ポスト京コンピューターの開発状況

石川　裕
理化学研究所計算科学研究機構

Yutaka Ishikawa @ RIKEN AICS

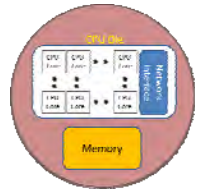
2015/07/01

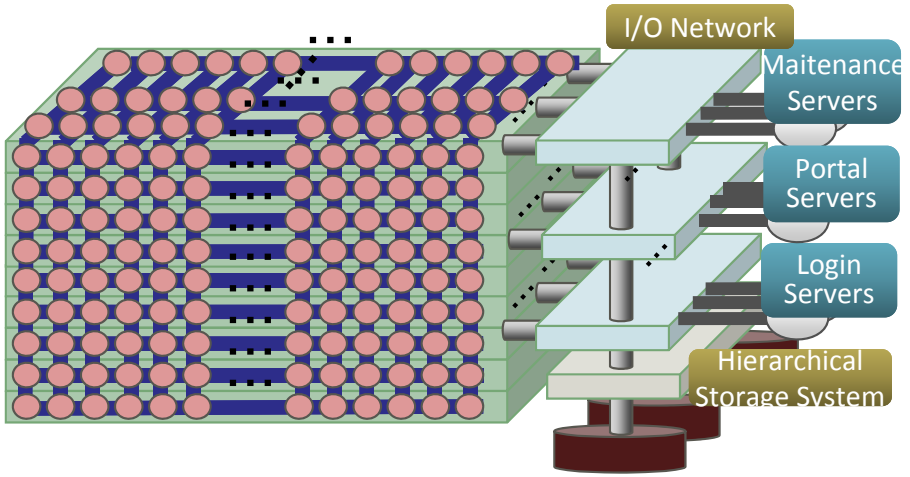
16:30 --- 17:05

バイオスーパーコンピューティング研究会

FLAGSHIP2020 Project

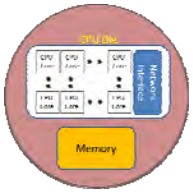
- Missions
  - Building the Japanese national flagship supercomputer, Post K, and
  - Developing wide range of HPC applications, running on Post K, in order to solve social and science issues in Japan
- Budget
  - 110 Billion JPY (about 0.91 Billion USD in case of 120 JPY/$)
  - including research, development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

