An Update of HPC at the JMA

Toshiharu Tauchi

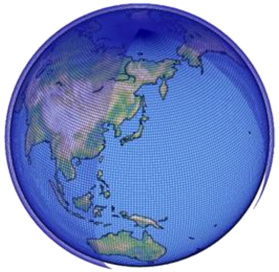
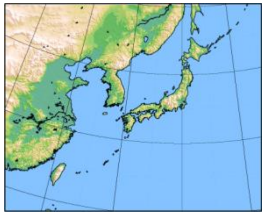
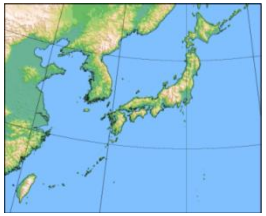
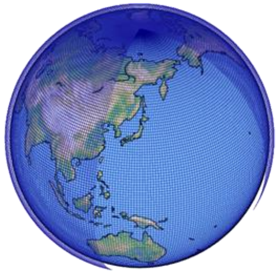
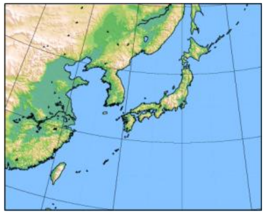
Numerical Prediction Division,

Japan Meteorological Agency

Contents

- Current NWP models of NPD/JMA
- HPC procurement
- New HPC configuration
- Migration
- Future plan

Current NWP models of NPD/JMA

	In Operation				Under Trial
	Global Spectral Model **GSM**	Meso-Scale Model **MSM**	Local Forecast Model **LFM**	Global Ensemble **GEPS**	Meso-scale Ensemble **MEPS**
objectives	Short- and Medium-range forecast	Disaster risk reduction Aviation forecast	Aviation forecast Disaster risk reduction	One-week forecast Typhoon forecast	Uncertainty and probabilistic information of MSM
Forecast domain	Global 	Japan and its surroundings (4080km x 3300km) 	Japan and its surroundings (3160km x 2600km) 	Global 	Japan and its surroundings (4080km x 3300km) 
Horizontal resolution	TL959(0.1875 deg)	5km	2km	TL479(0.375 deg)	5km
Vertical levels / Top	100 0.01 hPa	76 21.8km	58 20.2km	100 0.01 hPa	76 21.8km
Forecast Hours (Initial time)	132 hours (00, 06, 18 UTC) 264 hours (12 UTC)	39 hours (00, 03, 06, 09, 12, 15, 18, 21 UTC)	9 hours (00-23 UTC hourly)	264 h (00, 12 UTC) 132 h (06, 18 UTC)* 27 members	39h 21 members (6 hourly)
Initial Condition	Global Analysis (4D-Var)	Meso-scale Analysis (4D-Var)	Local Analysis (3D-Var)	Global Analysis with ensemble perturbations (SV, LETKF)	Meso-scale Analysis with ensemble perturbations (SV)

* when a TC of TS intensity or higher is present or expected in the RSMC Tokyo - Typhoon Center's area of responsibility (0º–60ºN, 100ºE–180º).

HPC PROCUREMENT

HPC procurement

- Requirements
  - Over 6x “effective” performance.
    - Not “theoretical (peak)” performance.
    - Evaluate by using benchmark test programs.
  - Restrictions on facilities (power).
    - Up to 4.4M Watts.
- Schedule
  - Request for information of materials … Nov. 2014
  - Request for submission of comments … Oct. 2015
  - Final RAPS release … Feb. 2016
  - Contract award … Apr. 2016

HPC procurement

- Benchmark Test
  - Programs
    - based on operational programs with next generation specs.
    - GA: Global 4D-Var(Inner:TL437)
    - GF: Global Forecast (TL1295)
    - GEPS: Global EPS (TL647)
    - MA: Meso 4D-Var(Inner:10km)
    - LA: Local 3D-Var(Inner:5km)
    - MF: Meso Forecast (asuca-5km)
    - LF :Local Forecast (asuca-2km)
  - Others … required for storage.
    - MDTEST, IOR and original I/O benchmark program.

HPC procurement

- Rules for evaluation
  - Execute the combination of two benchmark programs (e.g. Global EPS + asuca-2km).
    - Since many programs flow simultaneously in operational use.
  - Run more than each determined number of copies within a limited time by using full nodes.
    - Required to load the I/O and network as well as CPU by using whole system.
    - The time limit and the determined number were set based on the performance of SR16000 and the requested level (6x).
      - "SR16000" is the 9th generation HPC, which was in operation since June 2012.
  - Code optimizations allowed within the some predefined rules.
    - Directives (e.g. OpenACC) is OK, but code conversion (e.g. from Fortran to CUDA) is NOT.
      - Unrealistic to convert all the operational codes during the migration period.

- Pros and Cons
  - Pros: Can acquire enough "weak scaling" HPC.
  - Cons: "Strong scaling" could not be evaluated
    - We set the time limit necessary for operation.

