# Design Consideration for Data Assimilation in Post K Supercomputer

Yutaka Ishikawa
RIKEN AICS

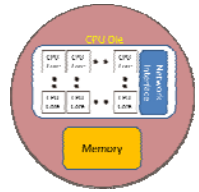
Yutaka Ishikawa @ RIKEN AICS

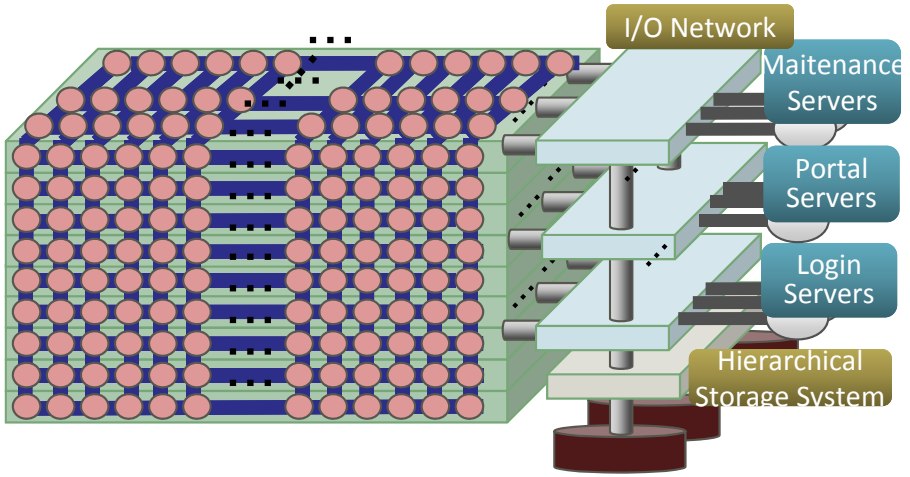
2015/02/23

14:30 – 14:50

# FLAGSHIP2020 Project

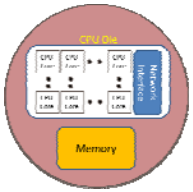
- Missions
  - Building the Japanese national flagship supercomputer, Post K, and
  - Developing wide range of HPC applications, running on Post K, in order to solve social and science issues in Japan
- Budget
  - 110 Billion JPY  (about 0.91 Billion USD in case of 120 JPY/$)
  - including research, development and acquisition, and application development

