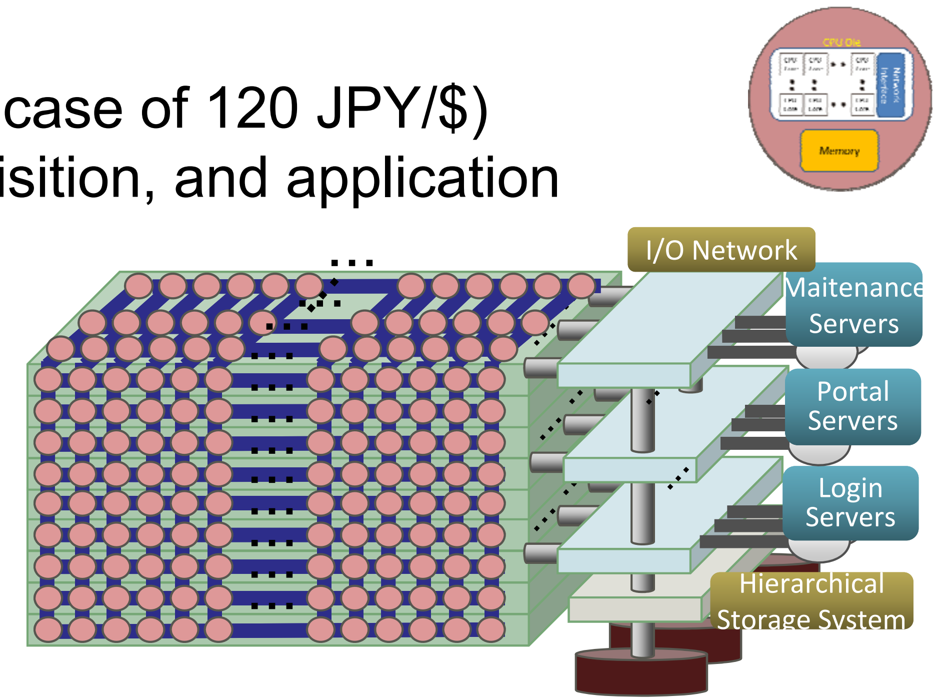
The Post K Project and Its Big Data Aspect

Yutaka Ishikawa

RIKEN AICS

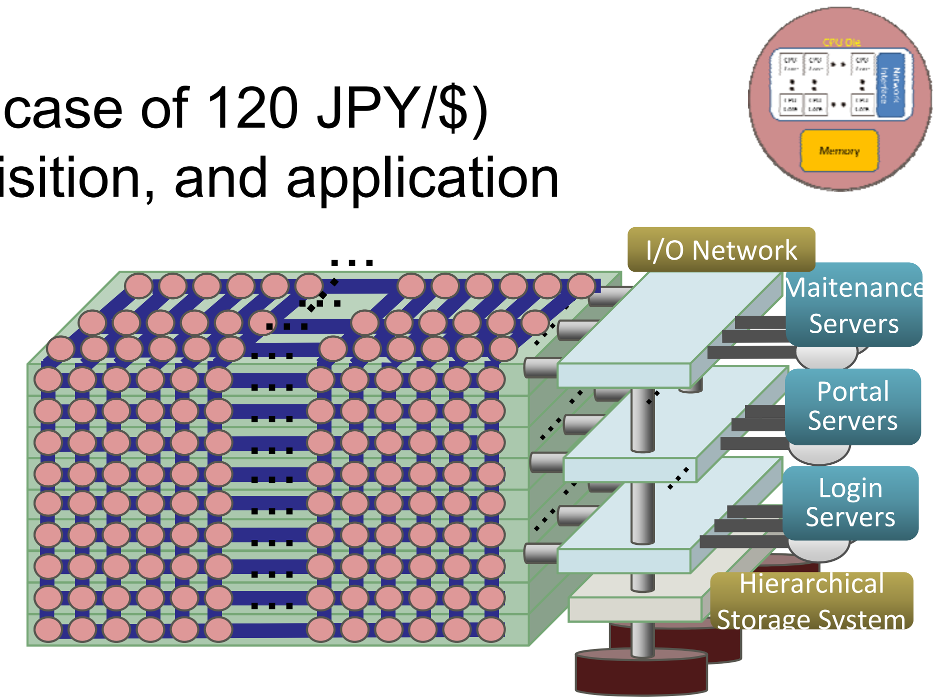
FLAGSHIP2020 Project

- **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



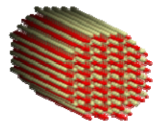
CY	2014				2015				2016				2017				2018				2019				2020			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4

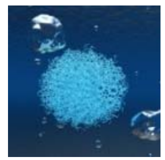
Basic Design | Design and Implementation | Manufacturing, Installation, and Tuning | Operation

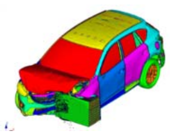
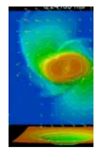
Social and scientific priority issues

