An Update on High-Performance Computing at the Japan Meteorological Agency

GANZU Katsuhiko

Numerical Prediction Division (NPD) , JMA

September 18, 2025

OUTLINE

- Update of our HPC systems
- The NWP operation on current systems
- Preparing for the next-generation HPC system

UPDATE OF OUR HPC SYSTEMS

Generation of HPCs at JMA

Generation	Date of implementation	Main computer
I	1959/3	IBM 704
II	1967/4	HITAC 5020/5020F
III	1973/8	HITAC 8700/8800
IV	1982/3	HITAC M-200H (2 units)
V	1987/9	HITAC M-680
	1987/12	HITAC S-810/20K
VI	1996/3	HITAC S-3800/480
VII	2001/3	HITACHI SR8000E1
VIII	2006/3	HITACHI SR11000K1 (2 units)
IX	2012/6	HITACHI SR16000M1 (2 units)
X	2018/6	Cray XC50 (2 units)
XI	2023/3	Fujitsu PRIMEHPC FX1000 (2 units)
	2024/3	Fujitsu PRIMERGY CX2550 M7 (2 units)

- Current systems are the 11th-Gen HPC systems
  - We operate two independent HPC systems

New HPC systems

- Fujitsu PRIMERGY **CX2550** (since 2024/03)
  - 484 Intel CPU nodes x 2 systems (main system and sub-system)
  - Introduced as the successor to the previous-generation HPC system
  - Most operations are carried out on this HPC system

- Fujitsu PRIMEHPC **FX1000** (since 2023/03)
  - 4224 A64fx(ARM-based CPU) nodes x 2 systems (main system and sub-system)
    - Fx1000 employs the same A64fx CPUs as supercomputer Fugaku (Japan's Flagship HPC)
  - Additionally introduced under a supplementary budget to support the prevention of disasters caused by heavy rain
  - Mainly used for executing Local Forecast Model (LFM)

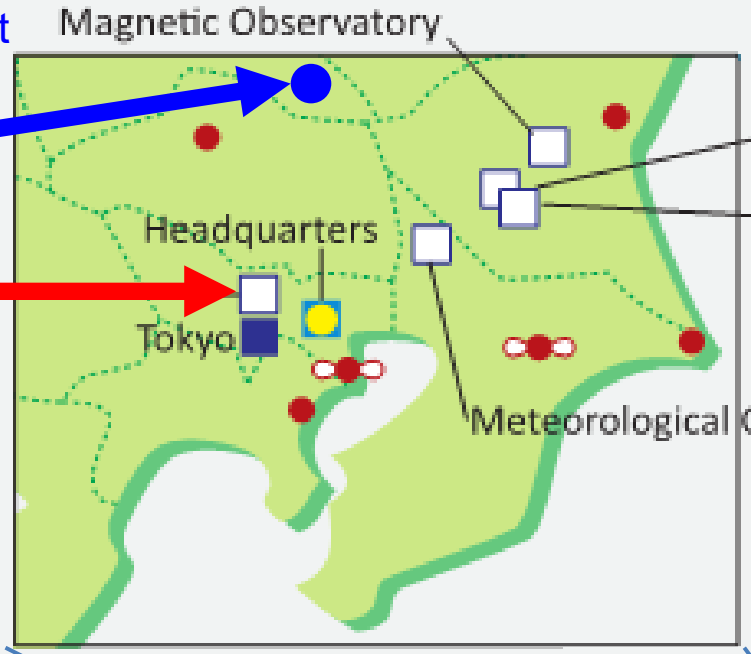
FX1000

CX2550

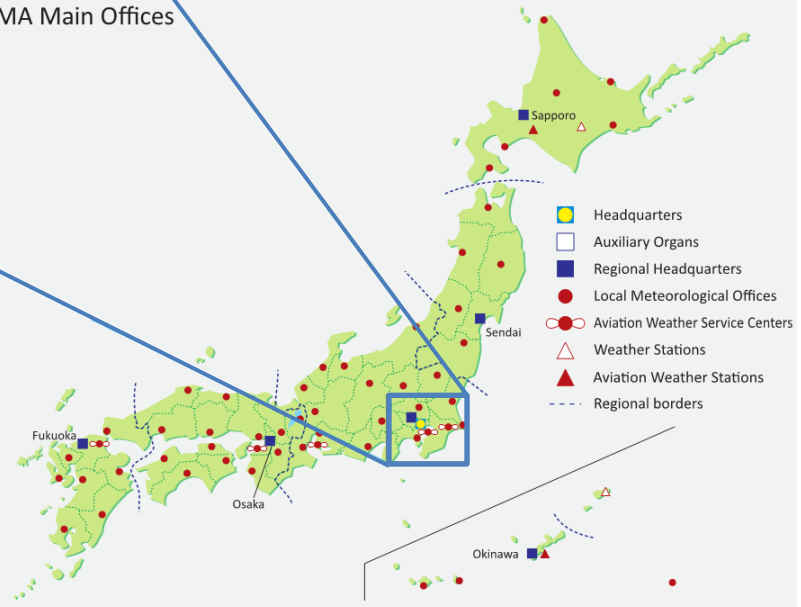
Bare-metal cloud system located at Fujitsu's Data Center

