US-Japan Joint Institute for Fusion Theory Workshop on
Innovations and co-designs of fusion simulations towards extreme scale computing
August 20-21, 2015
Nagoya University, Nagoya, Japan

FLAGSHIP 2020 project -- Development of Japanese National Flagship supercomputer "post K" --

Mitsuhisa Sato

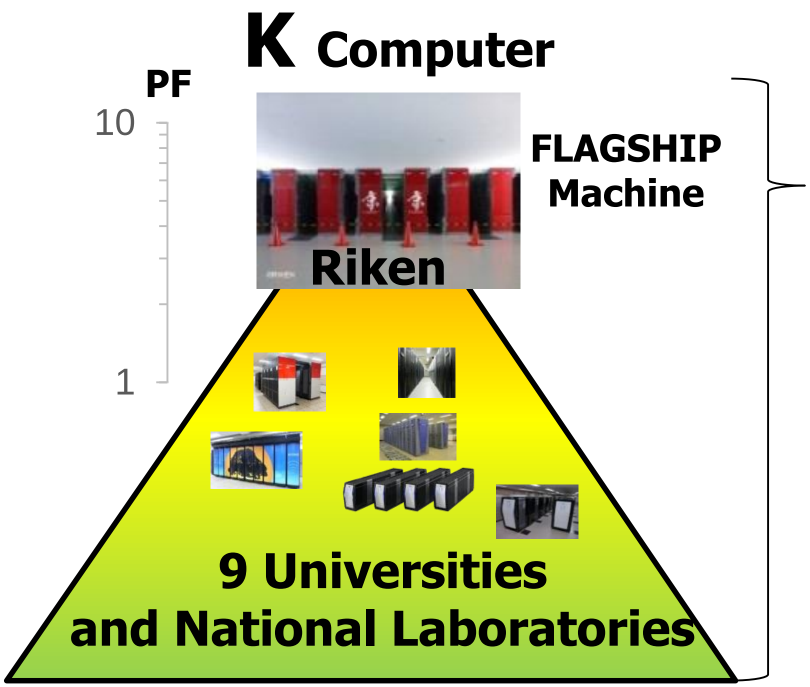
Team Leader of Architecture Development Team,
Exascale supercomputer project
RIKEN Advance Institute of Computational Science (AICS)

Outline

- Background: Japanese supercomputing infrastructure, HPCI
- FLAGSHIP 2020 project
  - to develop the next Japanese flagship computer system, “post-K”
- “co-design” effort to design the system
- International collaborations

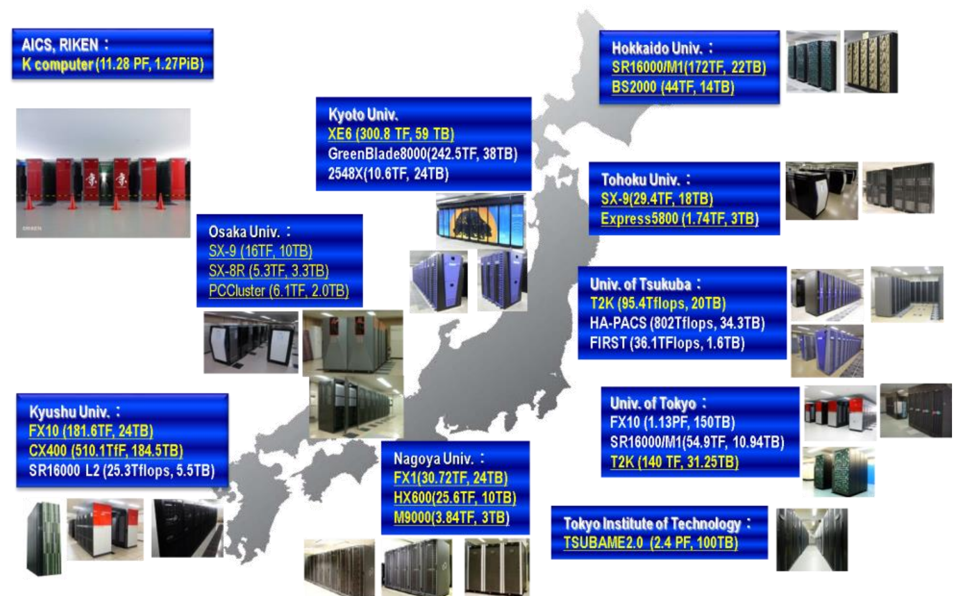
AICS and Supercomputer Centers in Japanese Universities

Supercomputers in Japan



HPCI (High Performance Computing Infrastructure)
is formed from those machines, called leading machines

Features: Single sign-on
Shared storage (Distributed file system)