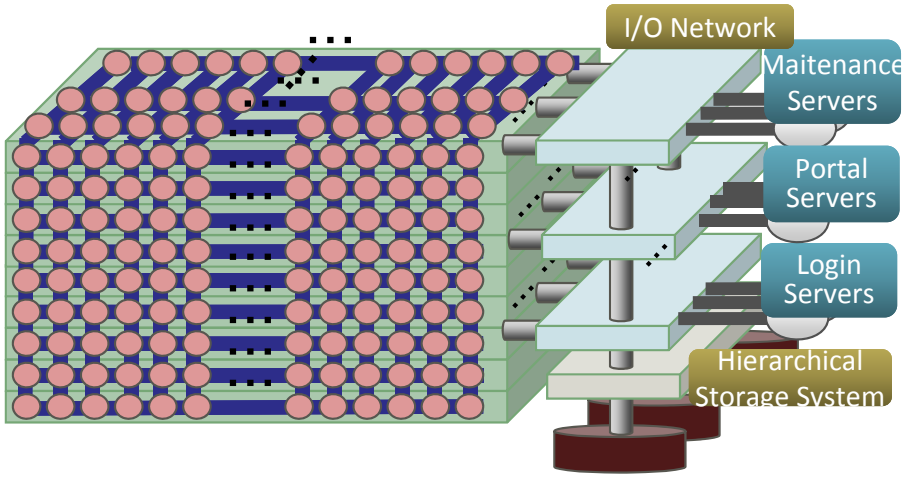


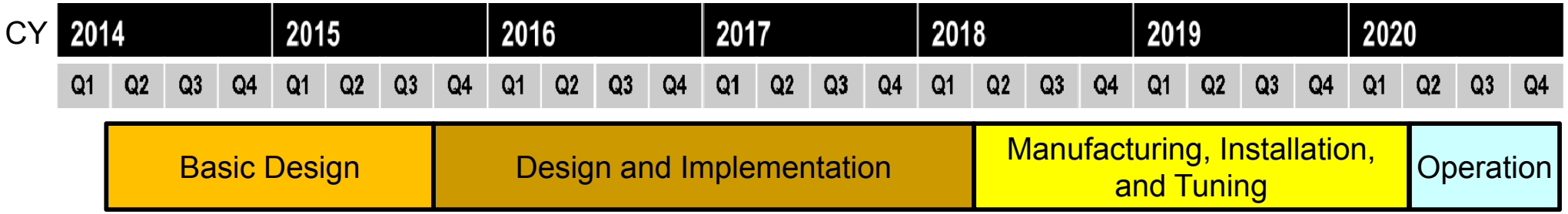



# FLAGSHIP2020 Project

- Missions
  - Building the Japanese national flagship supercomputer, Post K, and
  - Developing wide range of HPC applications, running on Post K, in order to solve social and science issues in Japan
- Budget
  - 110 Billion JPY (about 0.91 Billion USD in case of 120 JPY/$)
  - including research, development and acquisition, and application development
- Hardware and System Software
  - Post K Computer
    - RIKEN AICS is in charge of development
    - Fujitsu is vendor partnership