HPC procurement

- Result
  - Winner : HITACHI Inspire the Next
    - Supercomputer : CRAY

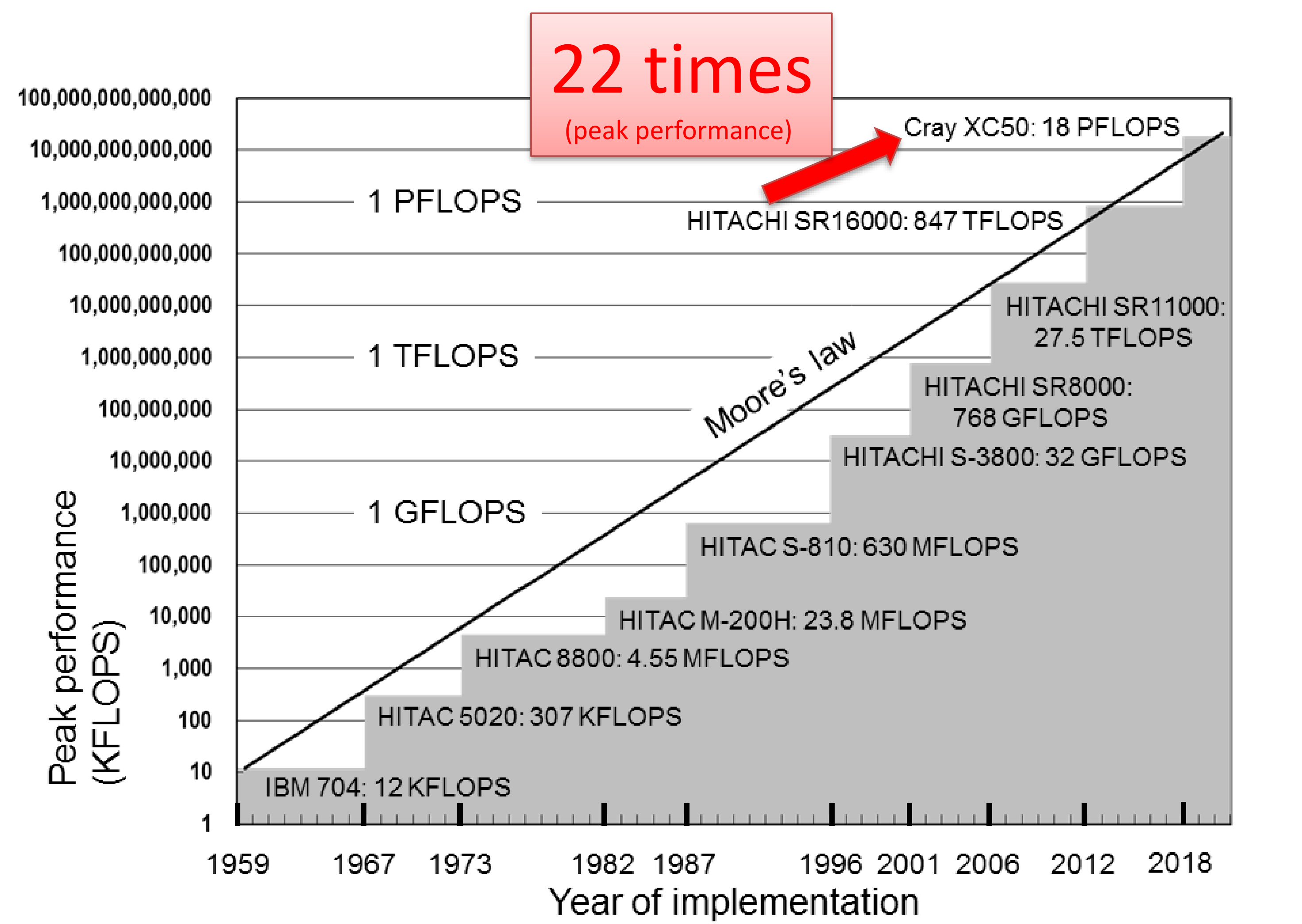
NEW HPC CONFIGURATION

Supercomputer System

- Supercomputer … Cray XC50
  - Two independent systems.
    - Main System : Operational NWP
    - Subsystem : Backup and Development
  - Specifications

Computational Node	CPU	Intel Xeon Platinum 8160 2.1GHz x2
	# of cores	24 x2
	Peak Performance	3.2256 TFlops
	Main Memory	96 GiB
Total	Num. of Nodes	2,816 (15 cabinets) x2 （ESM:2,741　MAMU:75）
	Peak Performance	9.083 PFlops x2
	Main Memory	264TiB x2
Operating system		Cray Linux Environment

HPC Growth at JMA


22 times
(peak performance)

Peak performance (Top 500)

- HPL performance of XC50
  - RMAX: 5,730.5 TFlops (62.8% of peak)
    - Ranked 25th and 26th in Top 500 (June 2018.)
  - Power: 1.354kW -> 4.232GFlops/W
    - Ranked 33rd and 34th in Green 500 (June 2018.)

Supercomputer System

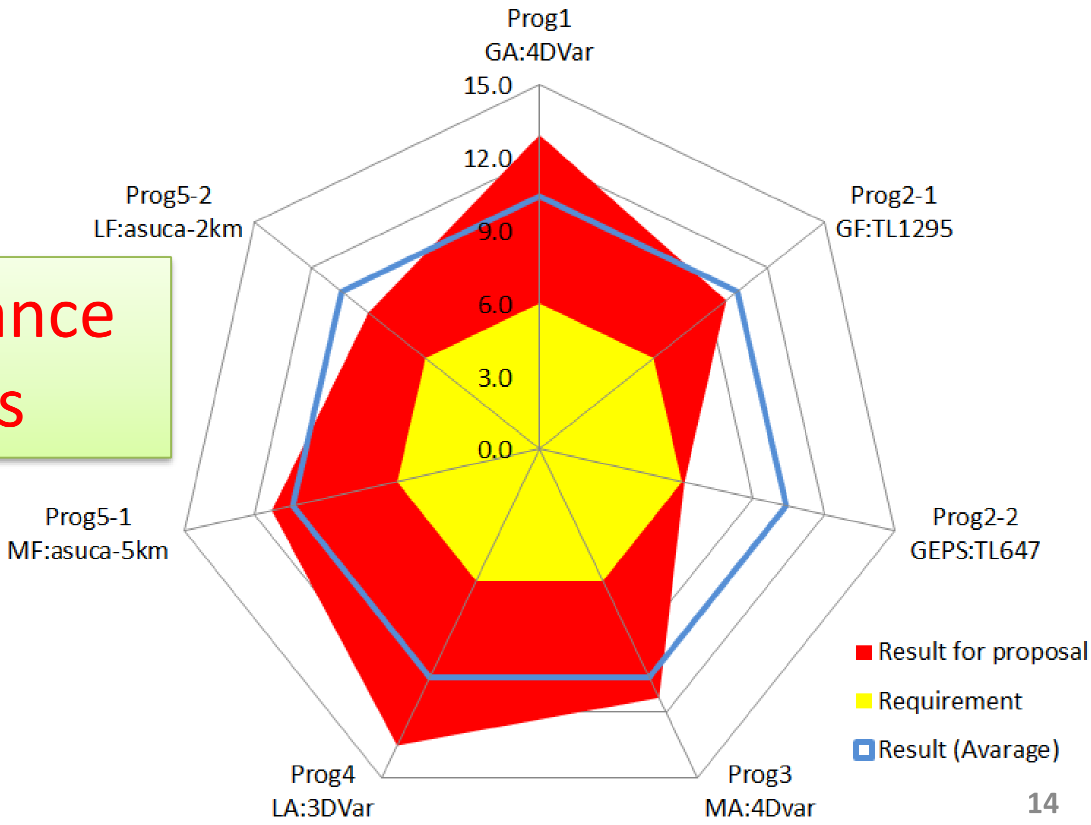
- Storage for HPC … DDN SFA14KXE
  - 3 sets for each Main system and Subsystem.
  - Filesystem: Lustre
  - Specifications (for 1 set).
    - OST … (RAID6(8D+2P) x 28 + 10S) x 2
      - 4TB 7,200rpm NL-SAS x 290 x 2
    - MDT … RAID6 (4D+2P) x 2 + 2S
      - 900GB 10Krpm SAS x 14
  - Performance.
    - Total capacity : 4.8PB/system
      - 1.6PB for each set.
    - Total I/O throughput : 135GB/s/system
      - 45GB/s read/write for each set.