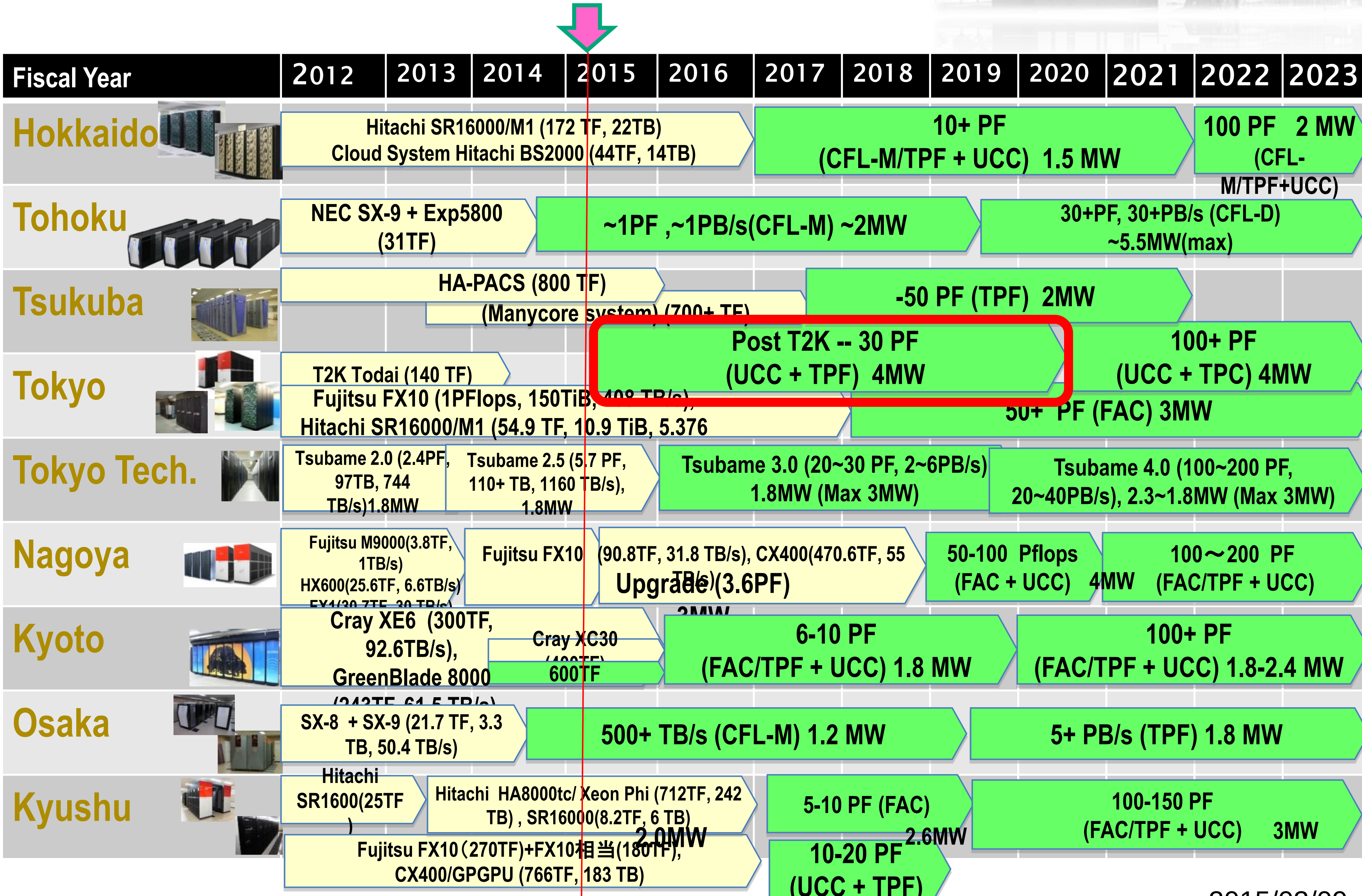


As of Jun 2012

- **Each supercomputer center has one, two or more supercomputers.**
- **Each supercomputer center replaces their machines every 4.5 to 6 years.**

京コンピュータ “The K computer"

Supercomputer Centers operated at Japanese Universities and Plan/schedule



FLAGSHIP 2020 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 our country.

- Planned Budget
  - 110 Billion JPY (about 0.91 Billion USD at the rete 120 JPY/$)
  - including research, development (NRE) and acquisition/deploy, and application development

- Post K Computer: System and Software
  - RIKEN AICS is in charge of development
  - Fujitsu is selected as a vendor partner
  - Started from 2014

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

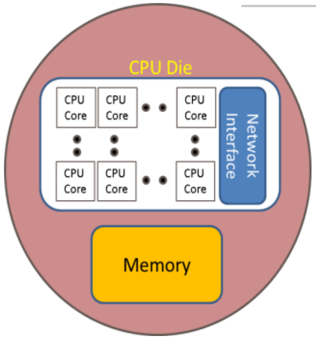
Current status of the post-K project

- The procurement for the development of the post-K computer system was done.
  - Fujitsu was selected as the vender partner.

- In the specification of RFP:
  - Constraints are:
    - 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.
- We are now working on the “basic design” of the system with the vender partner.
- The system should be designed to maximize the performance of applications in each computational science field.
  - "Co-design" is a keyword!