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

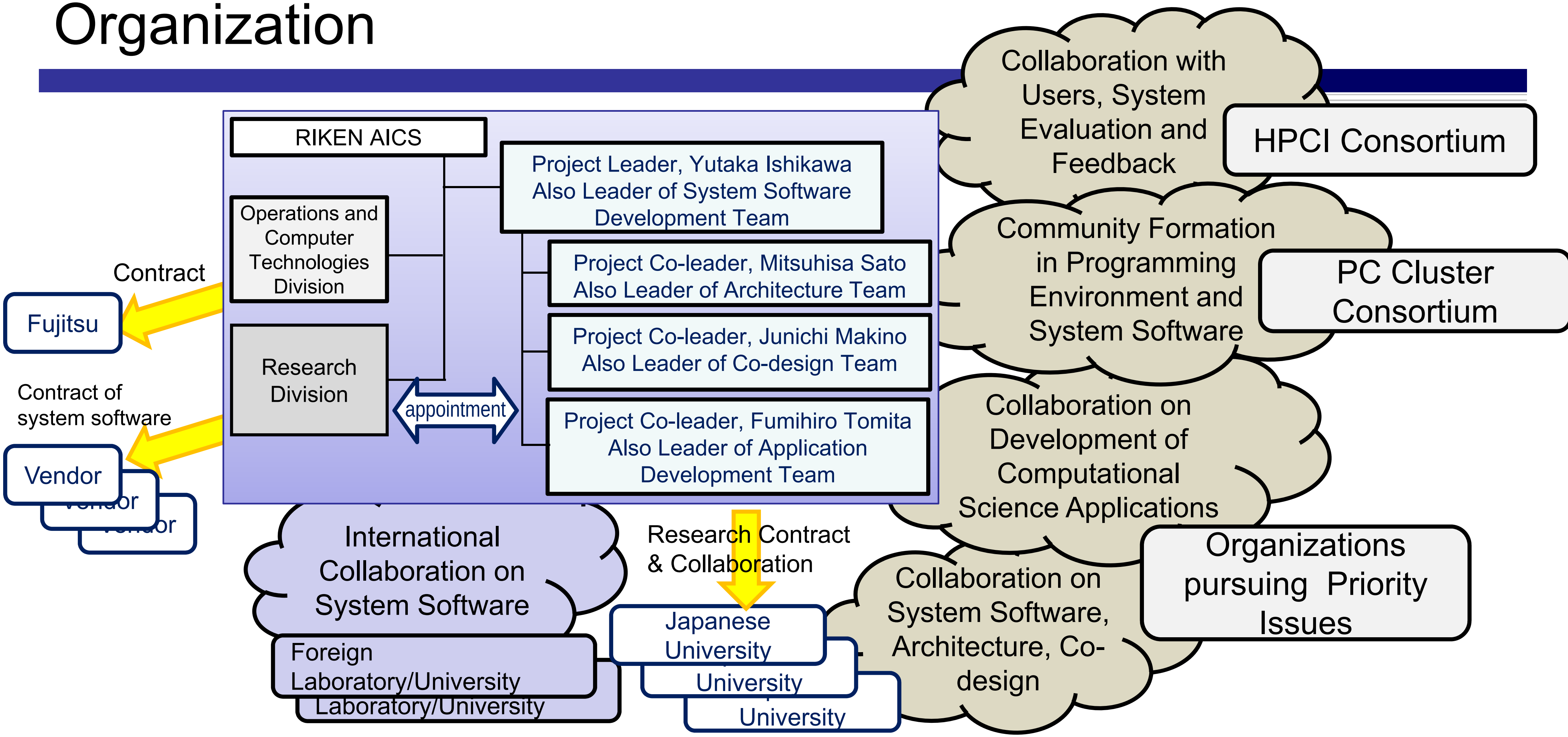




Organization



Foreign Laboratories and Universities

- Sys. Soft. (OS, Comm., …) ・Prog. Env.
- Low Power, FT, … ・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

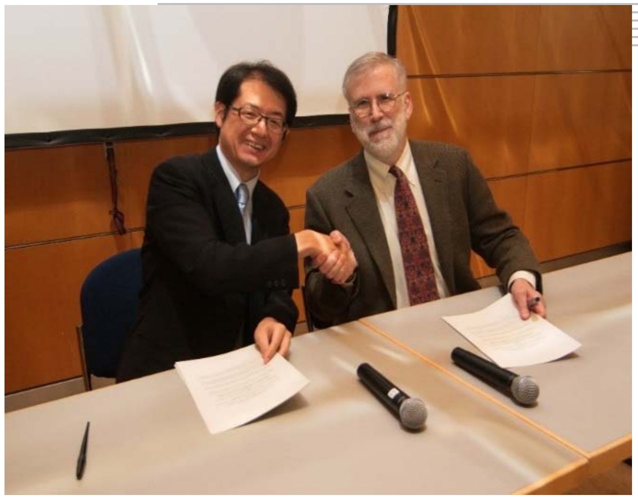
PC Cluster Consortium

- Community Formation in Programming Environment and System Software

International Collaboration between DOE and MEXT

PROJECT ARRANGEMENT
UNDER THE IMPLEMENTING ARRANGEMENT
BETWEEN
THE MINISTRY OF EDUCATION, CULTURE, SPORTS, SCIENCE AND TECHNOLOGY OF JAPAN
AND
THE DEPARTMENT OF ENERGY OF THE UNITED STATES OF AMERICA
CONCERNING COOPERATION IN RESEARCH AND DEVELOPMENT IN ENERGY AND RELATED FIELDS