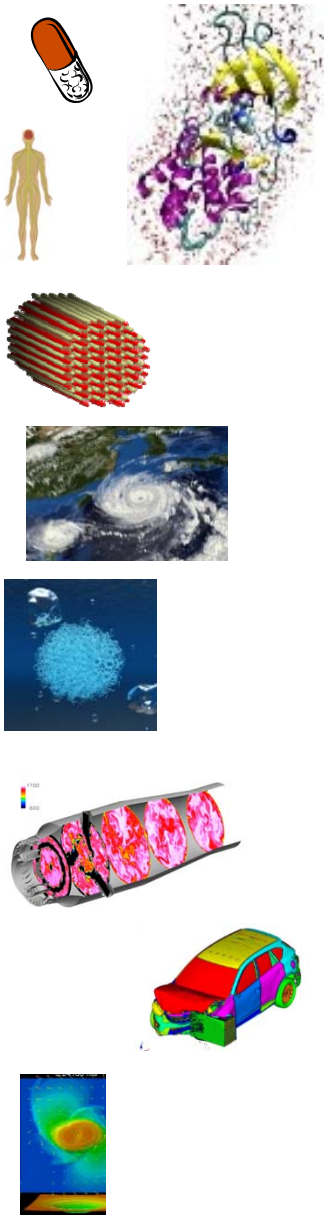


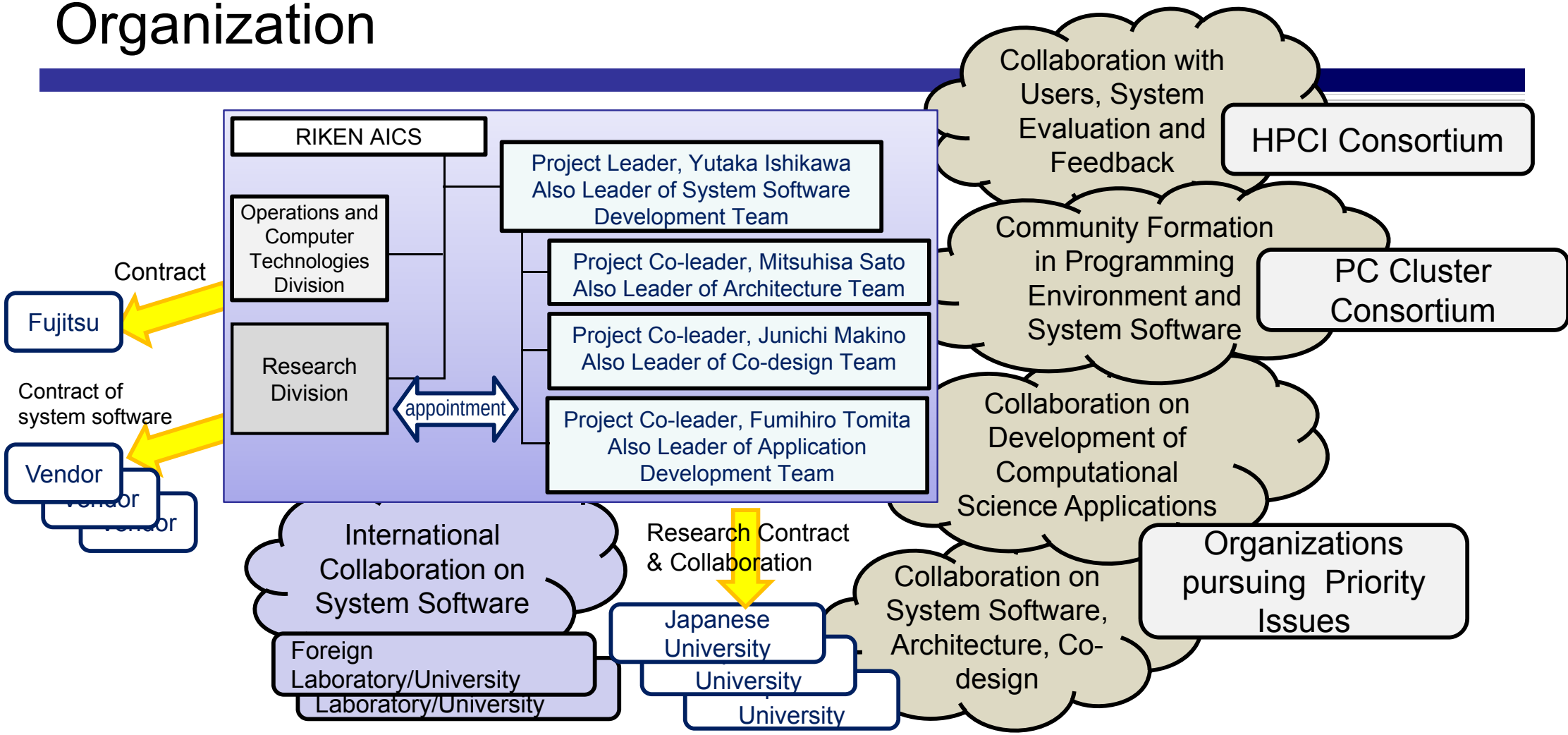



# Social and scientific priority issues

| Category | Priority issues |
|---|---|
| Achievement of a society that provides health and longevity | ① Innovative drug discovery infrastructure through functional control of biomolecular systems |
| | ② Integrated computational life science to support personalized and preventive medicine |
| Disaster prevention and global climate problem | ③ Development of integrated simulation systems for hazard and disaster induced by earthquake and tsunami |
| | ④ Advancement of meteorological and global environmental predictions utilizing observational “Big Data” |
| Energy problem | ⑤ Development of new fundamental technologies for high-efficiency energy creation, conversion/storage and use |
| | ⑥ Accelerated Development of Innovative Clean Energy Systems |
| Enhancement of industrial competitiveness | ⑦ Creation of new functional devices and high-performance materials to support next-generation industries |
| | ⑧ Development of Innovative Design and Production Processes that Lead the Way for the Manufacturing Industry in the Near Future |
| Development of basic science | ⑨ Elucidation of the fundamental laws and evolution of the universe |



# Organization




**Foreign Laboratories and Universities**
- Sys. Soft. (OS, Comm., …)
- Low Power, FT, …
- Prog. Env.
- Mini Apps.

**Japanese Universities**
(Pre-)Standardization of API/SPI, Benchmarks, etc.
- Power Control API, FT API, etc.
- Evaluation of Architecture & Co-design