Meteorological Satellite Center

Located at Meteorological Satellite Center



JMA Main Offices



CX2550

Office of Computer Systems Operations / Meteorological Satellite Center (Kiyose)

Main system of HPC

High performance storage (9.5PB)

Sub-system of HPC

High performance storage (17.7PB)

Other systems

FX1000 system

Kiyose network devices

Decode servers

Routine control servers

Business processing servers

Common storage (9.5PB)

Large storage (19.0PB)

Front disk of choki storage (8.6PB)

Choki storage (max 160PB)

Satellite data processing part (Kiyose)

Internet

Osaka network devices

Numerical prediction BCP servers

Satellite data processing part (Osaka)

Other systems

Osaka Regional Headquarters (Osaka)

HQ network devices

Terminal computers

Other systems

JMA Headquarters (Tokyo)

Computational processing part

48 GPUs A100

High performance storage (9.5PB)

Cloud Data Center

FX1000

Cloud Data Center

Main system of HPC

High performance storage (20PB)

Sub-system of HPC

High performance storage (23.5PB)

Data Center network devices

Server computers

Server computers

Common storage (28TB)

Internet

Kiyose network devices

CX2550 system

Terminal computers

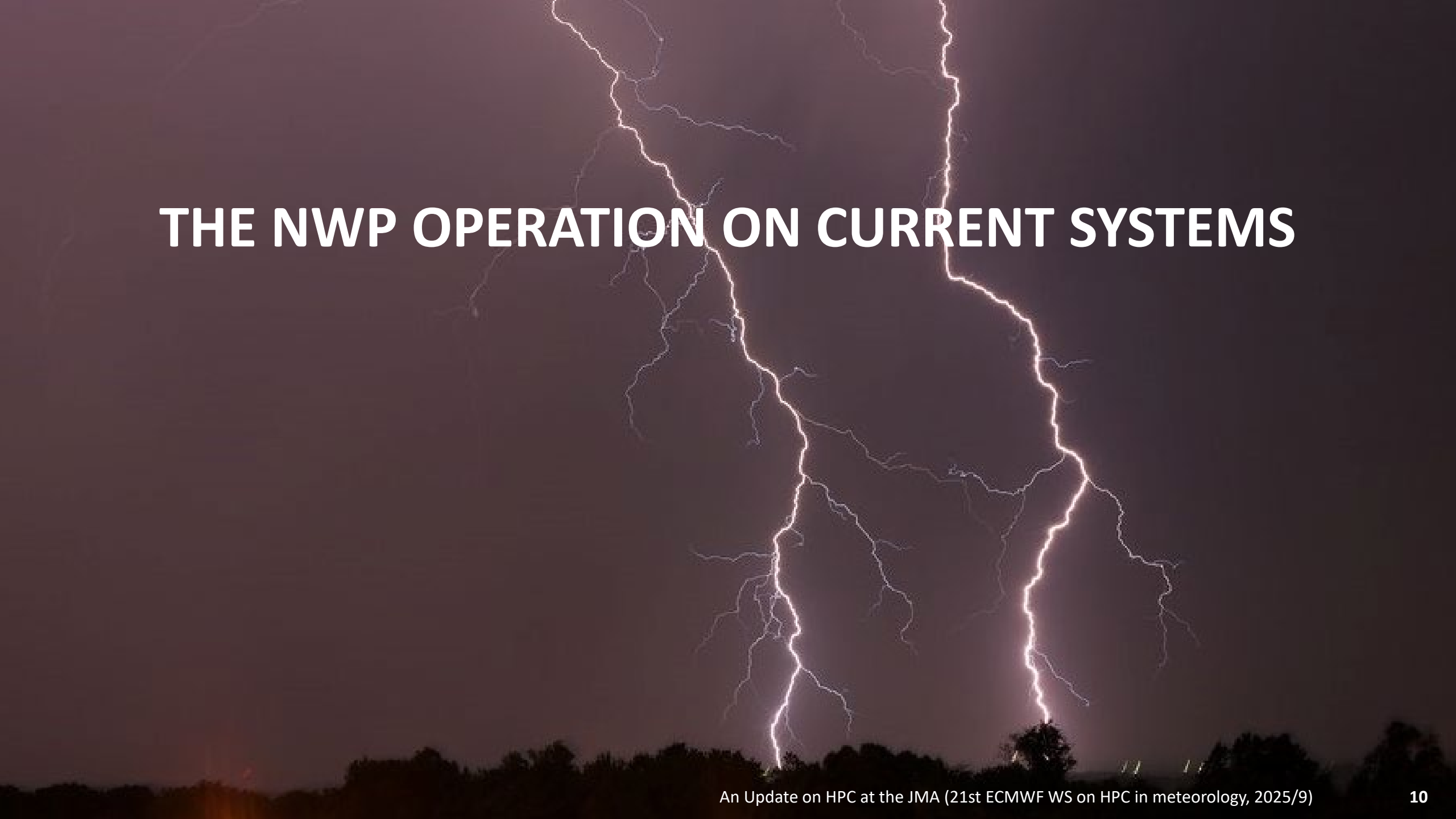
Other systems

Office of Computer Systems Operations / Meteorological Satellite Center (Kiyose)

Brief Specifications

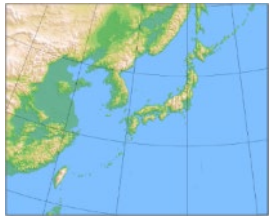
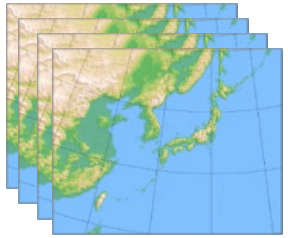