CONCERNING COMPUTER SCIENCE AND SOFTWARE RELATED TO CURRENT AND FUTURE HIGH PERFORMANCE COMPUTING FOR OPEN SCIENTIFIC RESEARCH



Yoshio Kawaguchi (MEXT, Japan) and William Harrod (DOE, USA)

Purpose: Work together where it is mutually beneficial to expand the HPC ecosystem and improve system capability

- Each country will develop their own path for next generation platforms
- Countries will collaborate where it is mutually beneficial

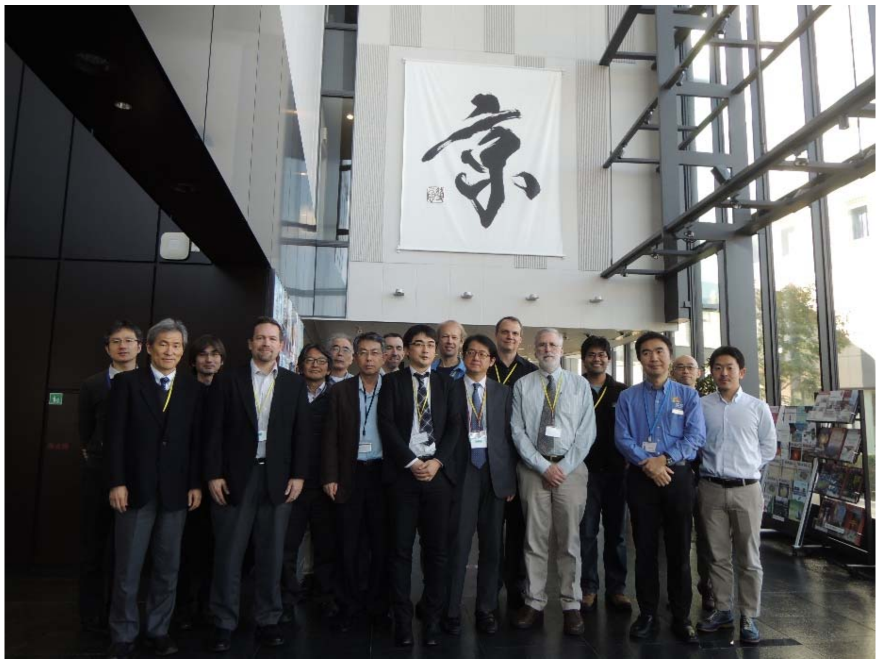
- Joint Activities
  - Pre-standardization interface coordination
  - Collection and publication of open data
  - Collaborative development of open source software
  - Evaluation and analysis of benchmarks and architectures
  - Standardization of mature technologies

Technical Areas of Cooperation

- Kernel System Programming Interface
- Low-level Communication Layer
- Task and Thread Management to Support Massive Concurrency
- Power Management and Optimization
- Data Staging and Input/Output (I/O) Bottlenecks
- File System and I/O Management
- Improving System and Application Resilience to Chip Failures and other Faults
- Mini-Applications for Exascale Component-Based Performance Modelling

List of Presentations at the first coordination committee

1. Operating System and Runtime
   - Coordinators: Pete Beckman (ANL) and Yutaka Ishikawa (RIKEN)
   - Leaders: Kamil Iskra (ANL) and Balazs Gerofi (RIKEN)
2. Power Monitoring, Analysis and Management
   - Coordinators: Martin Schulz (LLNL) and Hiroshi Nakamura (U. Tokyo)
   - Leaders: Martin Schulz (LLNL), Barry Rountree (LLNL), Masaaki Kondo (U. tokyo), and Satoshi Matsuoka (TITECH)
3. Advanced PGAS runtime and API
   - Coordinators: Peter Beckman (ANL) and Mitsuhisa Sato (RIKEN)
   - Leaders: Laxmikant Kale (UIUC), Barbara Chapman (U. Huston)
4. Storage and I/O
   - Coordinators: Rob Ross (ANL) and Osamu Tatebe (U. Tsukuba)
   - Leaders: Rob Ross (ANL) and Osamu Tatebe (U. Tsukuba)
5. I/O Benchmarks and netCDF implementations for Scientific Big Data
   - Coordinators: Choudary (North Western U.) and Yutaka Ishikawa (RIKEN)
   - Leaders: Choudary (North Western U.) and Yutaka Ishikawa (RIKEN)
6. Enhancements for Data Movement in Massively Multithreaded Environments
   - Coordinators: Peter Beckman (ANL) and Satoshi Matsuoka (TITECH)
   - Leaders: Pavan Balaji (ANL) and Satoshi Matsuoka (TITECH)
7. Performance Profiling Tools, Modeling and Database
   - Coordinators: Jeffery Vetter (ORNL) and Satoshi Matsuoka (TITECH)
   - Leaders: Jefery Vertter (ORNL), Martin Shultz (LLNL), Satoshi Matsuoka (TITECH), and Naoya Maruyama (RIKEN)
8. Mini- /Proxy-Apps for Exascale Codesign
   - Coordinators: Jefery Vetter (ORNL) and Satoshi Matsuoka (TITECH)
   - Leaders: <TBA> and Naoya Maruyama (RIKEN)
9. Extreme-Scale Resilience for Billion-Way Parallelism
   - Coordinators: Martin Schulz (LLNL) and Satoshi Matsuoka (TITECH)
   - Leaders:
10. Scalability and performance enhancements to communication library
    - Coordinators: Pete Beckman (ANL) and Yutaka Ishikawa (RIKEN)
    - Leaders: Pavan Balaji (ANL) and Masamichi Takagi (RIKEN)
11. Communication Enhancements for Irregular/Dynamic Environments
    - Coordinators: Pete Beckman (ANL) and Yutaka Ishikawa, RIKEN
    - Leaders: Pavan Balaji (ANL) and Atsushi Hori (RIKEN)