Effective performance

- Result of benchmark test
  - Prog1(GA:4DVar) and Prog4(LA:3DVar) achieved more than twice the performance of our requirement.
  - Prog2-2(GEPS:TL647) achieved almost same of our requirement.

Averaged performance

: about 10 times

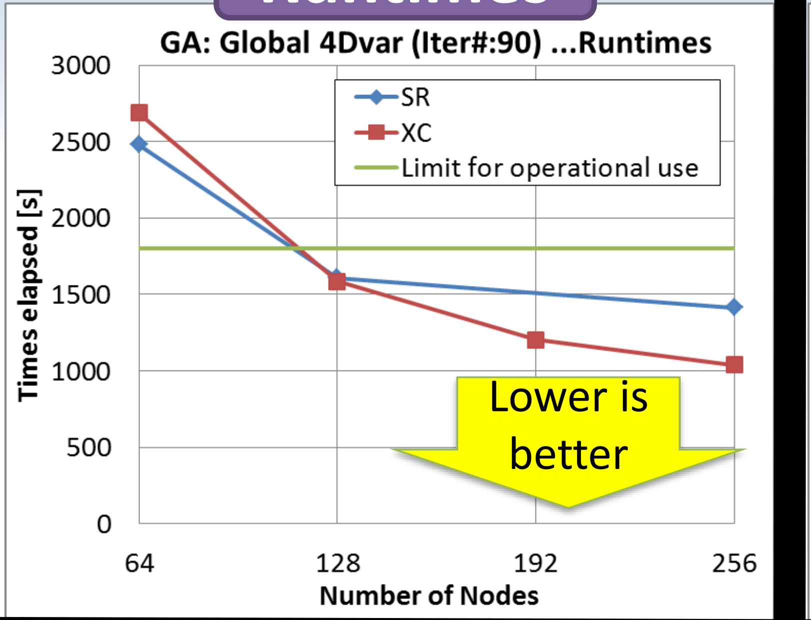


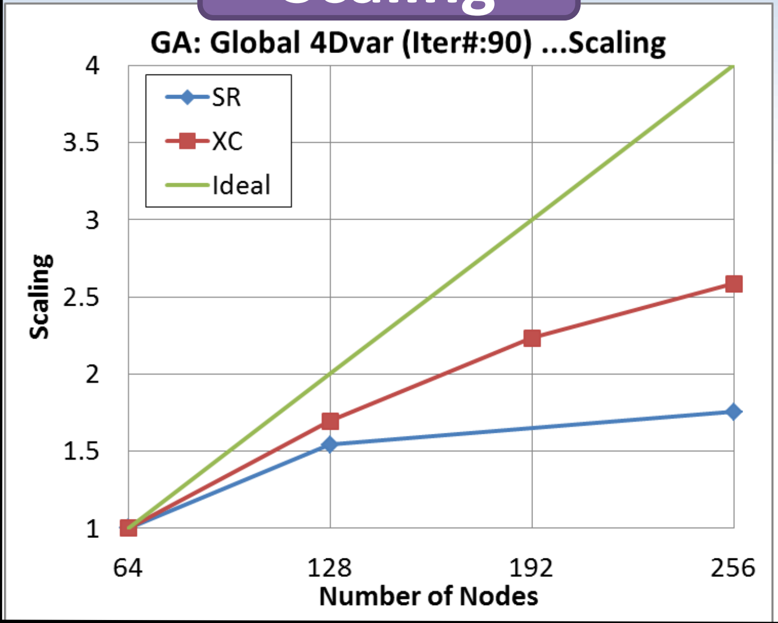
Strong scaling (Preliminary results)