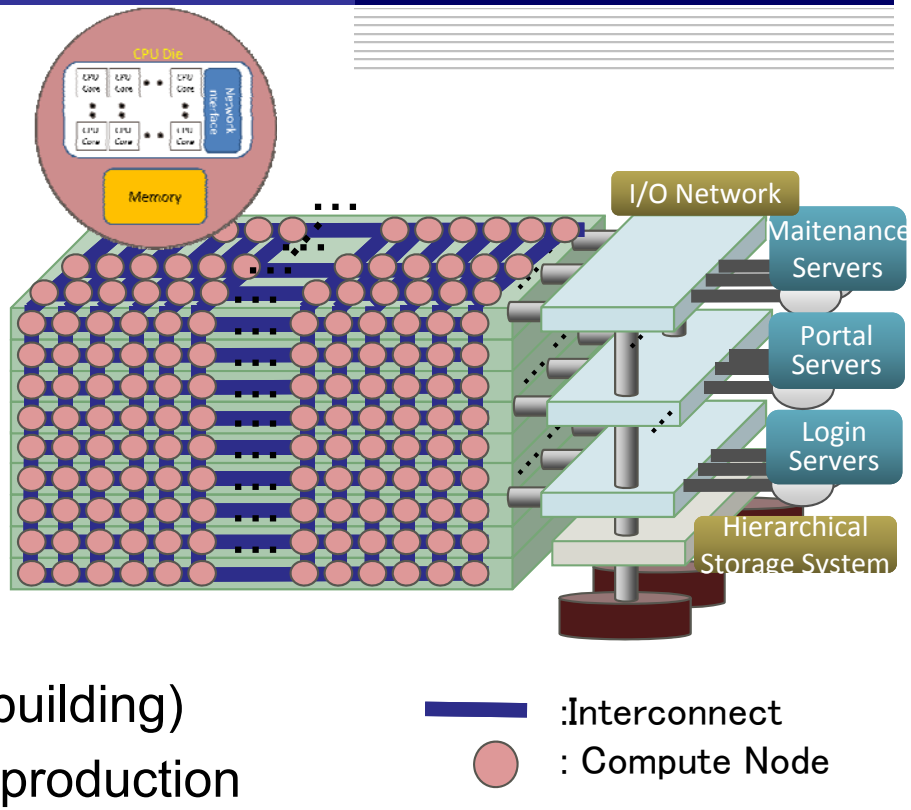
**Organizations pursuing Priority Issues**
- Co-design using target applications and optimization of primary applications
- Development novel algorithms

**HPCI Consortium**
- Feedback

**PC Cluster Consortium**
- Community Formation in Programming Environment and System Software

# An Overview of Post K

- ✓ CPU
  - Many-core with Interconnect interface
  - Power Knob feature
- ✓ Interconnect
  - TOFU (mesh/torus network)




- ✓ Constraints
  - Power capacity (about 30MW)
  - Space for system installation (in Kobe AICS building)
  - Budget (money) for development (NRE) and production
  - ... some degree of compatibility to the current K computer.
- ✓ The system should be designed to maximize the performance of applications in each computational science field.
  - "Co-design" is a keyword!

# Co-design Elements in Architecture

| | Target Application | | Co-design elements |
|---|---|---|---|
| | Program | Brief description | |
| ① | GENESIS | MD for proteins | Local and collective comm. and floating point(FP) performance |
| ② | Genomon | Genome processing (Genome alignment) | Workflow and file I/O |
| ③ | GAMERA | Earthquake simulator (FEM in unstructured & structured grid) | Comm. and memory bandwidth |
| ④ | NICAM+LETK | Weather prediction system using Big data (structured grid stencil & ensemble Kalman filter) | Comm. and memory bandwidth, SIMD, file I/O |
| ⑤ | NTChem | molecular electronic (structure calculation) | FP perf., SIMD, collective comm. |
| ⑥ | FFB | Large Eddy Simulation (unstructured grid) | Comm. and memory bandwidth, SIMD |
| ⑦ | RSDFT | an ab-initio program (density functional theory) | FP perf., collective comm. |
| ⑧ | Adventure | Computational Mechanics System for Large Scale Analysis and Design (unstructured grid) | Comm. and memory bandwidth, SIMD |
| ⑨ | CCS-QCD | Lattice QCD simulation (structured grid Monte Carlo) | Comm. and memory bandwidth, Local and collective comm. |

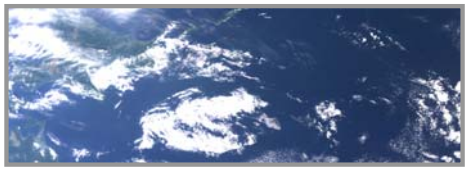
# File I/O for Big data: An Example

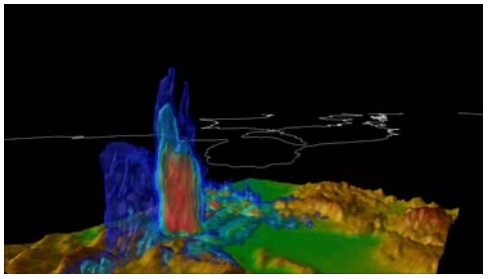
PI: Takemasa Miyoshi, RIKEN AICS
“Innovating Big Data Assimilation technology for revolutionizing very-short-range severe weather prediction”

**An innovative 30-second super-rapid update numerical weather prediction system for 30-minute/1-hour severe weather forecasting will be developed, aiding disaster prevention and mitigation, as well as bringing a scientific breakthrough in meteorology.**