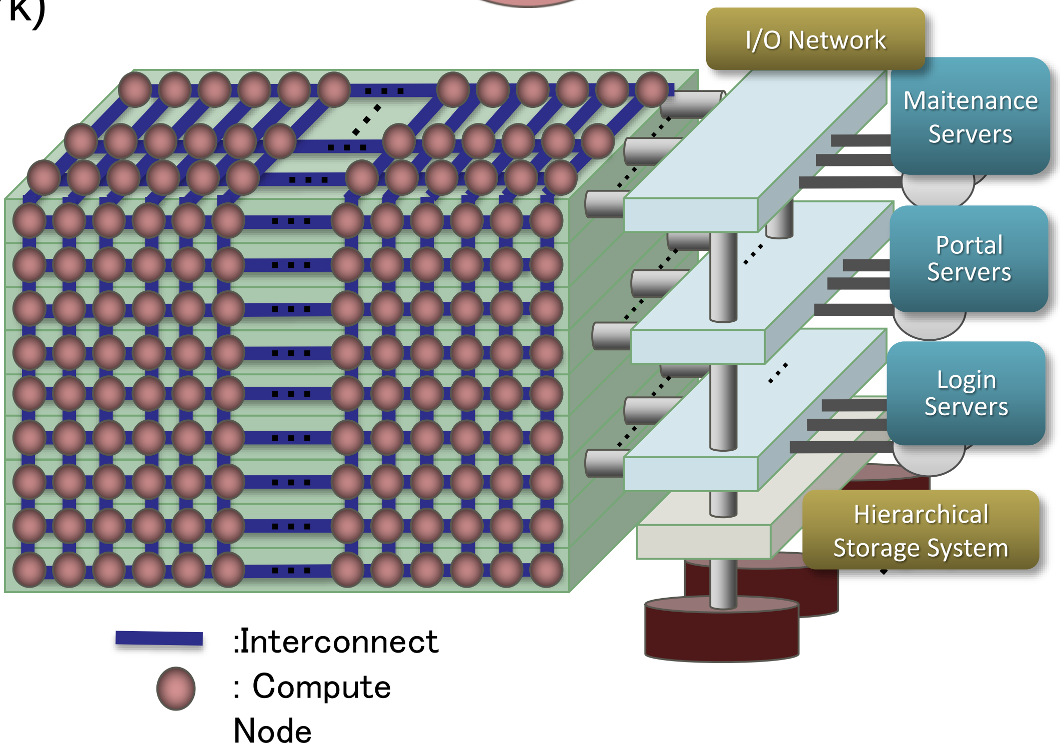
Post K Computer

- ✓ CPU
  - Many-core with Interconnect interface integrated on chip
  - Power Knob feature for saving power
- ✓ Interconnect
  - TOFU (mesh/torus network)



Co-design may include:

- Compute Node Features
  - Core architecture, FP performance
  - Memory hierarchy, control, capacity, and bandwidth
- Network Performance
- I/O Performance



Co-design in HPC

- "Co-design" in Wikipedia
  - "**Co-design** or **codesign** is a product, service, or organization development process where design professionals empower, encourage, and guide users to develop solutions for themselves."

    … ….

  - "The phrase co-design is also used in reference to the simultaneous development of interrelated software and hardware systems. The term co-design has become popular in mobile phone development, where the two perspectives of hardware and software design are brought into a co-design process"

- The co-design of HPC must optimize and maximize the benefits to cover many applications as possible.
  - different from "co-design" in embedded systems. For example, in embedded field, co-design sometimes includes "specialization" for a particular applications.
  - On the other hands, in HPC, the system will be shared by many applications.

Why “co-design” is needed in very high-end HPC and exascale?

- In modern very high-end parallel system, more performance can be delivered (even upto “exascale”) by increasing the number of nodes, but …

- We need to design the system by trade-off between “energy/power” and “cost” and performance
  - to satisfy constraints of “energy/power” and “cost”
  - to maximize the performance.

We need to design the system by taking characteristics of applications in account

⇒ “codesign” in HPC

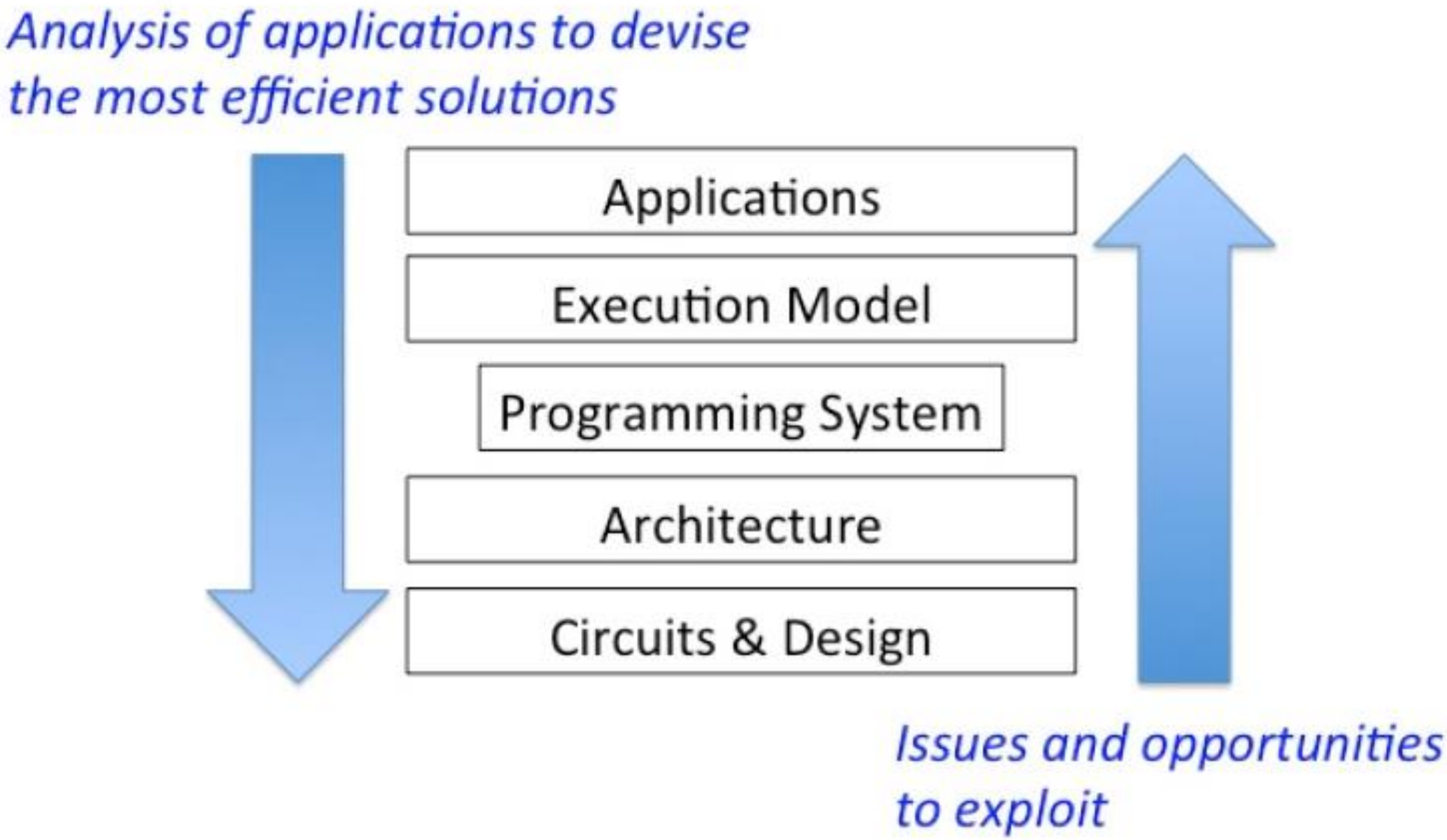
- The elements of "co-design" in our post-K project may include ....
  - Note that we are going to design processor/network and system with **t**he selected partner vender.
  - Different from supercomputer acquisition in universities.