Runtimes

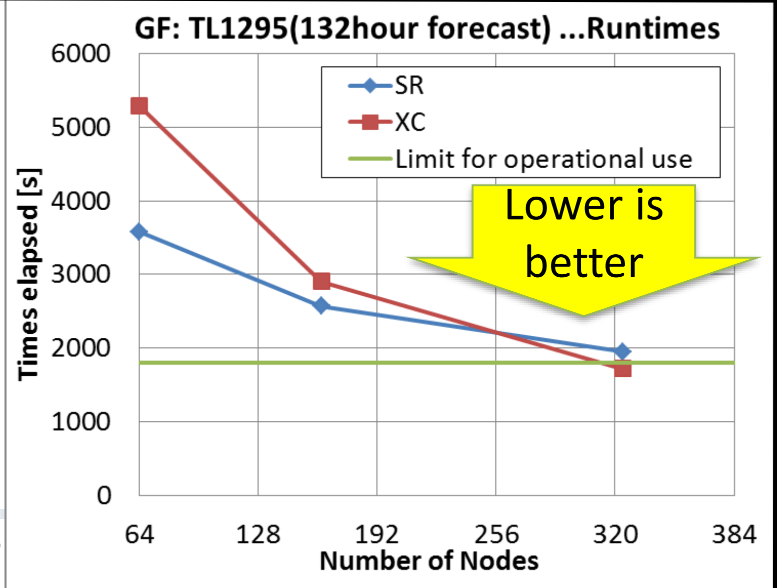
Scaling

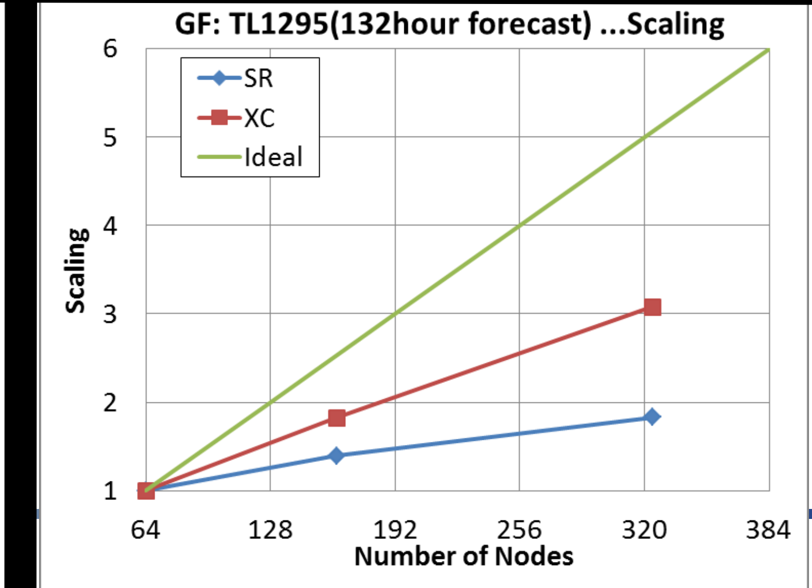
Global 4DVar


GA: Global 4Dvar (Iter#:90) ...Runtimes


GA: Global 4Dvar (Iter#:90) ...Scaling

Global Forecast


GF: TL1295(132hour forecast) ...Runtimes

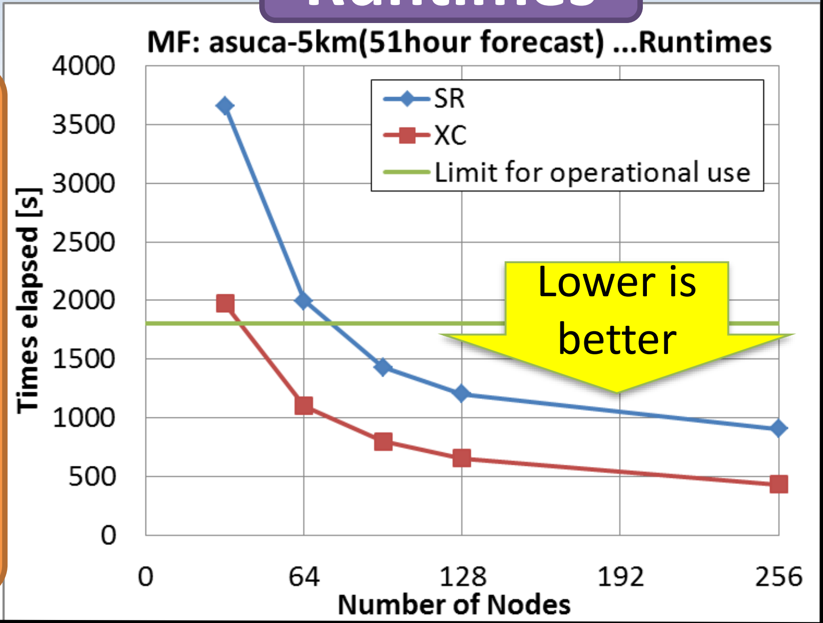

GF: TL1295(132hour forecast) ...Scaling

Strong scaling (Preliminary results)