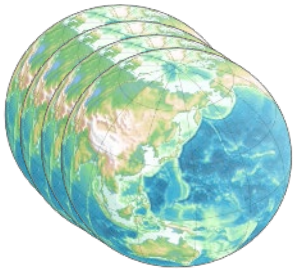
			CX2550	FX1000
Compute Node				
	Specification per system (main system or sub-system)			
		Compute Node	Fujitsu PRIMERGY CX2550 M7	Fujitsu PRIMEHPC FX1000
		Number of Nodes	**484**(active)+12(spare)	**4224**(active)+384(spare)
		FLOPS	**3.295 PFLOPS** (active)	**14.273 PFLOPS** (active)
		Memory Capacity	60.5 TiB (active)	135.2 TiB (active)
		Memory Band Width	1587 TB/s (active)	4325 TB/s (active)
		Inter-Connect	InfiniBand NDR200 (200Gbps)	InfiniBand HDR (200Gbps)
		Topology	Fat-Tree	Tofu interconnect D
	Specification per node			
		CPU	**Intel Xeon Max 9480**	**Arm A64fx**
		FLOPS (number of cores and sockets)	6.810TFLOPS (56core x 2socket)	3.379TFLOPS (48core x 1socket)
		Memory Type / Capacity	**HBM2e** / 128.0 GiB	**HBM2** / **32.0 GiB**
		Memory Band Width	**3280 GB/s**	**1024 GB/s**
Storage				
	High performance storage			
		Capacity (main system / sub-system)	DDN ES400NVX2(Lustre), 9.5PB / 17.7PB	DDN SFA7990XE(FEFS/Lustre) , 20.0PB / 23.5PB

More details: Outline of the Operational NWP at the JMA March 2025, section 1.1
https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline-latest-nwp/pdf/outline2025_01.pdf



THE NWP OPERATION ON CURRENT SYSTEMS

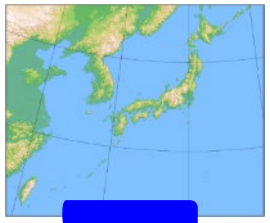
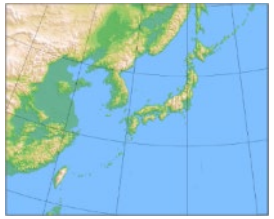
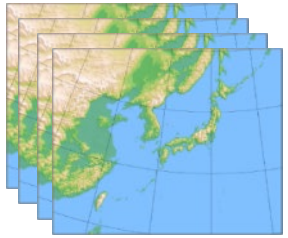
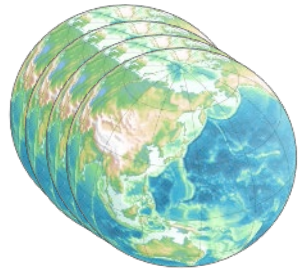
Major Operational NWP models at JMA as of March 2024

