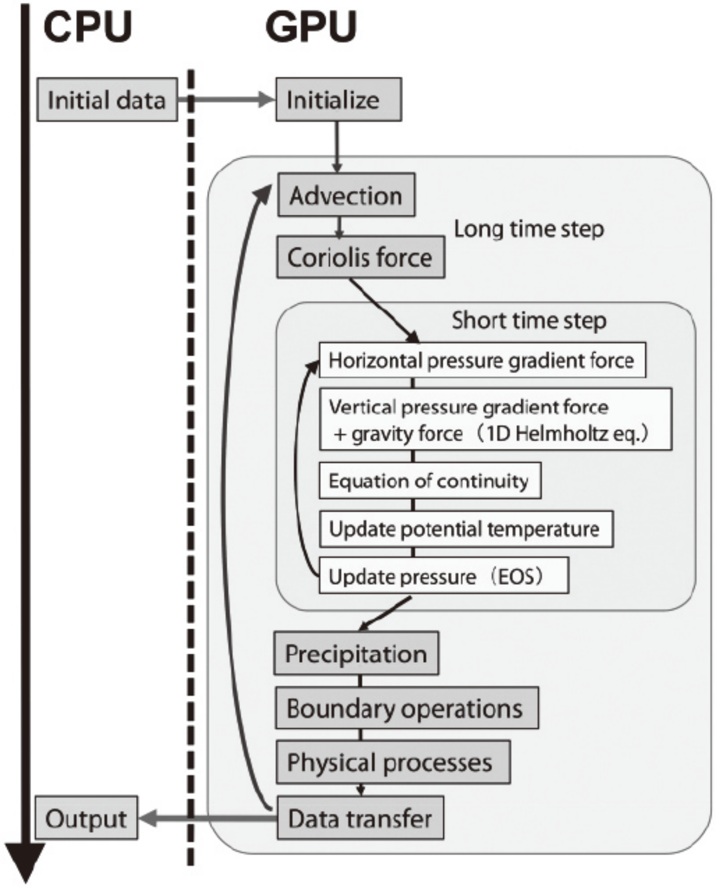
CPU(Total) : 2.56TFLOPS

PCI-E gen2 x16 x2slot/node

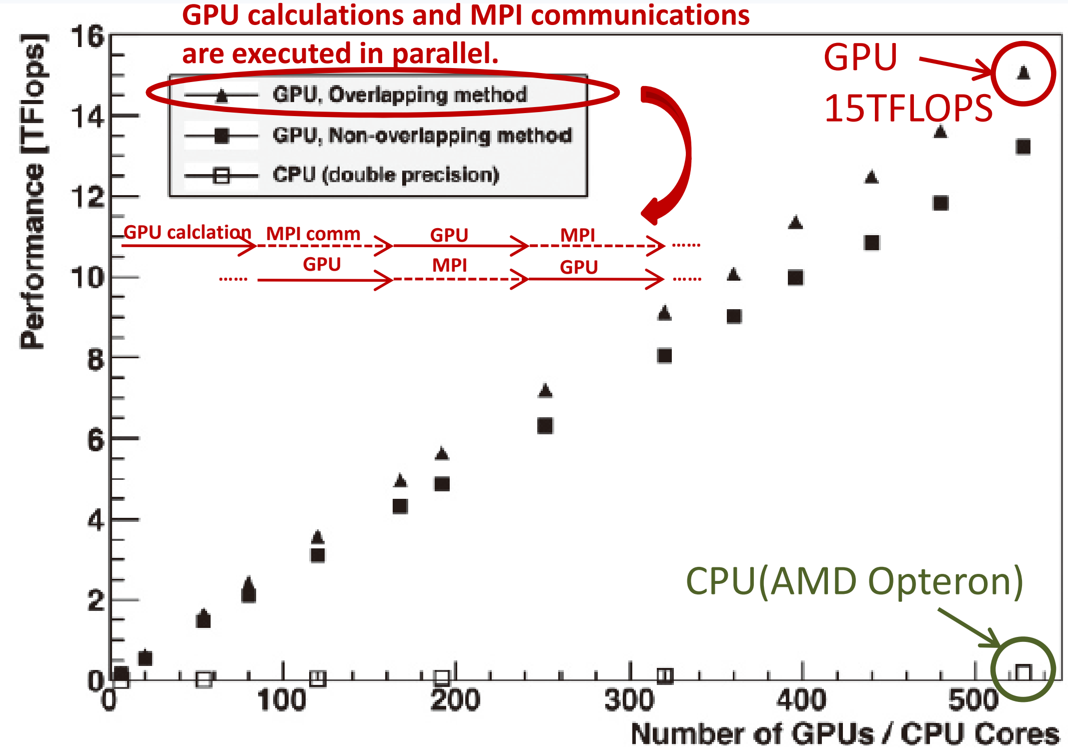
GSIC:NVIDIA Tesla S1070GPU (34 units)

NWP model on GPU supercomputer

- Joint research between **Tokyo-tech and JMA**
- **ASUCA** on **TSUBAME**
- Conversion process
  - Original : **Fortran90**
  - Rewrite to **C/C++**
  - Implement with **CUDA**
  - All time integration (dynamics & physics) is calculated on GPUs.