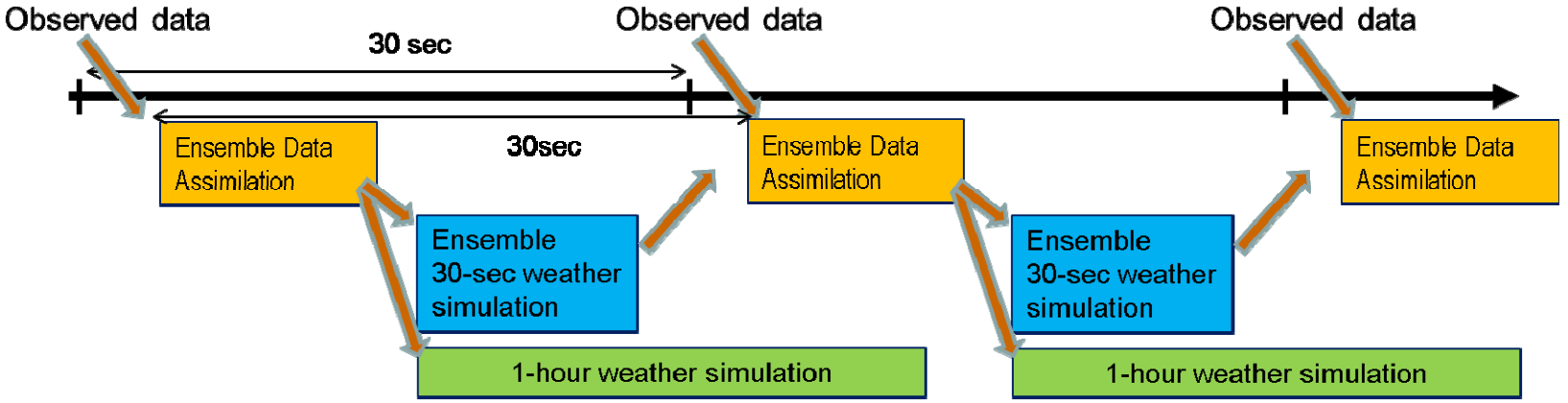
Phased array weather radar
151.2 MB/30sec

Satellite
500 MB/2.5min

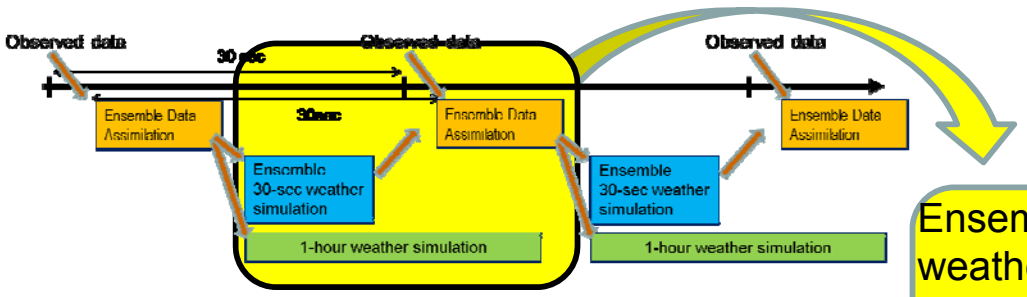




Rain particle



# File I/O patterns in Ensemble simulation and data assimilation




Observed data via Internet

Ensemble 30-sec weather simulation

DA (data assimilation) uses 7 time slots (17GB x 7)