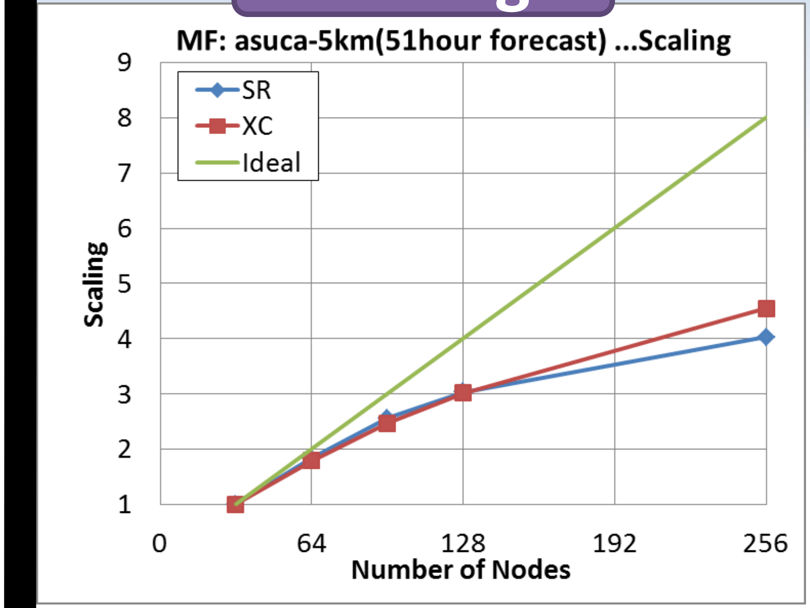
Runtimes

Scaling

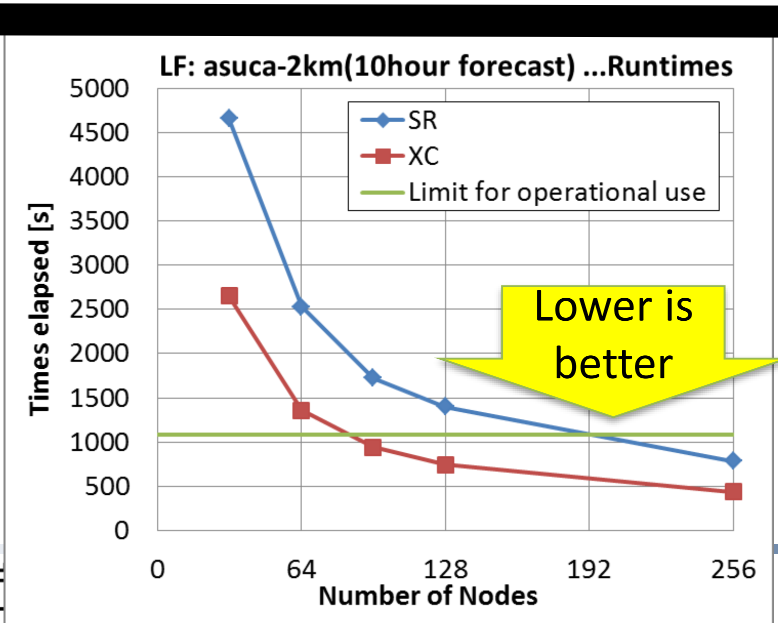
Meso-Scale forecast (asuca-5km)


MF: asuca-5km(51hour forecast) ...Runtimes

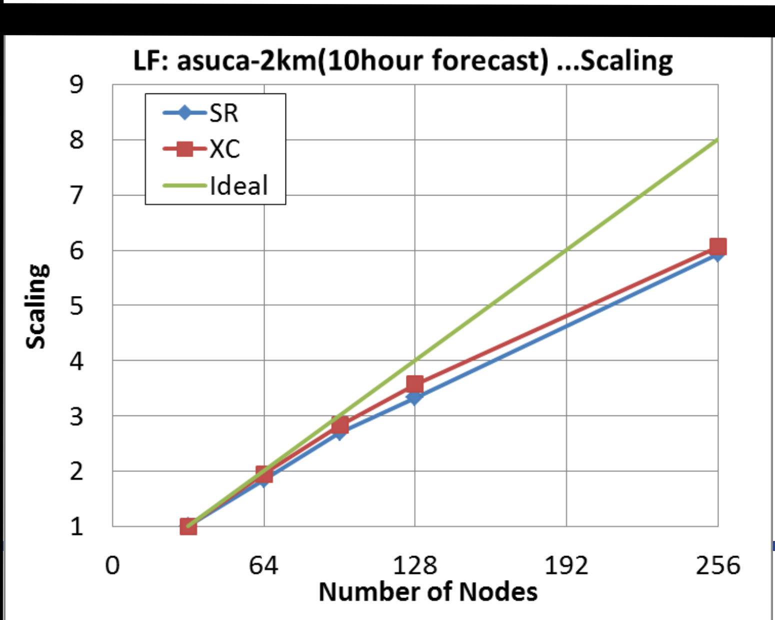
Lower is better


MF: asuca-5km(51hour forecast) ...Scaling

Local forecast (asuca-2km)


LF: asuca-2km(10hour forecast) ...Runtimes

Lower is better


LF: asuca-2km(10hour forecast) ...Scaling

MIGRATION FROM SR TO XC

Difficulties of Migration

- From Hitachi SR series to Cray XC series
  - Brand new CPU (Big change since 2001)
    - From IBM POWER to Intel Xeon.
  - Brand new compiler (Big change since the mid 1960's)
    - From Hitachi compiler to Cray (or Intel) compiler.
  - Migration from "Hitachi Service Subroutine"
    - These are provided by Hitachi along with his compiler but not supported on Cray system.

- The number of programs to be migrated are increasing.
  - Over 1,300 operational programs.
    - Total lines of programs: ~6,000,000
- The number of staffs are also increasing.
  - Gap of knowledge and ability for porting.
  - Difficulty in sharing information.

- Supposed to be the biggest migration challenge of this century.
  - We introduced a small-scale Xeon "training" server in advance and accumulated know-how.
  - Also, by using "Redmine", we shared information about bugs and tips of new system.

Schedule of Migration

- Before Jun. 2017
  - – Porting test at Xeon “training” server
- Jun. 16 2017
  - – Start using XC40 at U.S.A
- Aug. 1 2017
  - – Start using XC50 (2 cabinets) at U.S.A
- Sep. 5 2017
  - – Start using full XC50 (15 cabinets) at U.S.A
- Dec. 1 2017
  - – Start using XC50 Main system set in the JMA site.
- Jun. 5 2018
  - – Start XC50 operation.

Migration from SR to XC

- Basic Policy
  - – Prohibit the upgrade of specifications at the time of migration.
    - In case that new bugs add in the upgrade, it would be difficult to isolate reasons from system.
- Result
  - – Success in migration, and started in operation in June 5 2018.
  - – However, some system trouble occurred during the migration period, and it was investigated and resolved with the vendors.
    - CPU, Memory and I/O…

Thrash out some issues

- CPU issues
  - After starting to use full XC50, trouble “calculation result does not reproduce” occurred with 3 CPUs.
    - 3cases: case “O”, case “Ku” and case “Ka”
      - These case were named each from the Initial of first discoverers.
  - JMA staffs identified troubled CPU and Hitachi exchanged it in all cases.
    - The exchanged CPU is investigating with the vendor side.