	Local Forecast Model (LFM)	Meso-scale Model (MSM)	Meso-scale EPS (MEPS)	Global Spectral Model (GSM)	Global EPS (GEPS)	Seasonal EPS (JMA/MRI-CPS3)
Domain						
Horizontal resolution	2 km	5 km	5 km	approx. 13 km	approx. 27 km (up to 18 days) approx. 40 km (up to 34 days)	Atmosphere: approx. 55 km Ocean: approx. 25 km
Forecast length (initial hours)	18 hours(00, 03, 06, 09, 12, 15, 18, 21 UTC) 10 hours(hourly except for the above)	78 hours(00,12 UTC) 39 hours(03,06,09, 15,18,21 UTC)	39 hours(00,06,12,18 UTC)	264 hours(00,12 UTC) 132 hours(06,18 UTC)	5.5 days(06,18 UTC) 11 days(00 UTC) 18 days(12 UTC) 34 days(12 UTC on Tue. and Wed.)	7 months (00 UTC)
Ensemble size	1	1	21	1	51(up to 18 days) 25(up to 34 days)	5
Main Products	Aviation Weather Forecasts and Warnings, Weather Warnings/Advisories, Precipitation Forecasts	Weather Warnings/Advisories, Precipitation Forecasts, Aviation Weather Forecasts and Warnings, Three-hourly Forecasts, Daily Forecasts	Weather Warnings/Advisories, Aviation Weather Forecasts and Warnings, Three-hourly Forecasts, Daily Forecasts	Typhoon Forecasts, Three-hourly Forecasts, Daily Forecasts, Aviation Weather Forecasts and Warnings	Typhoon Forecasts, One-week Forecasts, Early Warning Information on Extreme Weather, Two-week Temperature Forecasts, One-month Forecasts	Three-month Forecasts, Warm/Cold Season Forecasts, El Niño Outlook
Initial conditions	Local Analysis Hybrid 3D-Var	Meso-scale Analysis Atmosphere: 4D-Var Ocean: HIMSST+ Climatological Profile	Meso-scale Analysis + SV	Global Analysis Hybrid 4D-Var	Global Analysis + SV + LETKF	Atmos.: Global Analysis + BGM Ocean Analysis Ocean: 4D-Var + perturbations calculated using 4DVAR minimization history
Sea Surface Temperatures conditions	Fixed (HIMSST)	Fixed (HIMSST) + 1D ocean mixed layer model	Fixed (HIMSST) + 1D ocean mixed layer model	Anomaly-fixed SST (MGDSST)	Anomaly-fixed SST and ensemble-mean SST by CPS3 after 6 days (two-tier method)	Predicted SST in the fully coupled model (one-tier method)

Major Operational NWP models at JMA as of March 2024

FX2000

CX2550

	Local Forecast Model (LFM)	Meso-scale Model (MSM)	Meso-scale EPS (MEPS)	Global Spectral Model (GSM)	Global EPS (GEPS)	Seasonal EPS (JMA/MRI-CPS3)
Domain						
Horizontal resolution	2 km	5 km	5 km	approx. 13 km	approx. 27 km (up to 18 days) approx. 40 km (up to 34 days)	Atmosphere: approx. 55 km Ocean: approx. 25 km
Forecast length (initial hours)	18 hours (00, 03, 06, 09, 12, 15, 18, 21 UTC) 10 hours (hourly except for the above)	78 hours(00,12 UTC) 39 hours(03,06,09, 15,18,21 UTC)	39 hours(00,06,12,18 UTC)	264 hours(00,12 UTC) 132 hours(06,18 UTC)	5.5 days(06,18 UTC) 11 days(00 UTC) 18 days(12 UTC) 34 days(12 UTC on Tue. and Wed.)	7 months (00 UTC)
Ensemble size	1	1	21	1	51 (up to 18 days) (up to 34 days)	5
Main Products	Aviation Weather Forecasts and Warnings, Weather Warnings/Advisories, Precipitation Forecasts	Weather Warnings/Advisories, Precipitation Forecasts, Aviation Weather Forecasts and Warnings, Three-hourly Forecasts, Daily Forecasts	Weather Warnings/Advisories, Aviation Weather Forecasts and Warnings, Three-hourly Forecasts, Daily Forecasts	Typhoon Forecasts, Three-hourly Forecasts, Daily Forecasts, Aviation Weather Forecasts and Warnings	Typhoon Forecasts, One-week Forecasts, Early Warning Information on Extreme Weather, Two-week Temperature Forecasts, One-month Forecasts	Three-month Forecasts, Warm/Cold Season Forecasts, El Niño Outlook
Initial conditions	Local Analysis Hybrid 3D-Var	Meso-scale Analysis Atmosphere: 4D-Var Ocean: HIMSST+ Climatological Profile	Meso-scale Analysis + SV	Global Analysis Hybrid 4D-Var	Global Analysis + SV + LETKF	Atmos.: Global Analysis + BGM Ocean Analysis Ocean: 4D-Var + perturbations calculated using 4DVAR minimization history
Sea Surface Temperatures conditions	Fixed (HIMSST)	Fixed (HIMSST) + 1D ocean mixed layer model	Fixed (HIMSST) + 1D ocean mixed layer model	Anomaly-fixed SST (MGDSST)	Anomaly-fixed SST and ensemble-mean SST by CPS3 after 6 days (two-tier method)	Predicted SST in the fully coupled model (one-tier method)