Joining JLESC

Joint Laboratory for Extreme Scale Computing

- Members
  - University of Illinois at Urbana-Champaign, INRIA, Argonne National Laboratory, Barcelona Supercomputing Center and Jülich Supercomputing Centre

- RIKEN AICS Activity
  - RIKEN’s participation has ben approved by the executive committee
  - After signing MOU, RIKEN will propose collaboration areas

Co-design Elements in Architecture

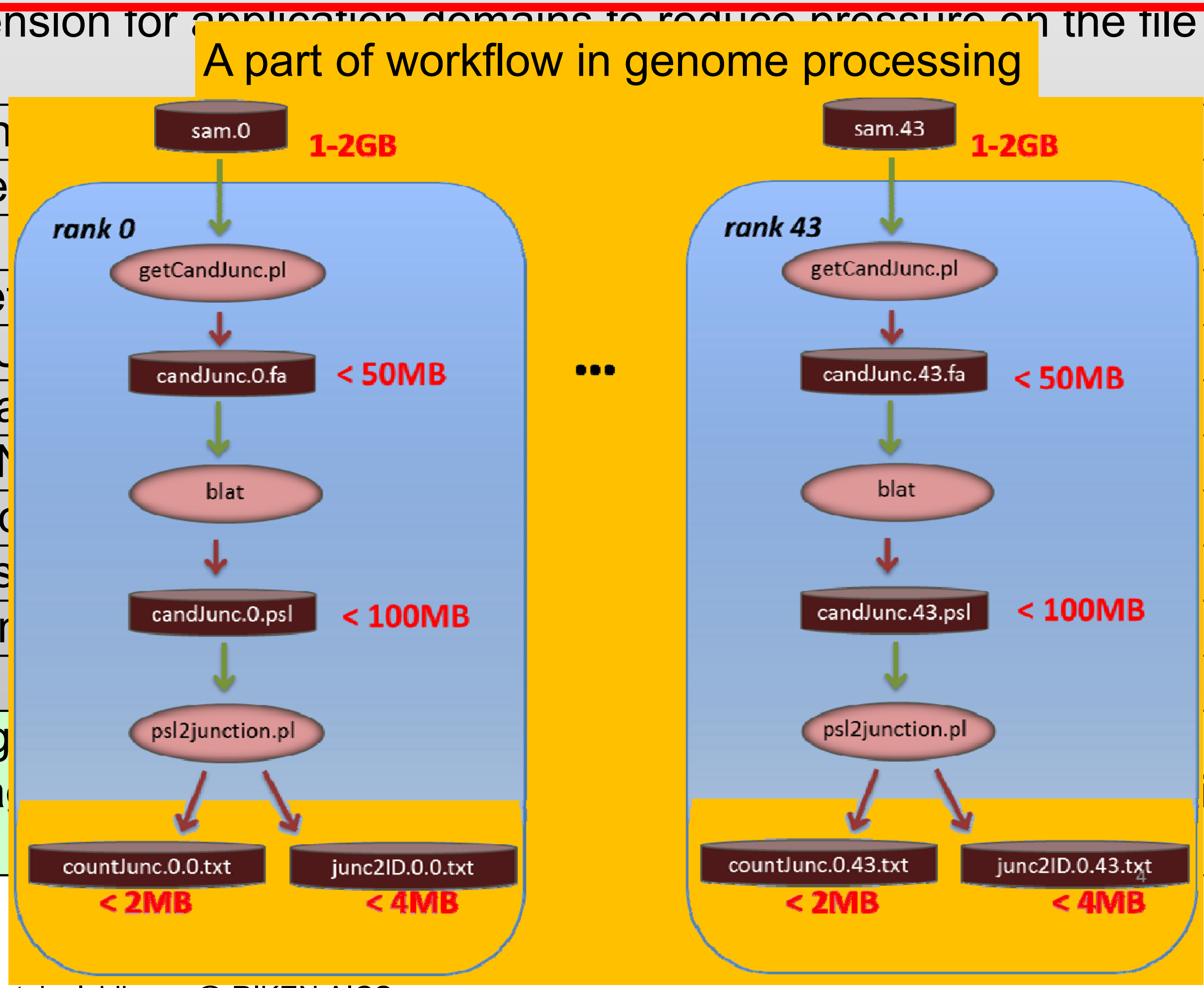
	Target Application		Co-design elements
	Program	Brief description	
①	GENESIS	MD for proteins	Local and collective comm. and floating point(FP) performance
②	Genomon	Genome processing (Genome alignment)	Workflow and file I/O
③	GAMERA	Earthquake simulator (FEM in unstructured & structured grid)	Comm. and 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.