N:1 processes 0.2K to 3.2 K node

Simulation Result

1:N

Data assimilation Result

100:N processes

N:100

100 cases

17 GB

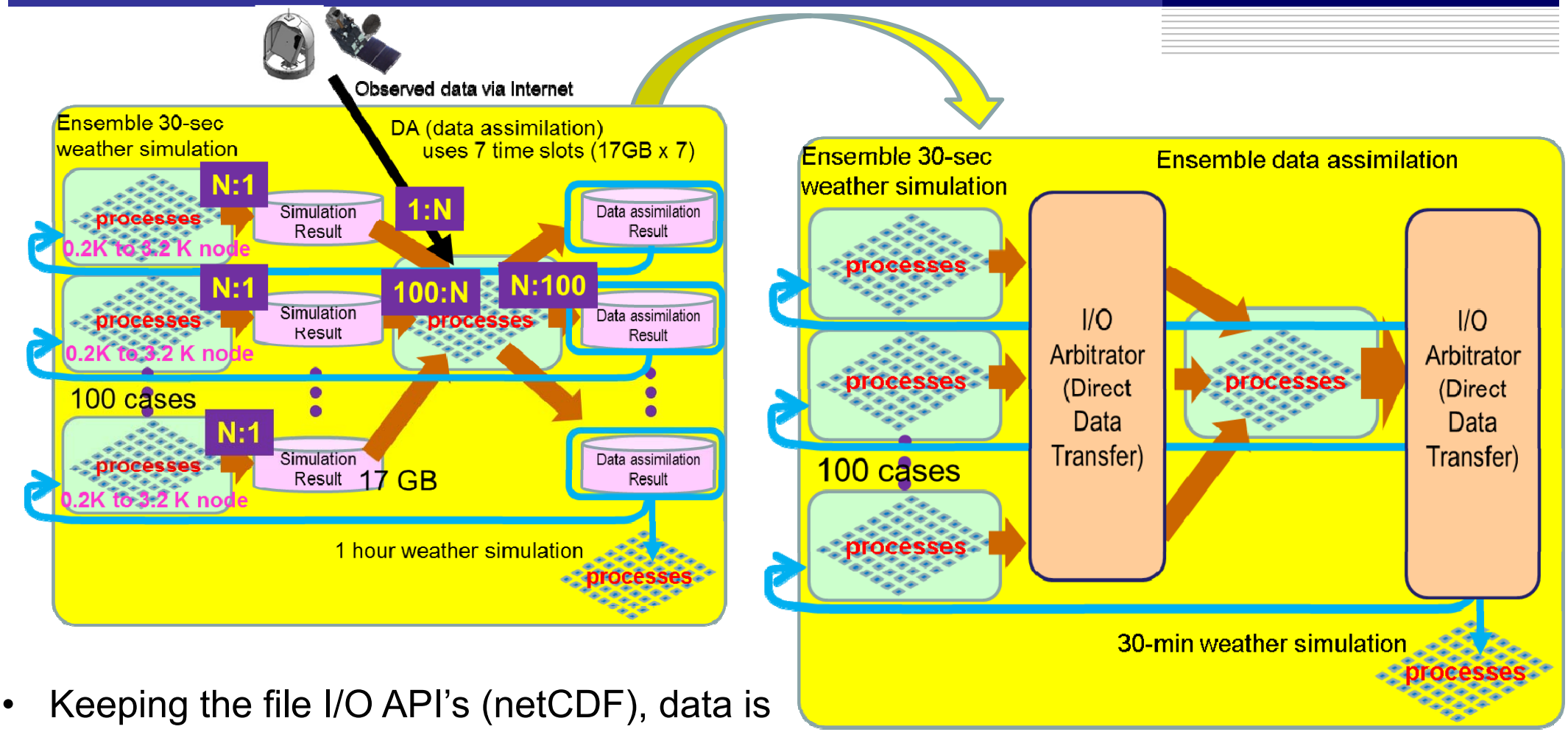
1 hour weather simulation processes

Every 30 second

1. Each weather simulation generates 17 GB data. 1.7 TB in total for 100 cases
2. Data assimilation code reads all results (17GB) and observed data (-- 1GB), and outputs assimilated data for 100 cases. That is, 1.7 TB in total for 100cases

If file I/O operations spend 10 second, the execution time of ensemble simulations must be shorter than 10 second.
If the file I/O time becomes shorter, execution time requirement of simulation and DA is relaxed