Comparison of Required Node Counts: CX2550 and Previous System

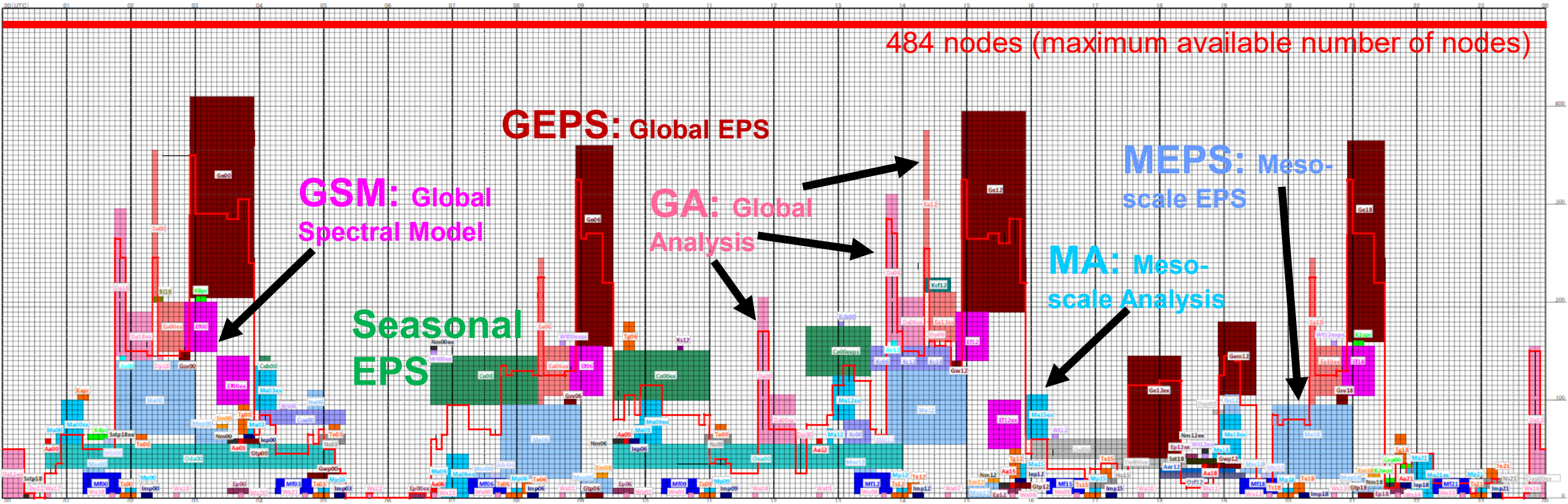
Specification per node	XC50(Previous System)	CX2550
CPU	Intel® Xeon® Platinum 8160	Intel® Xeon® CPU Max 9480
FLOPS	3.2256(2.10GHz x 24core x 2socket)	x2 → 6.810 (1.90GHz x 56core x 2socket)
Memory Type/Capacity	**DDR4 / 96.0 GiB**	**HBM2e / 128.0 GiB**
Memory Bandwidth	**255.9 GB/s**	More than x10 → **3,280.0 GB/s**

- For many of our NWP programs, the bottleneck lie in bandwidth, not FLOPS
- The programs were ported to CX2550 with minimal impact on execution time
- The high bandwidth of the new HBM-based system enables execution with significantly fewer nodes

Model	XC50	CX2550	ratio	Model	XC50	CX2550	ratio
Global Analysis	542	149	27%	Meso-scale Analysis	51	41	80%
Global Spectral Model	242	49	20%	Meso EPS	320	72	23%
Global EPS ~11days	1224	255	21%	Meso-scale Model(78H)	88	19	22%
Global EPS ~18days	624	130	21%	Meso-scale Model(39H)	51	11	22%
Global EPS ~34days	200	75	38%	Wave Ensemble System	72	18	25%
Seasonal EPS	240	50	21%	Asia-region Storm Surge EPS	104	27	26%