Co-design elements in HPC systems

- Hardware/architecture
  - Node architecture (#core, #SIMD, etc...)
  - cache (size and bandwidth)
  - network (topologies, latency and bandwidth)
  - memory technologies (HBM and HMC, ...)
  - specialized hardware
  - #nodes
  - Storage, file systems
  - ... system configurations
- System software
  - Operating system for many core architecture
  - communication library (low level layer, MPI, PGAS)
  - Programming model and languages
  - DSL, ...
- Algorithm and math lib
  - Dense and Sparse solver
  - Eigen solver
  - ... Domain-specific lib and framework
- And, Applications!

Co-design in HPC

- Richard F. BARRETT, et.al. "On the Role of Co-design in High Performance Computing", *Transition of HPC Towards Exascale Computing* より



What applications does our co-design target for?

- SPIRE (Strategic Programs for Innovative Research) Program for the K computer
  - The projects were organized around 2011.

- For the post-K system,
  - The committee (from academia and industry) was organized by our government to identify "priority research area" (9) and "frontier research area"(5) to be exploited by the post-K system.

  - The call for project proposals for these "priority research area" and "frontier research area" has been issued.
  - The projects for “priority research area” were accepted for the design of target apps and the co-design of the post-K system.

SPIRE
(Strategic Programs for Innovative Research)

Purpose

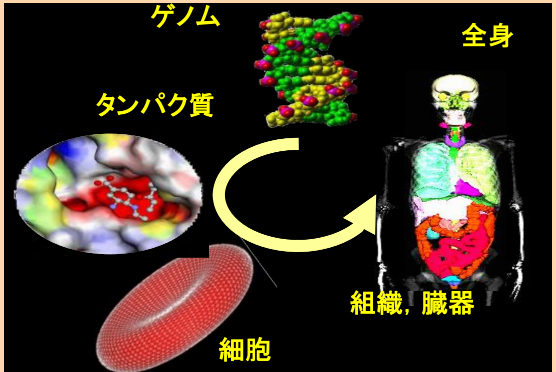
- ✓ To produce scientific results as soon as HPCI starts its operation
- ✓ To establish several core institutes for computational science

Outline of this program

- ✓ Identify the five strategic research areas which will contribute to produce results to scientific and social Issues
- ✓ A nation wide research groups are formed by funding the core organization designated by MEXT.
- ✓ The groups are to promote R&D using K computer and to construct research structures for their own area
- ✓ 50% computing resources of the K computer will be dedicated to this program

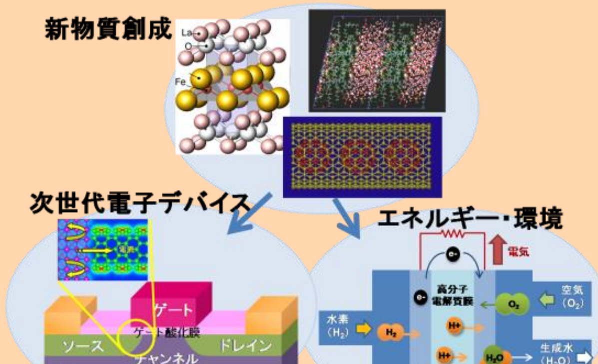
Five strategic areas of SPIRE

Life science/Drug manufacture



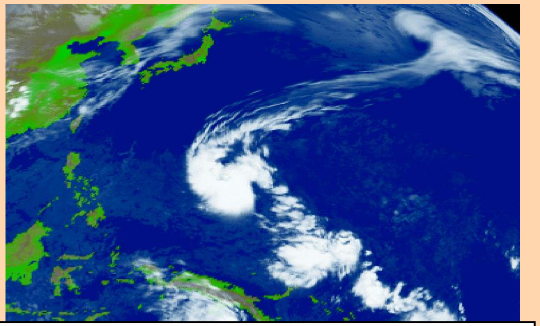
Toshio YANAGIDA (RIKEN)