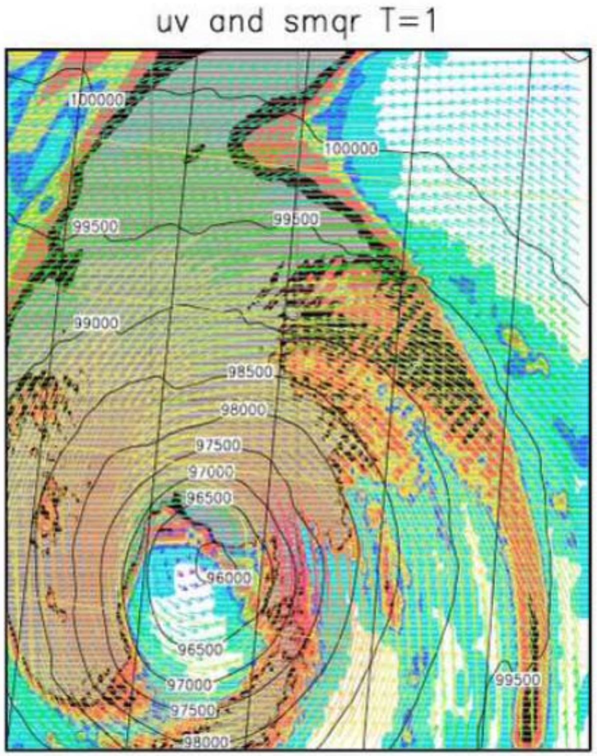
ASUCA on TSUBAME1.2



GPU calculations and MPI communications are executed in parallel.

GPU 15TFLOPS

CPU(AMD Opteron)

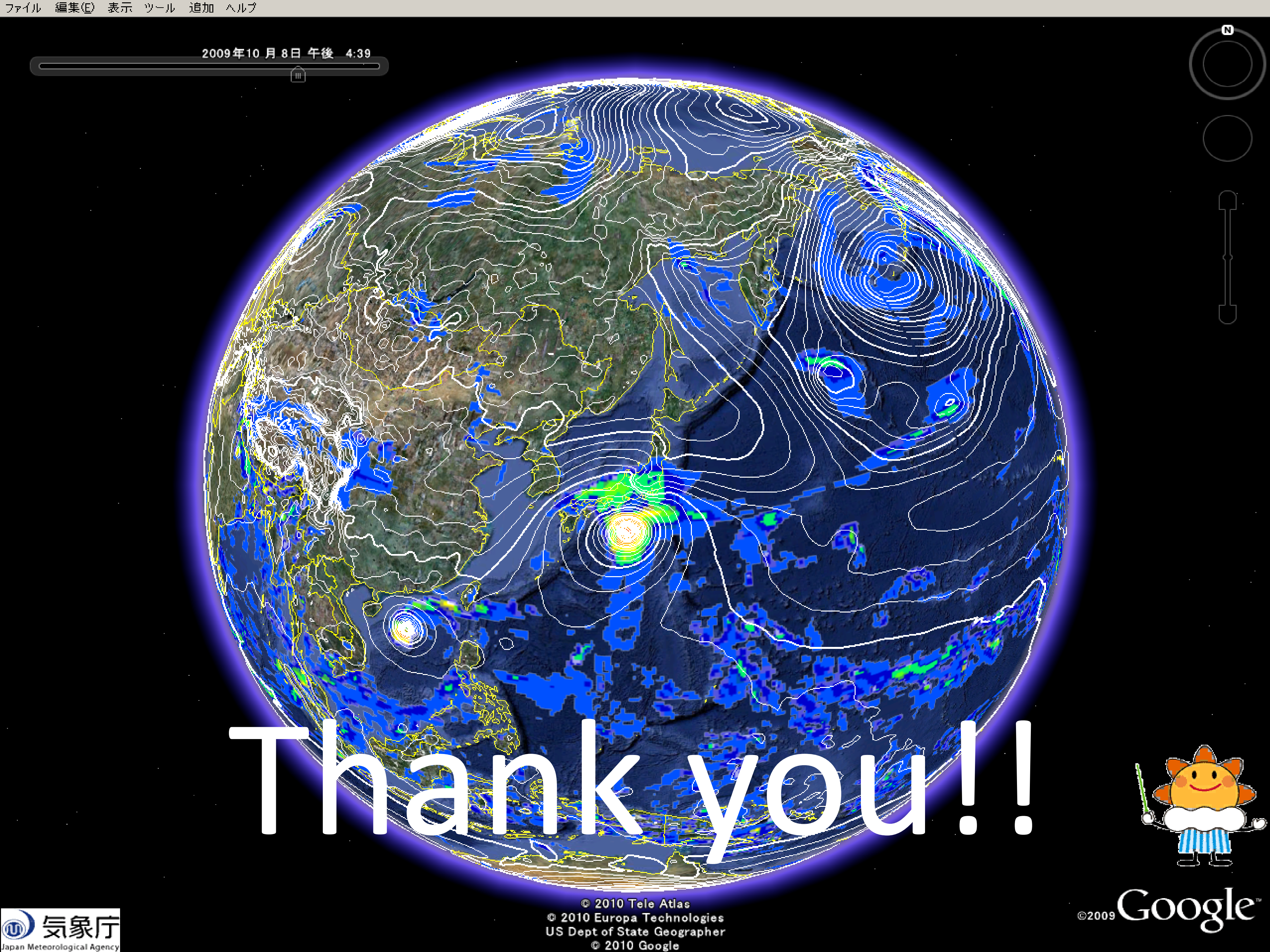


15 TFLOPS on 520 GPUs →

<u>150 TFLOPS</u> *<u>on TSUBAME2.0 (4,000 GPUs) ?</u>*

2 km mesh 3164x3028x48

(6 hours forecast / 70 min)



ファイル　編集(E)　表示　ツール　追加　ヘルプ
2009年10 月 8日 午後　4:39
N
© 2010 Tele Atlas
© 2010 Europa Technologies
US Dept of State Geographer
© 2010 Google
©2009 Google

Thank you!!

気象庁
Japan Meteorological Agency