Required Nodes: Previous vs. Current System

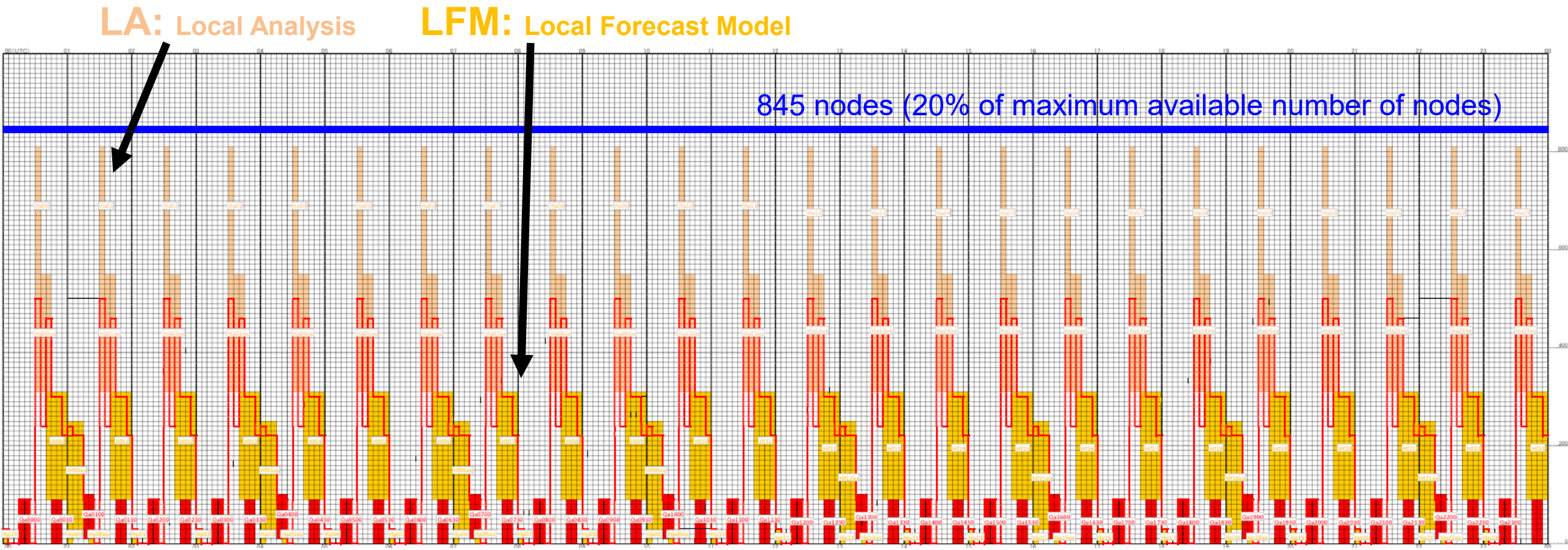
Node Utilization Diagram in CX2550 as of March 2025



- Daily schedule of job groups on the main system of the **CX2550** (March 2025)
- Height : the approximate number of nodes (5 nodes / grid)
- Width: time (5 minutes / grid)

More details: Outline of the Operational NWP at the JMA March 2025, section 1.3
https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline-latest-nwp/pdf/outline2025_01.pdf

Node Utilization Diagram in FX1000 as of April 2024



- Daily schedule of job groups on the main system of the **FX1000** (April 2024)
- Height : the approximate number of nodes (10 nodes / grid)
- Width: time (5 minutes / grid)

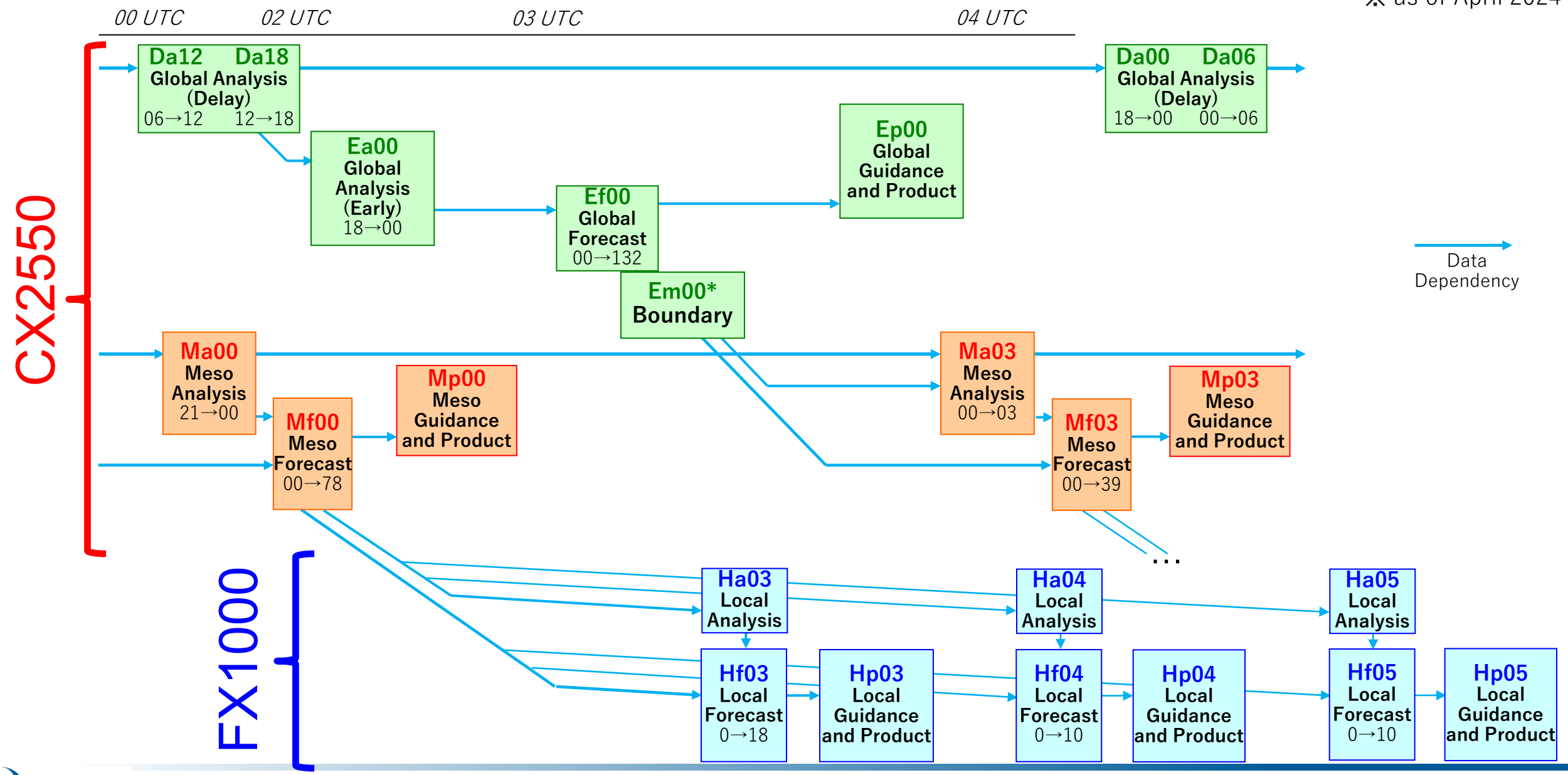
More details: Outline of the Operational NWP at the JMA March 2025, section 1.3
https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline-latest-nwp/pdf/outline2025_01.pdf