New material/energy creation



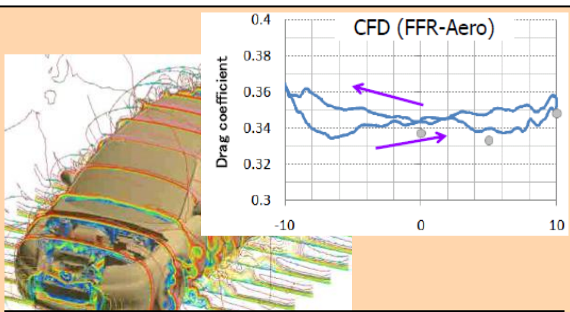
Shinji TSUNEYUKI (University of Tokyo)

Global change prediction for disaster prevention/mitigation



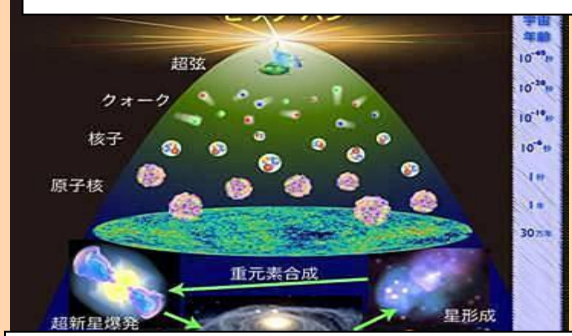
Shiro IMAWAKI (JAMSTEC)

Monodukuri (Manufacturing technology)



Chisachi KATO (University of Tokyo)

The origin of matter and the universe

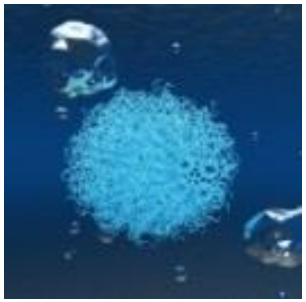
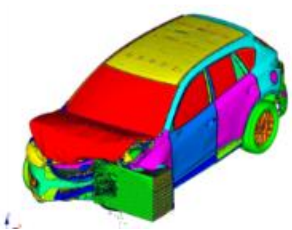


Shinya AOKI (University of

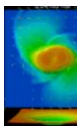
9 social and scientific priority issues (1/3)

Category	Priority issues
Achievement of a society that provides health and longevity 	**① Innovative drug discovery infrastructure through functional control of biomolecular systems** Develop ultra-high speed molecular simulations to achieve not only functional inhibition but also functional control of many biomolecules including factors that cause side-effects, in order to discover safe and highly effective drugs.
	② Integrated computational life science to support personalized and preventive medicine Exploit large-scale analysis of healthcare and medical "Big Data" and biomedical simulations (heart, brain and nervous system etc.) on the basis of optimal models obtained using these data, in order to support medicine tailored to each individual and preventive medicine that can extend healthy life expectancy.
Disaster prevention and global climate problem 	**③ Development of integrated simulation systems for hazard and disaster induced by earthquake and tsunami** Develop an integrated simulation system for hazard and disaster which are induced by earthquake and tsunami and are not estimated based on past experiences, by improving and strengthening a package of related analysis methods. The system is to be implemented in disaster management systems of the Cabinet Office and local governments, etc.
	④ Advancement of meteorological and global environmental predictions utilizing observational "Big Data" Build an infrastructure for a system that employs model calculations incorporating observational "Big Data" to accurately predict localized torrential rain, tornados, typhoons etc. and that also monitors and projects impacts of environmental changes due to human activity, in order to contribute to environmental policy, disaster prevention and health measures.

9 social and scientific priority issues (2/3)

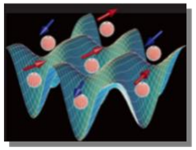
Category	Priority issues
Energy problem 	**⑤ Development of new fundamental technologies for high-efficiency energy creation, conversion/storage and use** Perform full-system simulations at the molecular level for complicated real-world complex systems to explain the entire process of high-efficiency energy creation, conversion/storage and use in coordination with experimentation, in order to develop new fundamental technologies to resolve energy-related problem.
	⑥ Accelerated Development of Innovative Clean Energy Systems Subject the complex physical phenomena that form the core of energy systems to first-principles analysis to predict their occurrence and explicate their comprehensive behavior for accelerating the practical application of innovative and clean energy systems that have ultra-high efficiency and low environmental impact.
Enhancement of industrial competitiveness 	**⑦ Creation of new functional devices and high-performance materials to support next-generation industries** Accelerate the development of electronics technologies, structural materials, functional chemical products etc. that have great international competitiveness, through coordination with large-scale massively parallel computing and the analysis of “Big Data” and data from measurement and experimentation, in order to create devices and 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 Conduct research and development for innovative design techniques, where the product concept is quantitatively assessed at the initial stage and optimization is performed. By implementing innovative manufacturing processes that reduce costs and by performing ultra-high speed integration simulations, both of which form the core of the research and development efforts, high value-added product development can be achieved.

9 social and scientific priority issues (3/3)

Category	Priority issues
Development of basic science 	**⑨ Elucidation of the fundamental laws and evolution of the universe** Realize precise calculations of the phenomena over wide range of scales from elementary particles to the universe. Combining with the data from large-scale experiments and observations, they play crucial roles to address the remaining problems in the history of the universe that extend across particle, nuclear and astro physics.