Co-design Elements in System Software

	Co-design Item
File I/O	Async. I/O, Caching/Buffering to reduce pressure on I/O network and the file system
	Location of temporal files based on workflow and memory availability (possibility of RAM disk)
	netCDF API and its extension for application domains to reduce pressure on the file system
	Data Exchange between applications (Coupling)
	Methods for massive files
Communication	In-situ Visualization
	Data transfer via Internet, e.g. genome sequencers, radars, satellites, XFEL
	Optimization of Many NUMA domains
	Applicability of RDMA-based Communication
OS Kernel	CPU Scheduler or not（NO OS Noise but no CPU scheduler)
	Special memory allocation for NUMA domain process/thread
	Supporting efficient consecutive job execution
	Efficient MPI environment within a node
	PGAS model
System Configuration	After the above co-designs, the system configuration, such as I/O network performance, local storage performance and capacity, hierarchical storage system, is decided

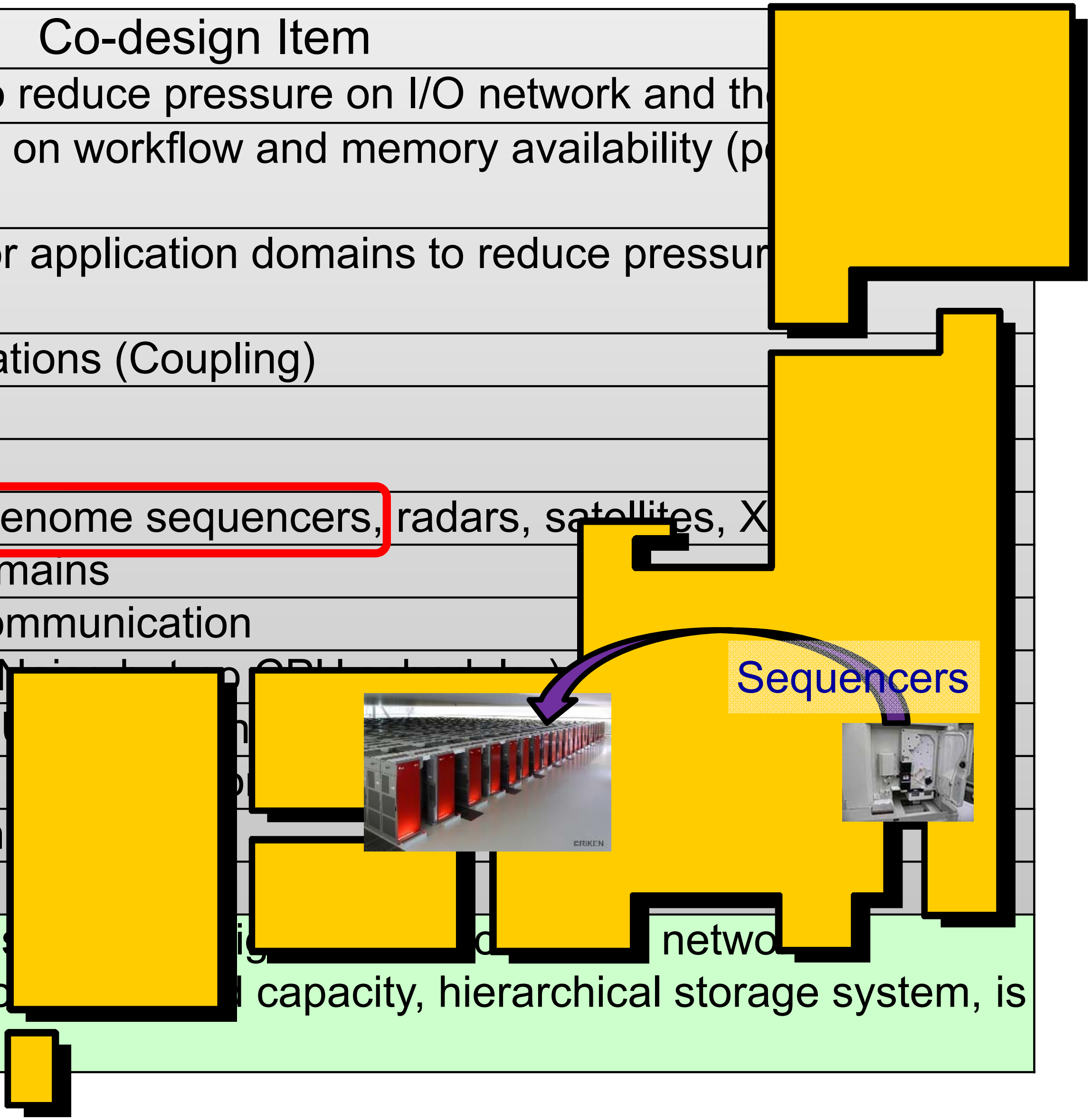
Co-design elements in System Software

	Co-design Item
File I/O	Async. I/O, Caching/Buffering to reduce pressure on I/O network and the file system
	Location of temporal files based on workflow and memory availability (possibility of RAM disk)
	netCDF API and its extension for application domains to reduce pressure on the file system
	Data Exchange between
	Methods for massive file
Communication	In-situ Visualization
	Data transfer via Interne
	Optimization of Many NU
	Applicability of RDMA-ba
OS Kernel	CPU Scheduler or not（N
	Special memory allocatio
	Supporting efficient cons
	Efficient MPI environmen
	PGAS model
System Configuration	After the above co-desig ... is performance, local stora ... decided



Co-design elements in System Software

	Co-design Item
File I/O	Async. I/O, Caching/Buffering to reduce pressure on I/O network and th
	Location of temporal files based on workflow and memory availability (p RAM disk)
	netCDF API and its extension for application domains to reduce pressur system
	Data Exchange between applications (Coupling)
	Methods for massive files
Communication	In-situ Visualization
	Data transfer via Internet, e.g. genome sequencers, radars, satellites, X
	Optimization of Many NUMA domains
	Applicability of RDMA-based Communication
OS Kernel	CPU Scheduler or not（NO OS N
	Special memory allocation for N
	Supporting efficient consecutive
	Efficient MPI environment within
	PGAS model
System Configuration	After the above co-designs, the s … netwo … performance, local storage perfo … capacity, hierarchical storage system, is decided



Co-design elements in System Software

	Co-design Item
File I/O	Async. I/O, Caching/Buffering to reduce pressure on I/O network and the file system
	Location of temporal files based on workflow and memory availability (possibility of RAM disk)
	netCDF API and its extension for application domains to reduce pressure on the file system
	Data Exchange between applications (Coupling)
	Methods for massive files
Communication	In-situ Visualization
	Data transfer via Internet, e.g. radars, satellites, genome sequencers, XFEL
	Optimization of Many NUMA domains
	Applicability of RDMA-based Communication
OS Kernel	CPU Scheduler or not（NO OS Noise but no CPU scheduler)
	Special memory allocation for NUMA domain process/thread
	Supporting efficient consecutive job execution
	Efficient MPI environment within a node
	PGAS model
System Configuration	After the above co-designs, the system configuration, such as I/O network performance, local storage performance and capacity, hierarchical storage system, is decided

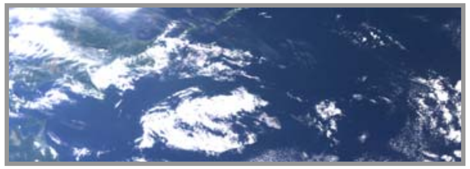
File I/O for Big data: An Example

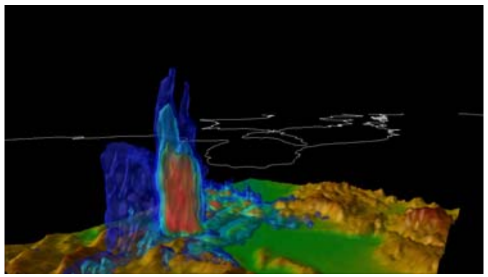
PI: Takemasa Miyoshi, RIKEN AICS
"Innovating Big Data Assimilation technology for revolutionizing very-short-range severe weather prediction"

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.