FX1000 Usage Overview

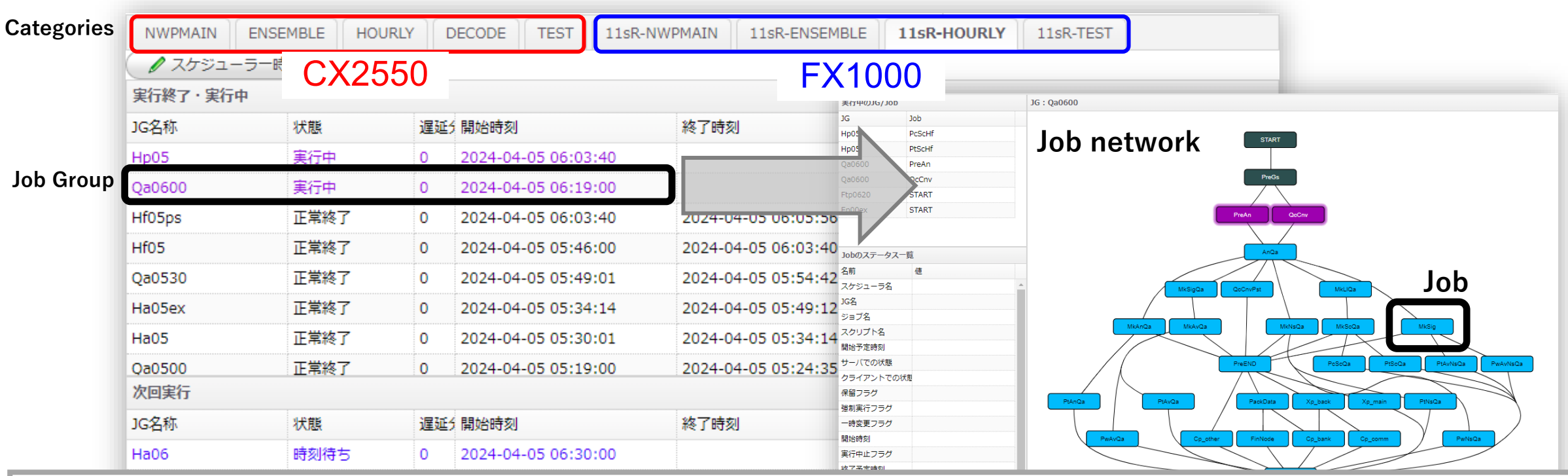
- FX1000 currently uses only a small portion of the entire system
- JMA is preparing to introduce new models:
  - Local Forecast Model (LFM)
    - Current Resolution: 2km
    - New 1km resolution model planned for 2026
  - Local EPS (LEPS)
    - New ensemble prediction system
    - 2km, 21-member EPS planned for 2026
  - These developments have also utilized the supercomputer Fugaku
    - Easy to port the development deliverables to FX1000
    - Optimization for A64fx requires effective utilization of SIMD and software pipelining
    - Need to be mindful of limited memory capacity and register count
    - Fortunately, rich set of Performance Monitoring Events (PMEs) makes profiling easy