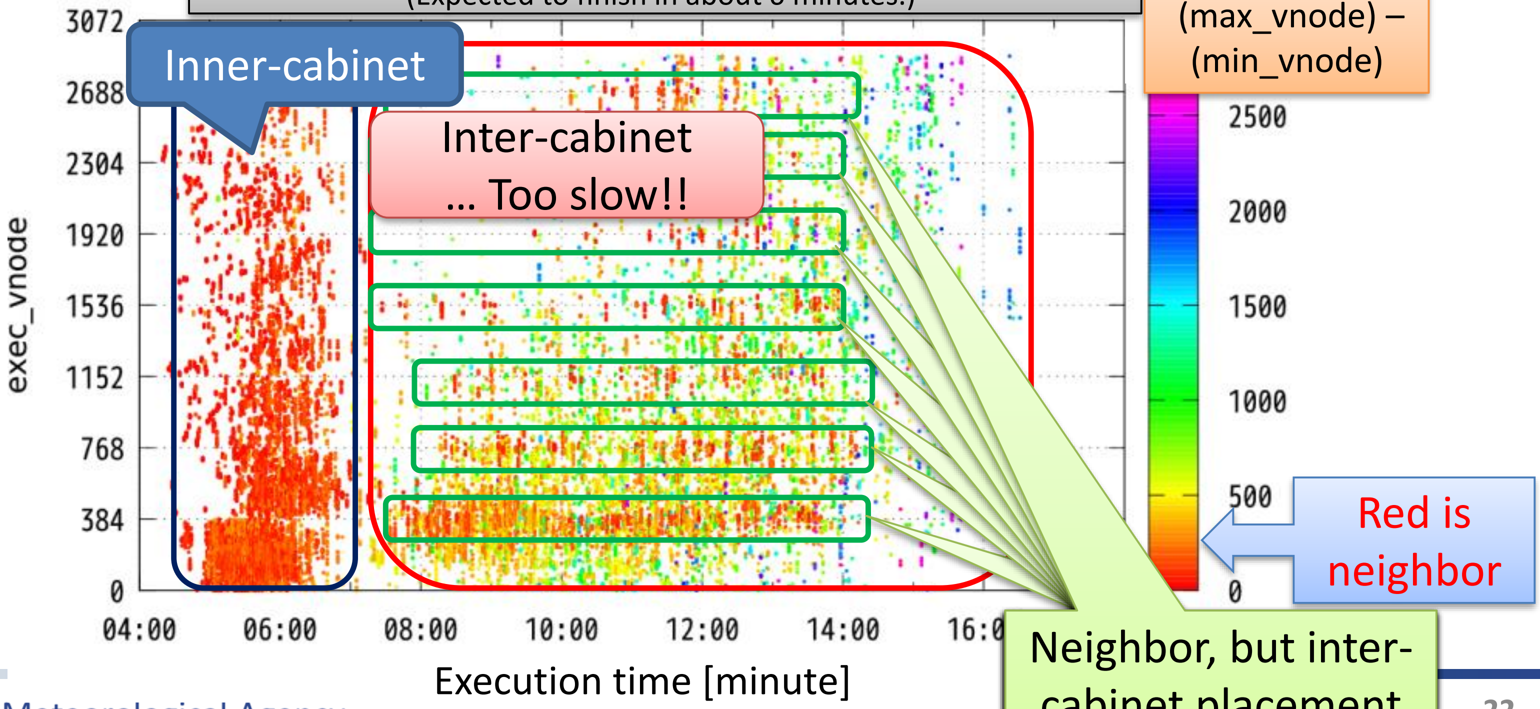
Case	Detection date	Occurrence condition (program)	Error frequency
O (xm_2446)	Nov. 2017	Various OpenMP parallel programs	< 10%
Ku (xm_914)	May. 2018	Only when using Cray compiler and Intel MKL in combination	100%
Ka (xs_1738)	Jul. 2018	Only in a specific program (Global 4D-Var) (This case was confirmed by Cray when execute HPL.)	~50%

Feel like looking for a needle in a haystack …

Thrash out some issues

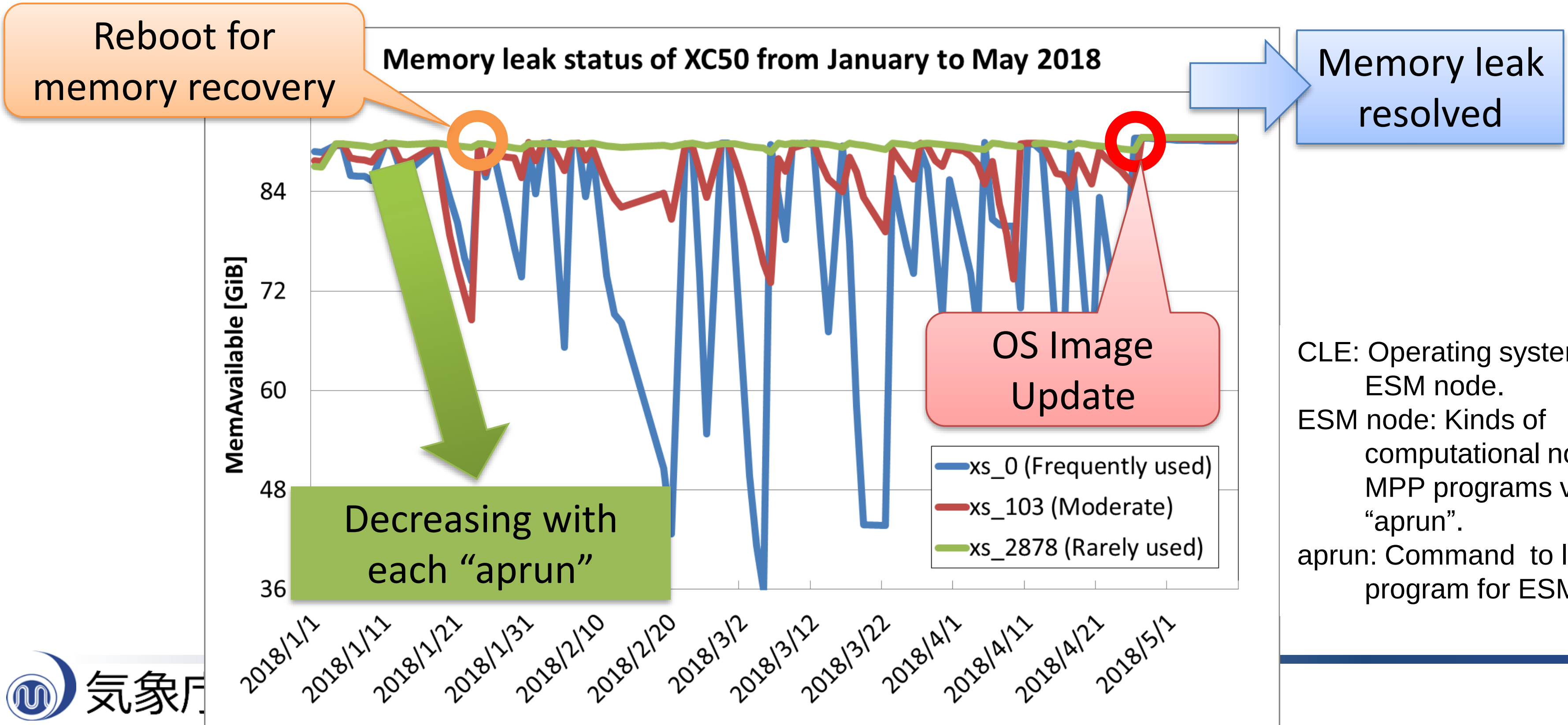
- I/O issues
  - Execution delay
    - In node arrangement using Rank3 (inter-cabinet) communication, some MPI program was significantly delayed compared to normal.
      - It seems that communication from I/O and user program(MPI) were congested.
    - By optimizing the parameters around Lustre, the issue were almost eliminated.