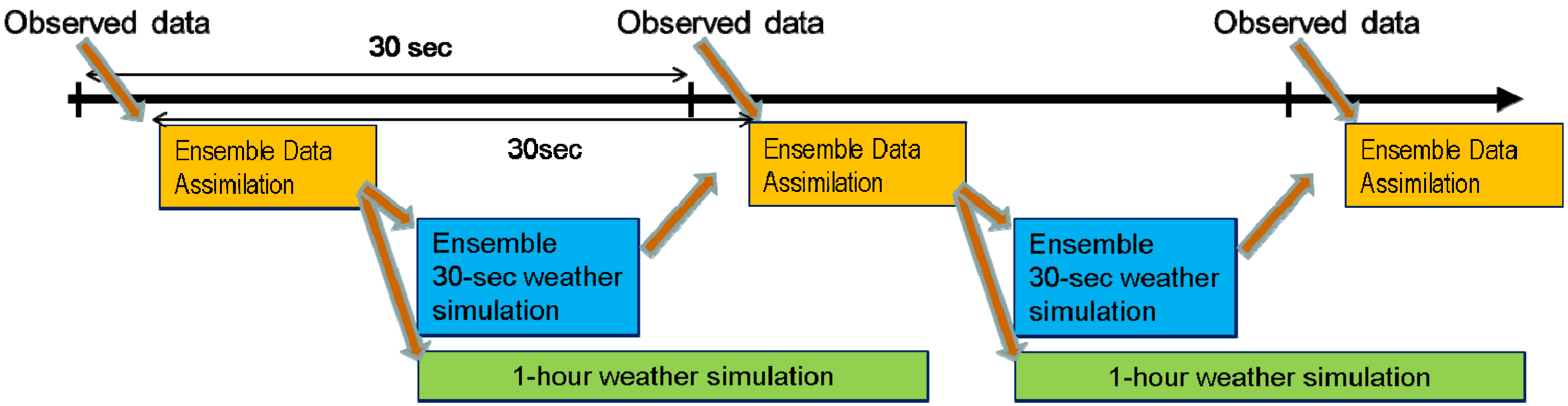
Phased array weather radar
151.2 MB/30sec