4 exploratory Challenges

Research organizations for 4 challenges have not been selected

Exploratory challenges
⑩ Frontiers of basic science: challenge to the limits  At the frontiers of basic science where researchers pursue the limits and extreme conditions, efforts will be made to resolve difficult problem and challenges that have been answered neither by experiments, observations nor even by individual achievements of computational science using the K Computer. Co-creation of new science and interdisciplinary collaboration using the "Post-K computer" is called for.
⑪ Construction of models for interaction among multiple socioeconomic To give policy and measures the agility to deal with various problems produced in our complex and rapidly changing modern society, research and development of systems for determination, analysis and prediction will be conducted, taking into account the effect of the mutual influence of individual elements of social activities such as transport and the economy.
⑫ Elucidation of the birth of exoplanets (Second Earths) and the environmental variations of planets in the solar system Through multidisciplinary approach under the collaboration of computational sciences (in the fields of astrophysics, geophysical/planetary science, meteorology, and molecular science), we achieve large-scale calculations, which can be directly confronted to observations and experiments, and explore the origin of terrestrial planets, the environment of the solar system, and interstellar molecular science.
⑬ Elucidation of how neural networks realize thinking and its application to artificial intelligence By integrating big data produced by innovative brain science technologies, large-scale multi-level models of the brain are constructed and through large-scale simulations using the "Post-K computer," the brain's mechanism of thinking by neural networks is reproduced and applied to artificial intelligence.  

19

Selected target apps from each area for “codesign”

	Target Application	
	Program	**Brief description**
①	GENESIS	MD for proteins
②	Genomon	Genome processing (Genome alignment)
③	GAMERA	Earthquake simulator (FEM in unstructured & structured grid)
④	NICAM+LETK	Weather prediction system using Big data (structured grid stencil & ensemble Kalman filter)
⑤	NTChem	molecular electronic (structure calculation)
⑥	FFB	Large Eddy Simulation (unstructured grid)
⑦	RSDFT	an ab-initio program (density functional theory)
⑧	Adventure	Computational Mechanics System for Large Scale Analysis and Design (unstructured grid)
⑨	CCS-QCD	Lattice QCD simulation (structured grid Monte Carlo)

Methodology and Tools for co-design

- Performance estimation tools provided by the vender
  - ① PA- information based performance estimation tool – estimate performance by using the PA information (derived by hardware counter to measure the time to issue inst., load, store, etc) obtained by FX100 (or K computer/FX10)
  - ② cycle-accurate CPU simulators and compiler for post K – it is limited to execution of kernel loops.

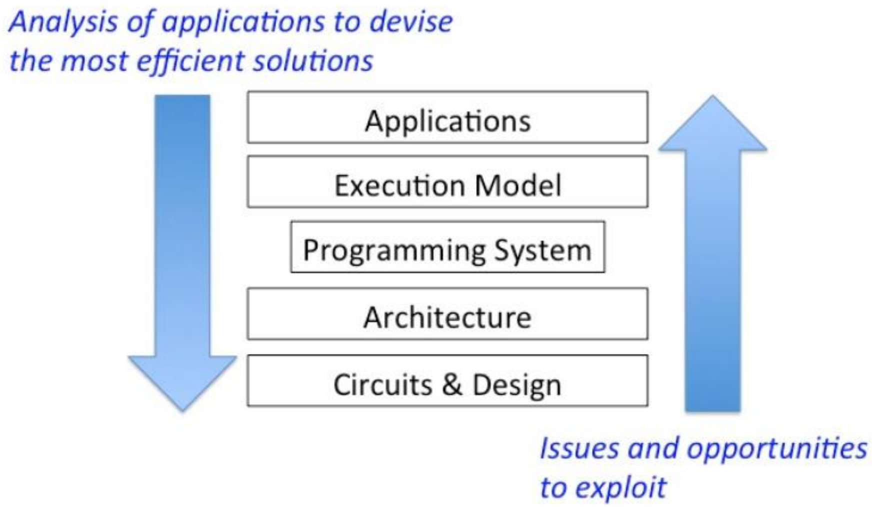
- Methodologies to estimate performance of post-K
  - (1) Building Analytical Performance model by roof-line model, etc …
  - (2) Performance estimation of whole application using ① PA- information based performance estimation tool
  - (3) Precious performance estimation by using ② cycle-accurate CPU simulators and compiler
  - Communication time by analytical model of network.

- Evaluate performance of each apps under combinations of various system parameter including number of core, SIMDs, configuration of cache and memory, …