Execution time of Ocean Assimilation using 16 nodes.
(Expected to finish in about 6 minutes.)



Thrash out some issues

- Memory issues 1
  - Memory leak
    - Due to a bug in CLE(6.0UP04), MemAvailable of ESM node decreased with each “aprun” command.
      - Needed periodic node reboot until OS image update (Apr. 2018).



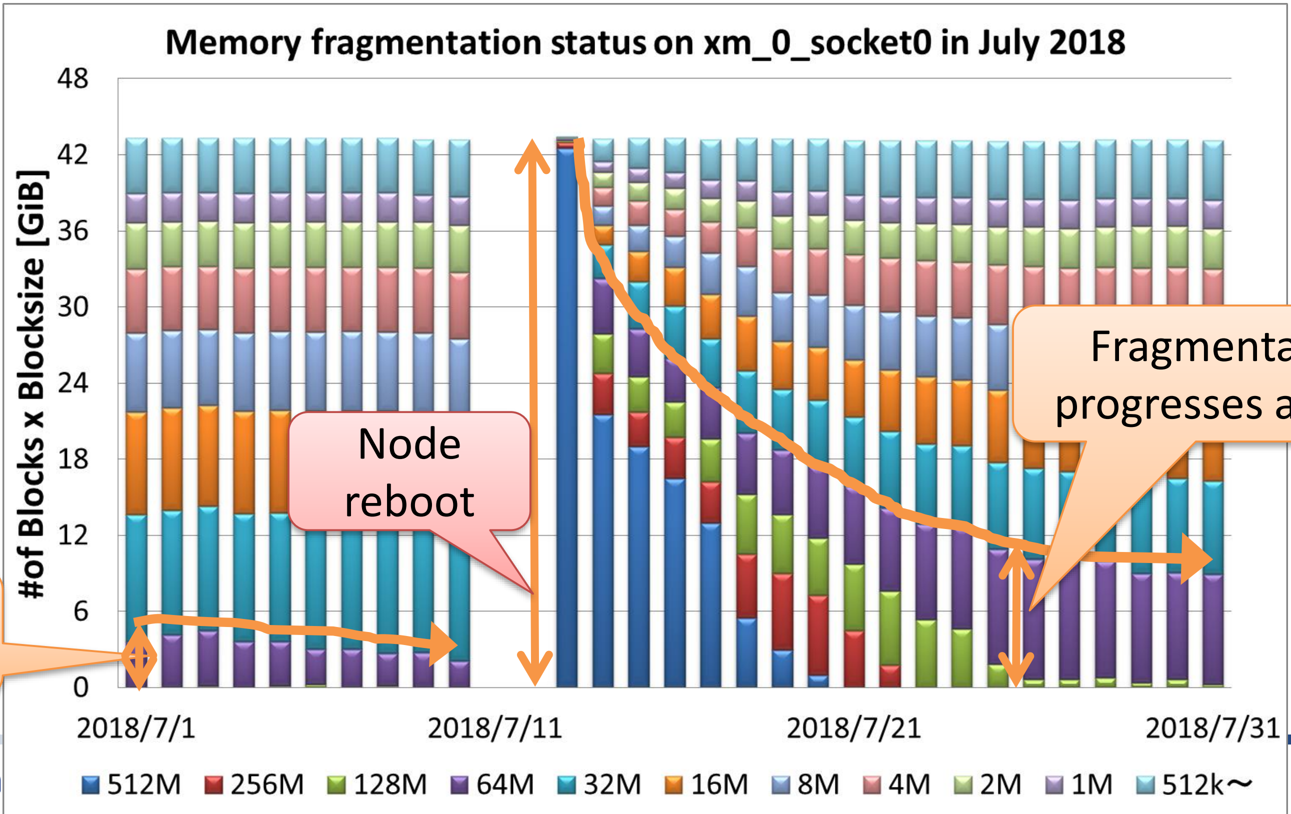
CLE: Operating system for ESM node.

ESM node: Kinds of computational node, run MPP programs via “aprun”.

aprun: Command to launch program for ESM nodes.

Thrash out some issues

- Memory issues 2
  - Memory Fragmentation
    - Execution delay occurred in programs using large hugepages.
      - Need periodic reboot even now.
        - » Question to XC users …
          - How often do you reboot your operational HPC?