Satellite
500 MB/2.5min

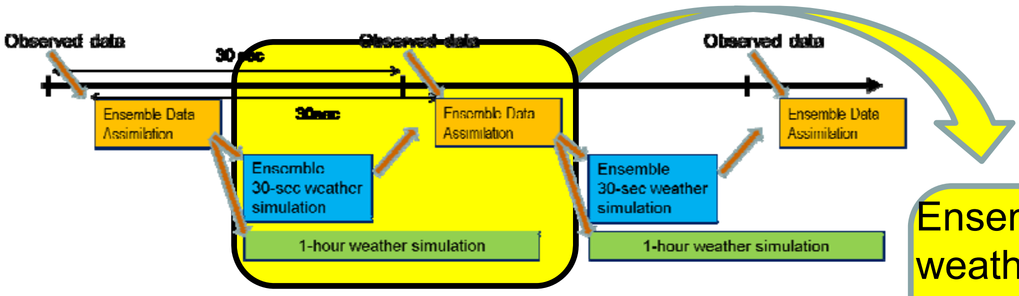


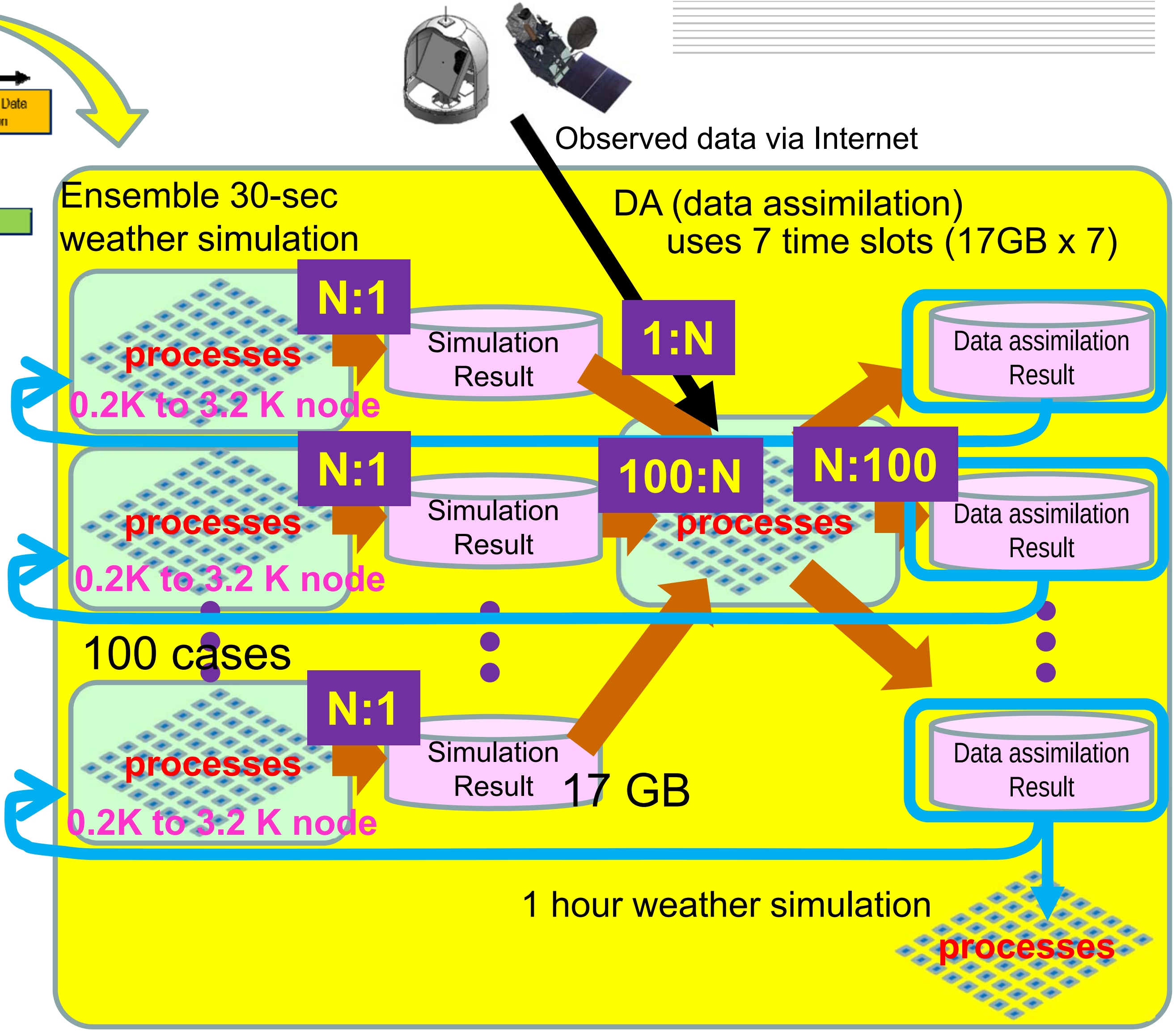


Rain particle



File I/O patterns in Ensemble simulation and data assimilation

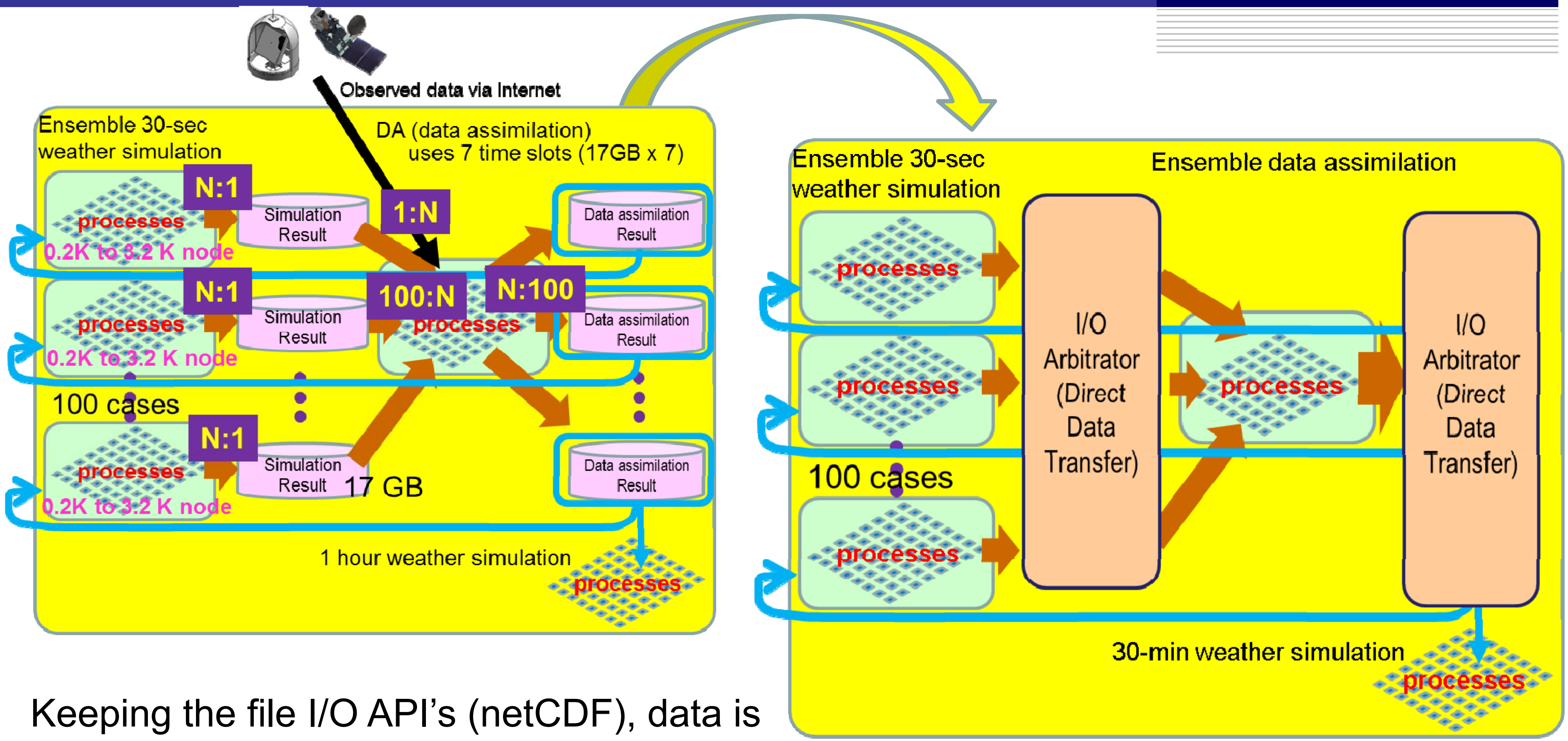




Every 30 second

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

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

Concluding Remark

- The basic architecture design and target application performances will be decided by 2015 Summer