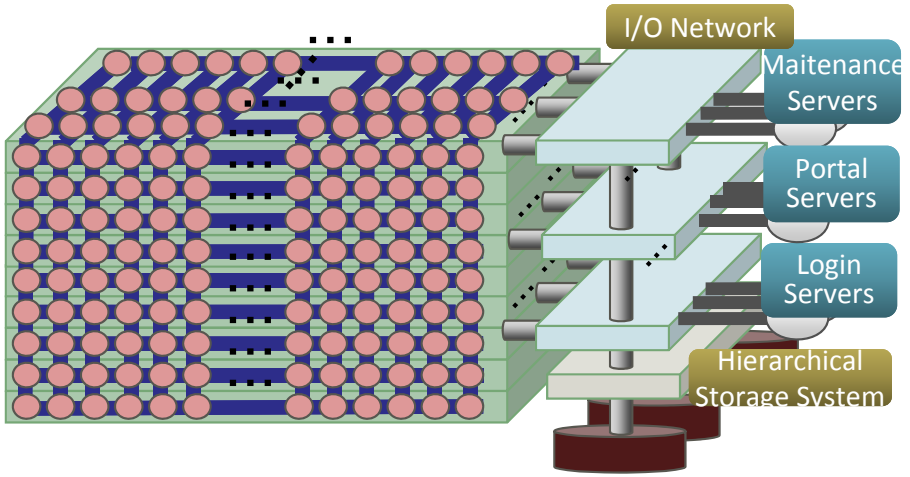


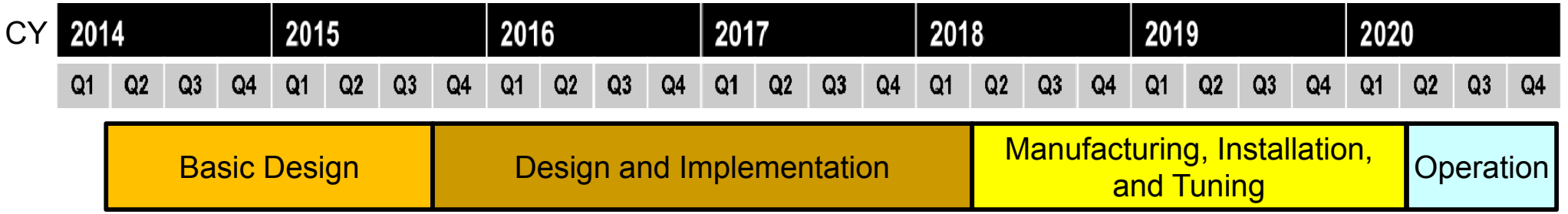


FLAGSHIP2020 Project

- 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
  - 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







ポスト京の設計方針

- Science-driven System
  - ポスト京運用後に成果が期待されるアプリケーションが必要とする性能要求
    - 大規模、精密、長時間発展といったcapability computingのニーズ
    - 複雑な現象を対象とした課題におけるensemble computingのニーズ
    - Big data computing、社会科学シミュレーションのニーズ

    → 広範なアプリ
  - Co-design
    - アプリケーション開発者と計算機システム開発者の協調によりアプリケーションおよびシステムを協調設計(co-design)していく
- Easy Migration
  - 京の後継機として京の資産が受け継げる
- TCO削減
  - 製造・運用・保守経費削減
- Upgradable System
  - ボード交換および機能拡張でポスト京の次世代CPUにアップグレード可能な設計

- **社会が欲するニーズに即応**

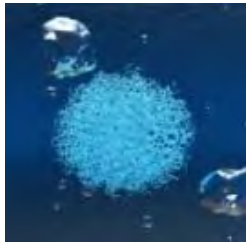
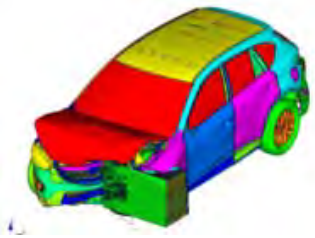
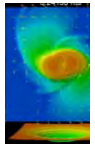
ビッグデータ、IoT、人工知能応用ニーズ
- ファイルI/O強化、ビッグデータ向けシステムソフトウェア整備
- 実時間処理

重点課題 (1/2)

①社会的・国家的見地から高い意義がある、
②世界を先導する成果の創出が期待できる、
③ポスト「京」の戦略的活用が期待できる課題を「重点課題」として選定。

カテゴリ	重点課題
健康長寿社会の実現 	**① 生体分子システムの機能制御による革新的創薬基盤の構築** 超高速分子シミュレーションを実現し、副作用因子を含む多数の生体分子について、機能阻害ばかりでなく、機能制御までをも達成することにより、有効性が高く、さらに安全な創薬を実現する。
	② 個別化・予防医療を支援する統合計算生命科学 健康・医療ビッグデータの大規模解析とそれらを用いて得られる最適なモデルによる生体シミュレーション（心臓、脳神経など）により、個々人に適した医療、健康寿命を延ばす予防をめざした医療を支援する。
防災・環境問題 	**③ 地震・津波による複合災害の統合的予測システムの構築** 内閣府・自治体等の防災システムに実装しうる、大規模計算を使った地震・津波による災害・被害シミュレーションの解析手法を開発し、過去の被害経験からでは予測困難な複合災害のための統合的予測手法を構築する。
	④ 観測ビッグデータを活用した気象と地球環境の予測の高度化 観測ビッグデータを組み入れたモデル計算で、局地的豪雨や竜巻、台風等を高精度に予測し、また、人間活動による環境変化の影響を予測し監視するシステムの基盤を構築する。環境政策や防災、健康対策へ貢献する。

重点課題 (2/2)