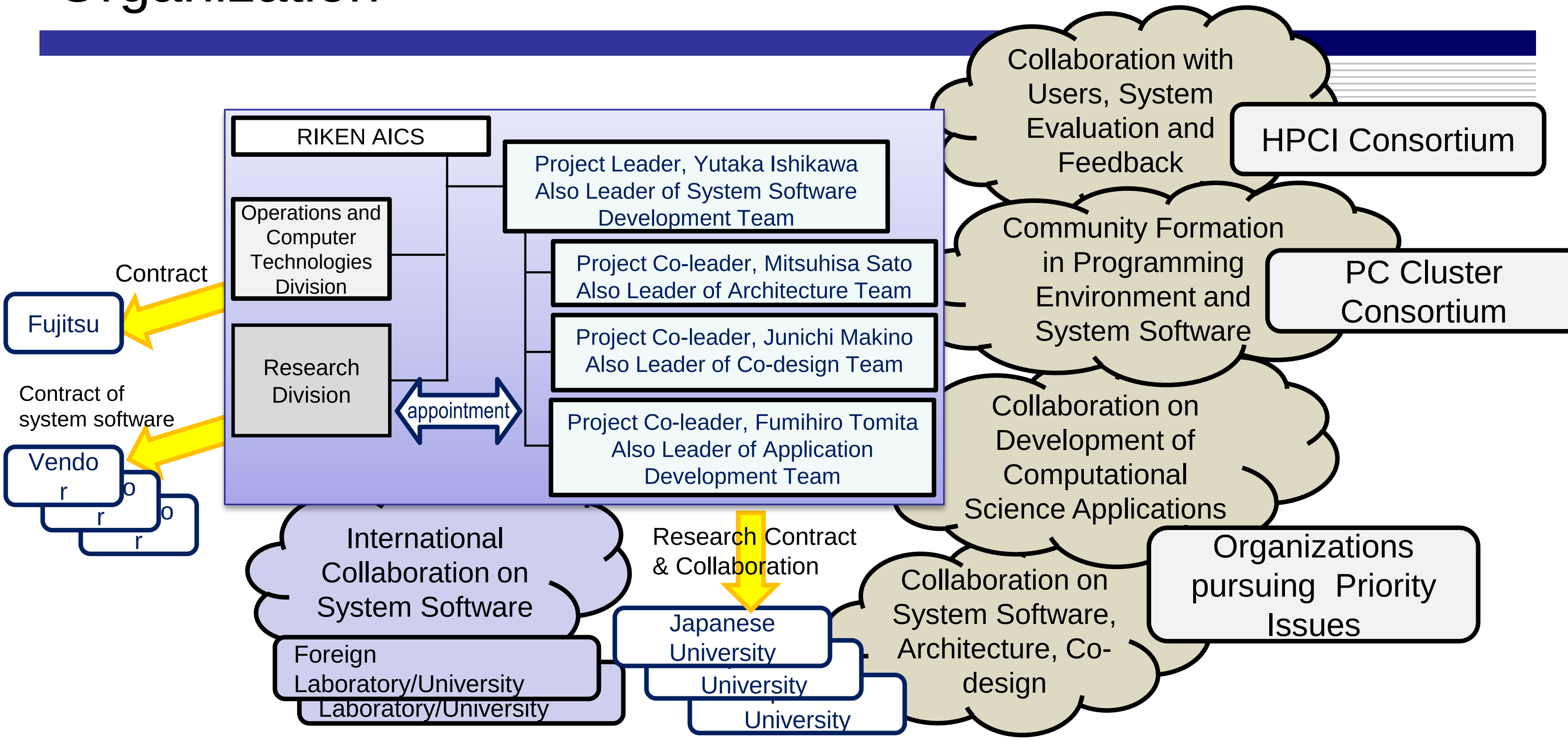
Topics of Co-design

- Trade-off against constraints of cost and total power/energy of the system
- Mechanism and middleware design for "power-shifting" by Power-knob to reduce the power consumption
  - Power-knob: mechanism to "switch" to on/off elements including memory-bandwidth and network-bandwidth, parts of CPUs
  - "Power-shift" between memory and CPUs – memory-intensive apps and CPU-hungry apps.

- Design of programming environment and math-libraries to exploit many-core architecture in each target apps.
- Design of DSL (Domain-Specific Language) for priority typical applications such as stencil apps and particle-based apps
- Design of System software such as file system for big-data apps



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

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)

Meeting in Dec/2014 at RIKEN AICS

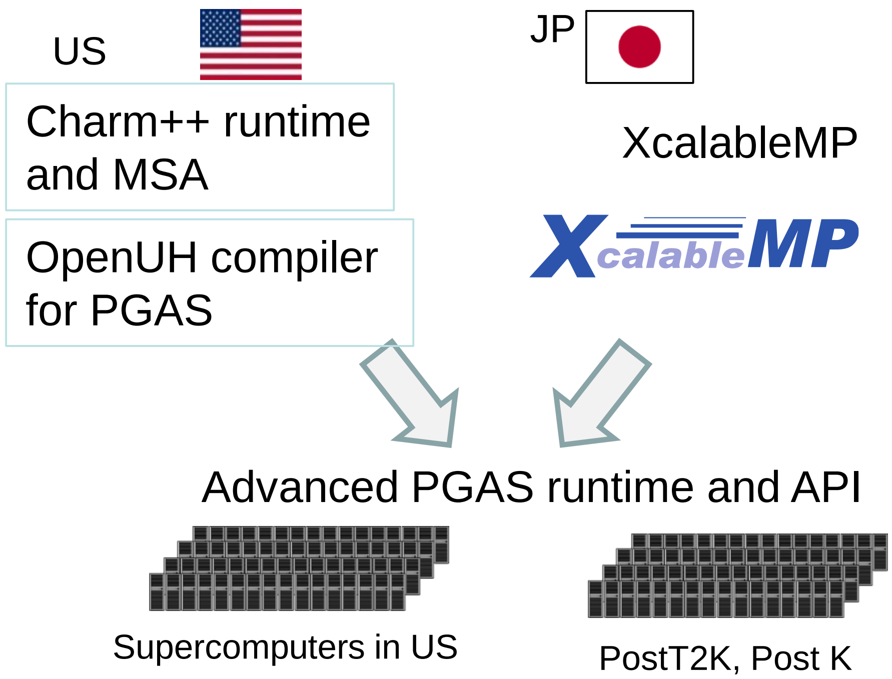
Collaborations with ORNL & RIKEN AICS: Co-design & programming models

- We are now starting discussions on collaborations in the context of US-JP DOE-MEXT collaborations. The two topics are under discussion:
  - Codesign methodology and tools
    - Use tools developed for codesign on both sides to analysis applications proposed in each side.
    - Applications performance studies of both sides.
    - Discussion on the future codesign methodologies
    - (We need to re-arrange some existing topics to avoid conflicts)
  - Advanced Programming models for manycore and accelerators
    - Research on advanced programming model for both of manycore and accelerators
    - To reduce the cost of application development cost by programming models covering both.