Example of Data Dependencies between CX2550 and FX1000

※ as of April 2024



Workflow tool in JMA



- In-house developed workflow tool
- Supports various batch schedulers, including PBS, LoadLeveler, and TCS(Fujitsu's scheduler)
- Enables job control across multiple HPC systems using different batch schedulers
- JMA's operational staff also monitor job status 24/7

More details: Outline of the Operational NWP at the JMA March 2025, section 1.3
https://www.jma.go.jp/jma/jma-eng/jma-center/nwp/outline-latest-nwp/pdf/outline2025_01.pdf
Development of this tool began in Oct. 2008, and it has since been called "ROSE(Routine Operation and Scheduling Environment)"

Unexpected Porting Issues and Their Remedies

- **CX2550**
  - Execution delays may occur when a large job uses a node immediately after an I/O-intensive job
    - Due to residual page cache
    - Resolved by vendor setting that clears the cache just before job start
  - Allocation failures on shared nodes with multiple job submissions
    - Scheduler uses cgroups; page cache reserved within one cgroup may not be released
    - Not fundamentally solved; currently avoided by reserving extra memory.
- **FX1000**
  - File cache limit on compute nodes was too small, slowing I/O
    - Resolved by vendor adding an environment variable to adjust the limit
  - I/O is slow when jobs are submitted while racks are not busy
    - This occurs because I/O nodes may remain idle when racks are not busy
    - Vendor setting applied to prevent I/O nodes from entering idle state
  - With SIMD enabled, certain optimization options caused registers to be unprotected in atan(), leading to unintended variable overwrites
    - Fixed in the updated compiler suite.

PREPARING FOR THE NEXT-GENERATION HPC SYSTEM

Procurement of the 12th-gen HPC

- JMA has started exploring the procurement of the next-generation HPC system
- The next-generation system is expected:
  - With operations scheduled to begin around the end of FY2028
- Considering recent trends, the 12th-gen HPC system will likely include GPUs.
  - CPU nodes are expected to remain a significant part of the system
  - AI applications may be an important purpose of this system

FY in Japan: from April to March of the following year

Porting our NWP models to GPUs

- Fortran directive-based (OpenAcc)
  - NVIDIA HPC SDK
    - Drastic loop restructuring required for optimization
    - Optimization of data transfer between CPU and GPU
      - Connection between CPU and GPU is a key issue (e.g. NVLink-C2C)
  - Other essential techniques for adequate performance:
    - with CUDA aware MPI, GPU Direct RDMA
    - Math Libraries (cuFFT, cuBLAS)
- Challenges
  - Adaptation to AMD and Intel GPU
    - Adaptation to NVIDIA environments is ahead
    - DSLs are considered a candidate solution, though their long-term maintainability needs further evaluation

Converting NWP models to FP32

- Ongoing work to convert NWP models to FP32:
  - Improve performance
  - Reduce memory usage
  - Prepare for potential decline in FP64 capabilities on future GPUs
- Testing of Single-Precision Models:
  - Execution stability is a critical issue
    - Daily runs and hindcast experiments help validation result
    - Failures due to loss of significance heavily depend on the physical field
    - Some issues only appear over years of simulation, requiring substantial resources
  - Use of Fujitsu TCS Compiler (A64fx):
    - TCS compiler provides precision-lowering function that sets the user-specified number of lower significand bits to zero during variable assignment
    - This function has helped identify and fix potential failure points due to insufficient precision

Additional GPUs for Development Use

- Introduction of a New GPU Environment:
  - Expected to be available in FY2025
  - Separate from the next-generation HPC
  - Intended for development, not operational use
  - Aimed for developing more accurate typhoon prediction
  - Contract signed with a cloud provider Sakura internet

- Overview:
  - Preparing NVIDIA H100 * 160 as a bare-metal cloud environment
  - In addition to porting to GPUs, it will also be used for AI applications
  - Considering Slurm support for our workflow tool

Conclusion

- JMA introduced new HPC systems in March 2023 (FX1000) and March 2024 (CX2550):
  - New systems are 11th-generation
  - HBM enables to reduce node count in CX2550
  - Plan to improve the resolution of LFM(1km) and introduce new Local EPS(2km, 21members) in FX1000
  - Interdependent jobs between CX2550 and FX1000 are managed via JMA's workflow tool

- Preparing for the future HPC:
  - Plan to introduce 12th-generation HPC
  - Next system may include large GPU nodes
  - We are porting our programs to GPU and converting to FP32
  - Separate GPU-based cloud environment for development will be available in FY2025