# Approach: I/O Arbitrator

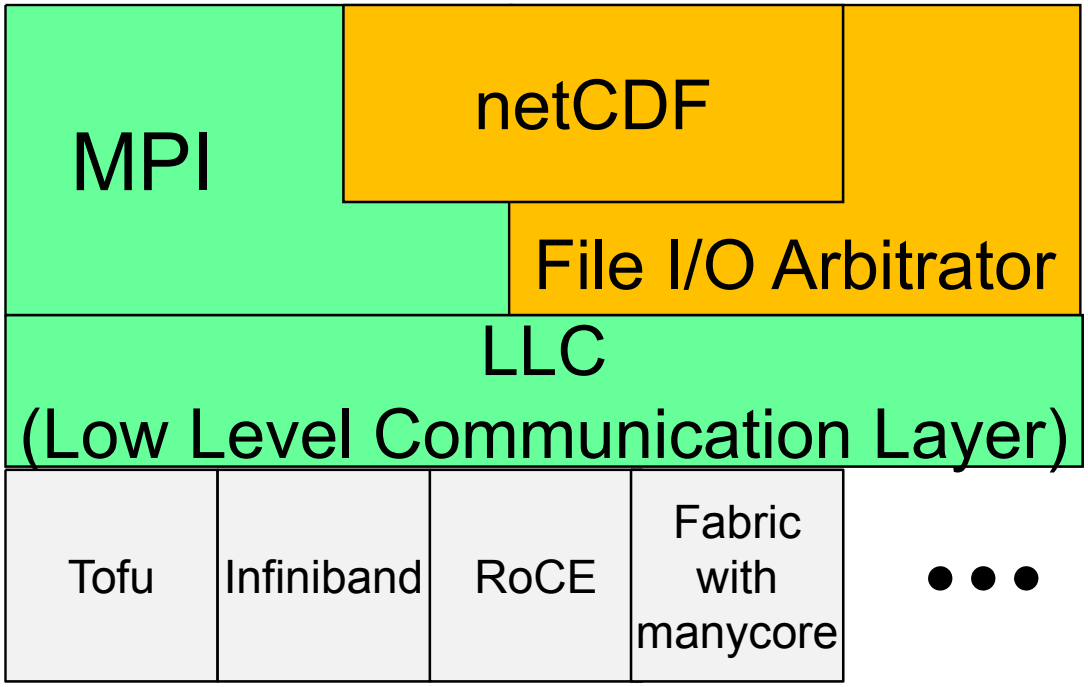



- Keeping the file I/O API's (netCDF), data is transferred from the ensemble simulation jobs to the data assimilation job without storing/loading data to/from the file system
  - Application programs are not modified

*Note: the current prototype system extends netCDF*

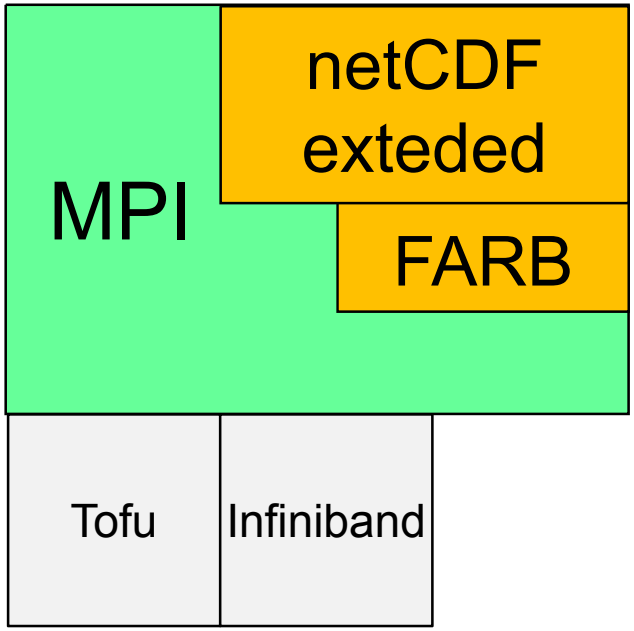
# Communication Architecture

## Target Communication Architecture




- MPICH: MPI Implementation at Argonne National Laboratory, USA
- Low Level Communication Layer: developed by RIKEN AICS under the Post K project

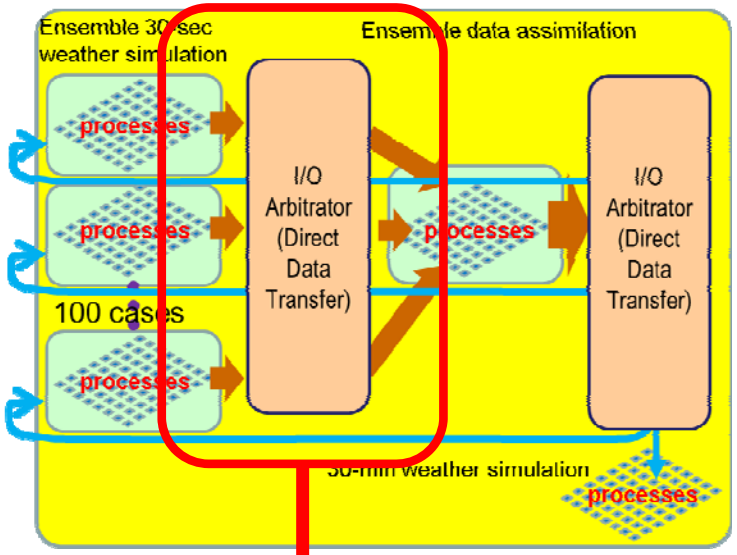
## Prototype System




- A prototype system extends the netCDF API. It has been implemented using the MPI communication library because LLC has not been available.
- The prototype system runs with any MPI implementations, such as MPICH and OpenMPI.