Over 64MB is very small

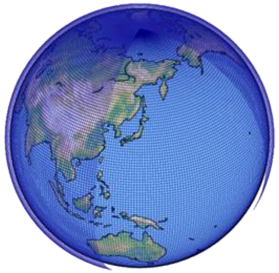
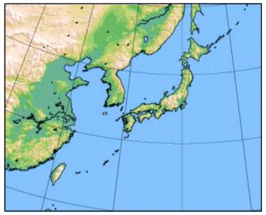
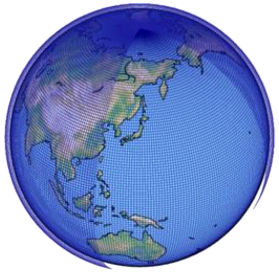
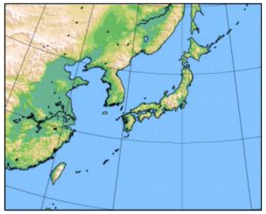
Node reboot

Fragmentation progresses again…

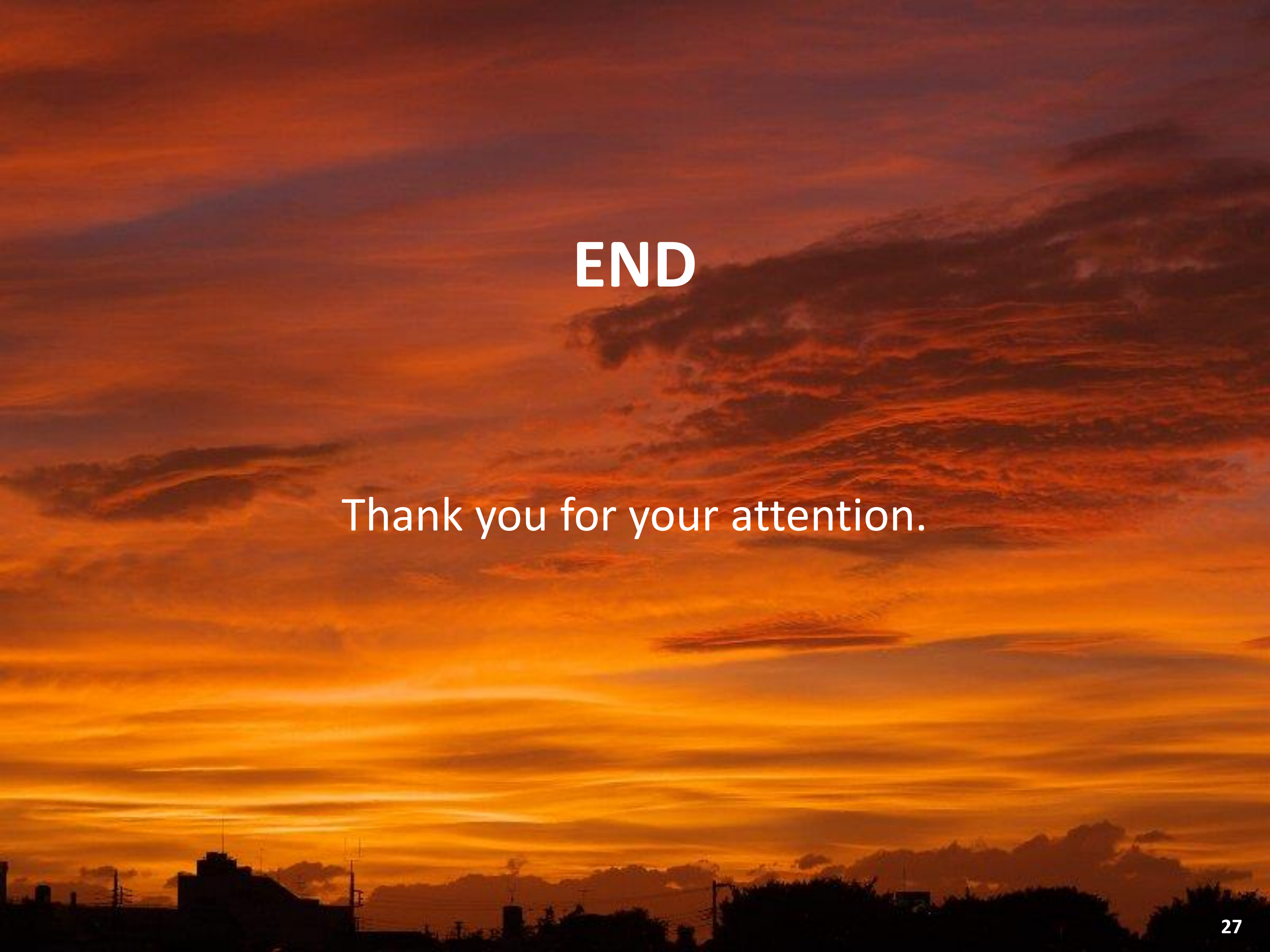
CRAY
気象庁
Japan Meteorological Agency
HITACHI
Inspire the Next

FUTURE PLAN

NWP models of NPD/JMA in future (plan)

	In Operation				
	Global Spectral Model **GSM**	Meso-Scale Model **MSM**	Local Forecast Model **LFM**	Global Ensemble **GEPS**	Meso-scale Ensemble **MEPS**
objectives	Short- and Medium-range forecast	Disaster risk reduction Aviation forecast	Aviation forecast Disaster risk reduction	One-week forecast Typhoon forecast	Uncertainty and probabilistic information of MSM
Forecast domain	Global 	Japan and its surroundings (4080km x 3300km) 	Japan and its surroundings (3160km x 2600km) 	Global 	Japan and its surroundings (4080km x 3300km) 
Horizontal resolution	TL1295(0.1389 deg)	5km	2km	TL647(0.2778 deg)	5km
Vertical levels / Top	128 0.01 hPa	96 37km	76 21.8km	128 0.01 hPa	96 37km
Forecast Hours (Initial time)	132 hours (06, 18 UTC) 264 hours (00, 12 UTC)	51 hours (00, 12UTC) 39 hours (03, 06, 09, 15, 18, 21 UTC)	10 hours (00-23 UTC hourly)	264 h (00, 12 UTC) 132 h (06, 18 UTC)* 27 members	39h 21 members (00, 06, 12, 18 UTC)
Initial Condition	Global Analysis (LETKF/4D-Var hybrid)	Meso-scale Analysis (LETKF/4D-Var hybid)	Local Analysis (3D-Var)	Global Analysis with ensemble perturbations (SV, LETKF)	Meso-scale Analysis with ensemble perturbations (SV)

* when a TC of TS intensity or higher is present or expected in the RSMC Tokyo - Typhoon Center's area of responsibility (0º–60ºN, 100ºE–180º).



END

Thank you for your attention.