カテゴリ	重点課題
エネルギー問題 	**⑤ エネルギーの高効率な創出、変換・貯蔵、利用の新規基盤技術の開発** 複雑な現実複合系の分子レベルでの全系シミュレーションを行い、高効率なエネルギーの創出、変換・貯蔵、利用の全過程を実験と連携して解明し、エネルギー問題解決のための新規基盤技術を開発する。
	⑥ 革新的クリーンエネルギーシステムの実用化 エネルギーシステムの中核をなす複雑な物理現象を第一原理解析により、詳細に予測・解明し、超高効率・低環境負荷な革新的クリーンエネルギーシステムの実用化を大幅に加速する。
産業競争力の強化 	**⑦ 次世代の産業を支える新機能デバイス・高性能材料の創成** 国際競争力の高いエレクトロニクス技術や構造材料、機能化学品等の開発を、大規模超並列計算と計測・実験からのデータやビッグデータ解析との連携によって加速し、次世代の産業を支えるデバイス・材料を創成する。
	⑧ 近未来型ものづくりを先導する革新的設計・製造プロセスの開発 製品コンセプトを初期段階で定量評価し最適化する革新的設計手法、コストを最小化する革新的製造プロセス、およびそれらの核となる超高速統合シミュレーションを研究開発し、付加価値の高いものづくりを実現する。
基礎科学の発展 	**⑨ 宇宙の基本法則と進化の解明** 素粒子から宇宙までの異なるスケールにまたがる現象の超精密計算を実現し、大型実験・観測のデータと組み合わせて、多くの謎が残されている素粒子・原子核・宇宙物理学全体にわたる物質創成史を解明する。

萌芽的課題

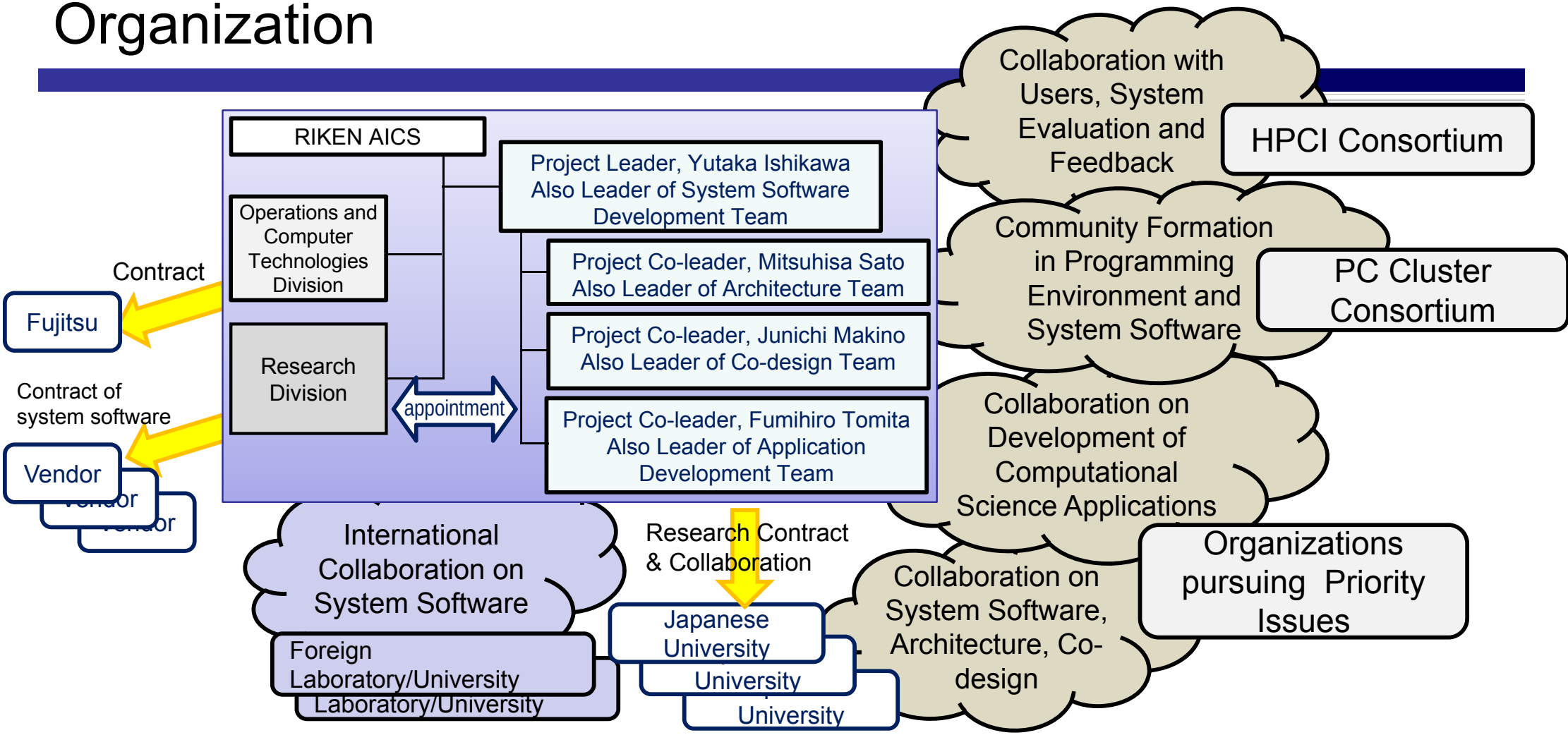
ポスト「京」で新たに取り組むチャレンジングな課題として、今後、**調査研究を通じて実現化を検討**する。調査研究終了後に、ポスト「京」における研究開発実施について決定する。