*Note: the current prototype system extends netCDF*

# Experimental Results

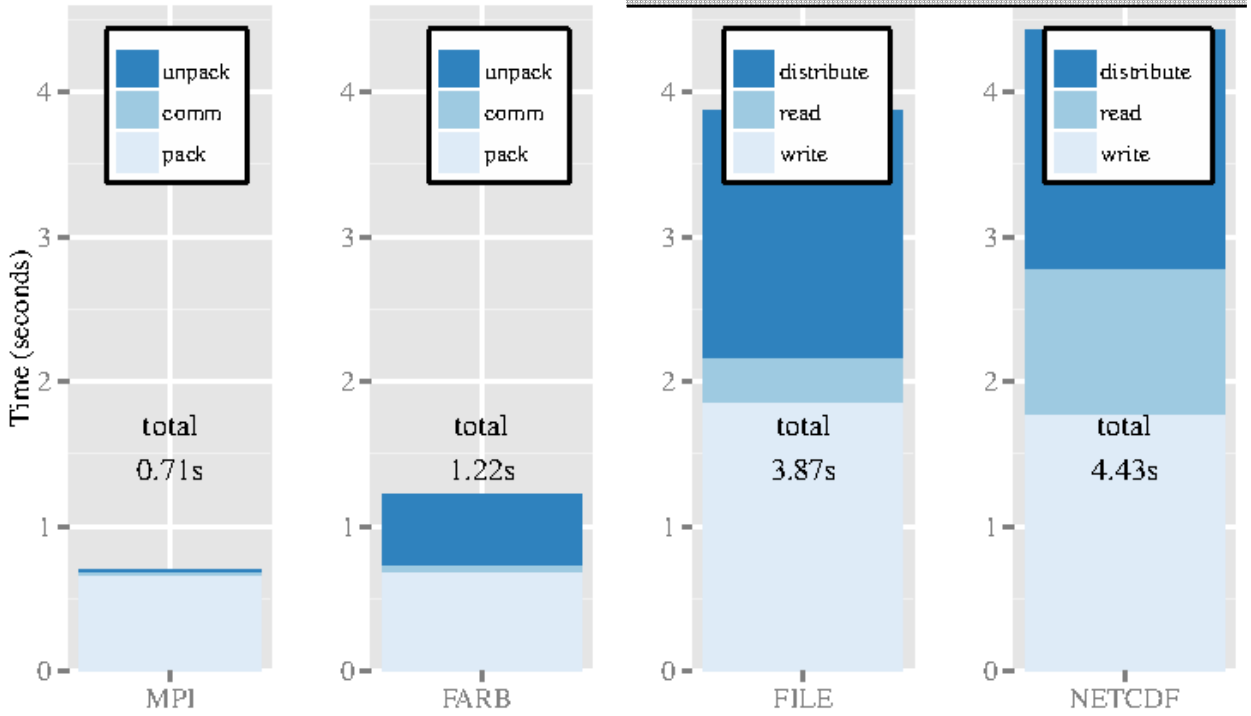



| netCDF | FARB | FILE |
|---|---|---|
| 4.43 sec | 1.22 sec | 3.87 sec |

- A benchmark program, performing the same communication pattern of data assimilation, has been implemented
- Four implementations:
  - FARB: using the proposed system
  - MPI: The same communication pattern is realized by MPI communication library
  - FILE: data is transferred via files
  - NETCDT: data is transferred via netCDF

Evaluation Environment
Oakleaf-fx (University of Tokyo)

| CPU | SPARC64 IXfx |
|---|---|
| Number of Cores | 16 |
| Clock | 1.848 GHz |
| Memory | 32 GB |
| Network Link | 5 GB/sec |
| Topology | 6-D Mesh/Torus |




MPI implementation does not have overheads to interpret the communication pattern requested by the benchmark code, and thus it is the best performance

# Concluding Remarks

- Based on experiments of the prototype system, a direct data transfer between jobs in netCDF will be developed
- The design of observed data transfer via Internet is also important to meet the requirement of real-timeness