- Plan and schedule
  - We will have one-day meeting at RIKEN AICS in Japan on 22th Aug.
  - Research topics of our collaborations will be proposed at the meeting at Cluster 2015, in Sep.

Advanced PGAS runtime and API, programming models

- Coordinators
  - US: Peter Beckman, Argonne National Lab.
  - JP: Mitsuhisa Sasto, RIKEN AICS
- Leaders
  - US: Laxmikant (Sanjay) Kale, UIUC
  - Barabara Chapman, Univ. of Huston
  - JP: Mitsuhisa Sato, RIKEN AICS
- Description
  - Each side (RIKEN and Univ. Huston) is developing the programming languages and compilers based on “coarray” PGAS model. US partner, UIUC is working on advanced dynamic runtime based on their experience of Charm++.
  - We plan to explore how exascale system software may be exploited to enable the execution of PGAS programs at extreme scales, and consider extensions to this model for facilitating asynchronous movement of tasks and data.
  - As our deliverables, and APIs are derived.
- How to collaborate
  - Twice meetings per year
  - Student / young researchers exchange, sharing codes
  - Funding:
    - US: X-stack(XPRESS)?, ARGO?
    - JP: FLAGSHIP 2020
- Deliverables
  - Advanced runtime for scale PGAS model for exasclae
  - Pre-standardization of Application Programming Interface for PGAS language runtime



Plan & Status

- Technical Research Collaboration
  - Exploit the possibility to use Charm++ runtime as a part of XMP runtime
    - Use the idea of Multi-phase Share Arrays (MSA) in XMP.
    - Integrations of Charm++ object on to XMP C++ (under development).
  - Share experiences of PGAS model ("coarray") and runtime technologies with UH group.
    - Extension with Dynamic tasking in node
- Proposed agenda for pre-standarization
  - What is a good API, what level API is appropriate, to implement PGAS models
  - How to mix with MPI and PGAS operations in multithreaded execution.
  - Execution models of PGAS within shared memory node. How to support one-sided comm within a node.
  - Memory consistency model and semantics of one-sided comm of PGAS models
- Status
  - In March, JP team visit UH. We agreed with moving to design of dynamic tasking (in node) with PGAS models (CAF and XMP)
    - JP invite Post-doc from UH for discussion
    - Plan to send student from U. Tsukuba
  - In March, JP team visited UIUC for the discussion on these topics:
    - Discussion with future programming model
    - Charm++ on Argobots
    - How to integrate with Charm++ and XMP
  - JP team visited Arogbots team at ANL. We agreed the followings
    - Design and prototype implementation of OpenMP (Omni OpenMP) using Argobots for the hybrid of OpenMP and XcalableMP
    - Extends this work to design XcalableMP 2.0 for dynamic tasking

Joining JLESC

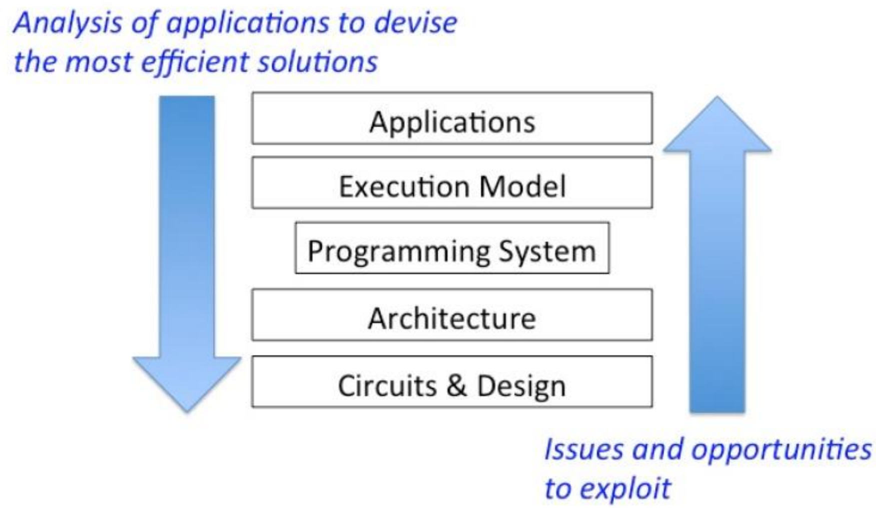
Joint Laboratory for Extreme Scale Computing

To initiate and facilitate international collaborations on research and state of the practice topics, related to computational and data focused simulation and analytics at scale. The JLESC will facilitate the production of original ideas, publications, discussion forums, research reports, products and open source software, aimed to address the most critical issues in advancing from petascale to extreme scale computing.

- Members
  - University of Illinois at Urbana-Champaign(NCSA), INRIA, Argonne National Laboratory, Barcelona Supercomputing Center and Jülich Supercomputing Centre
- RIKEN AICS Activity
  - MOU has been signed
  - RIKEN is going to propose collaboration areas

Summary

- FLAGSHIP 2020 project
  - To develop the next Japanese flagship computer system, post-K
  - The basic architecture design and target application performances will be decided by 2015 3Q
  - "Co-design" effort will be continued



- Schedule