萌芽的課題	
将来性を考慮し、今後、実現化を検討する課題	**⑩ 基礎科学のフロンティア　－　極限への挑戦** 極限を探究する基礎科学のフロンティアで、実験・観測や「京」を用いた個別計算科学の成果にもかかわらず答の出ていない難問に、ポスト「京」のみがなし得る新しい科学の共創と学際連携で挑み、解決を目指す。
	⑪ 複数の社会経済現象の相互作用のモデル構築とその応用研究 複雑且つ急速に変化する現代社会で生じる様々な問題に政策・施策が俊敏に対応するために、交通や経済など社会活動の個々の要素が互いに影響し合う効果を取り入れて把握・分析・予測するシステムを研究開発する。
	⑫ 太陽系外惑星（第二の地球）の誕生と太陽系内惑星環境変動の解明 宇宙、地球・惑星、気象、分子科学分野の計算科学と宇宙観測・実験が連携する学際的な取り組みにより、観測・実験と直接比較可能な大規模計算を実現し、地球型惑星の起源、太陽系環境、星間分子科学を探究する。
	⑬ 思考を実現する神経回路機構の解明と人工知能への応用 革新技術による脳科学の大量のデータを融合した大規模多階層モデルを構築し，ポスト「京」での大規模シミュレーションにより思考を実現する脳の大規模神経回路を再現し，人工知能への応用をはかる。

重点実施機関

カテゴリ	重点課題名	選定実施機関
健康長寿社会の実現	①生体分子システムの機能制御による革新的創薬基盤の構築	理化学研究所生命システム研究センター （課題責任者：奥野 恭史・客員主管研究員） 他5機関
	②個別化・予防医療を支援する統合計算生命科学	東京大学医科学研究所 （課題責任者：宮野 悟・教授） 他5機関
防災・環境問題	③地震・津波による複合災害の統合的予測システムの構築	東京大学地震研究所 （課題責任者：堀 宗朗・教授） 他4機関
	④観測ビッグデータを活用した気象と地球環境の予測の高度化	海洋研究開発機構地球情報基盤センター （課題責任者：高橋 桂子・センター長） 他3機関
エネルギー問題	⑤エネルギーの高効率な創出、変換・貯蔵、利用の新規基盤技術の開発	自然科学研究機構分子科学研究所 （課題責任者：岡崎 進・教授） 他8機関
	⑥革新的クリーンエネルギーシステムの実用化	東京大学大学院工学系研究科 （課題責任者：吉村 忍・教授） 他11機関
産業競争力の強化	⑦次世代の産業を支える新機能デバイス・高性能材料の創成	東京大学物性研究所 （課題責任者：常行 真司・教授） 他8機関
	⑧近未来型ものづくりを先導する革新的設計・製造プロセスの開発	東京大学生産技術研究所 （課題責任者：加藤 千幸・教授） 他6機関
基礎科学の発展	⑨宇宙の基本法則と進化の解明	筑波大学計算科学研究センター （課題責任者：青木 慎也・客員教授） 他7機関

Organization



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

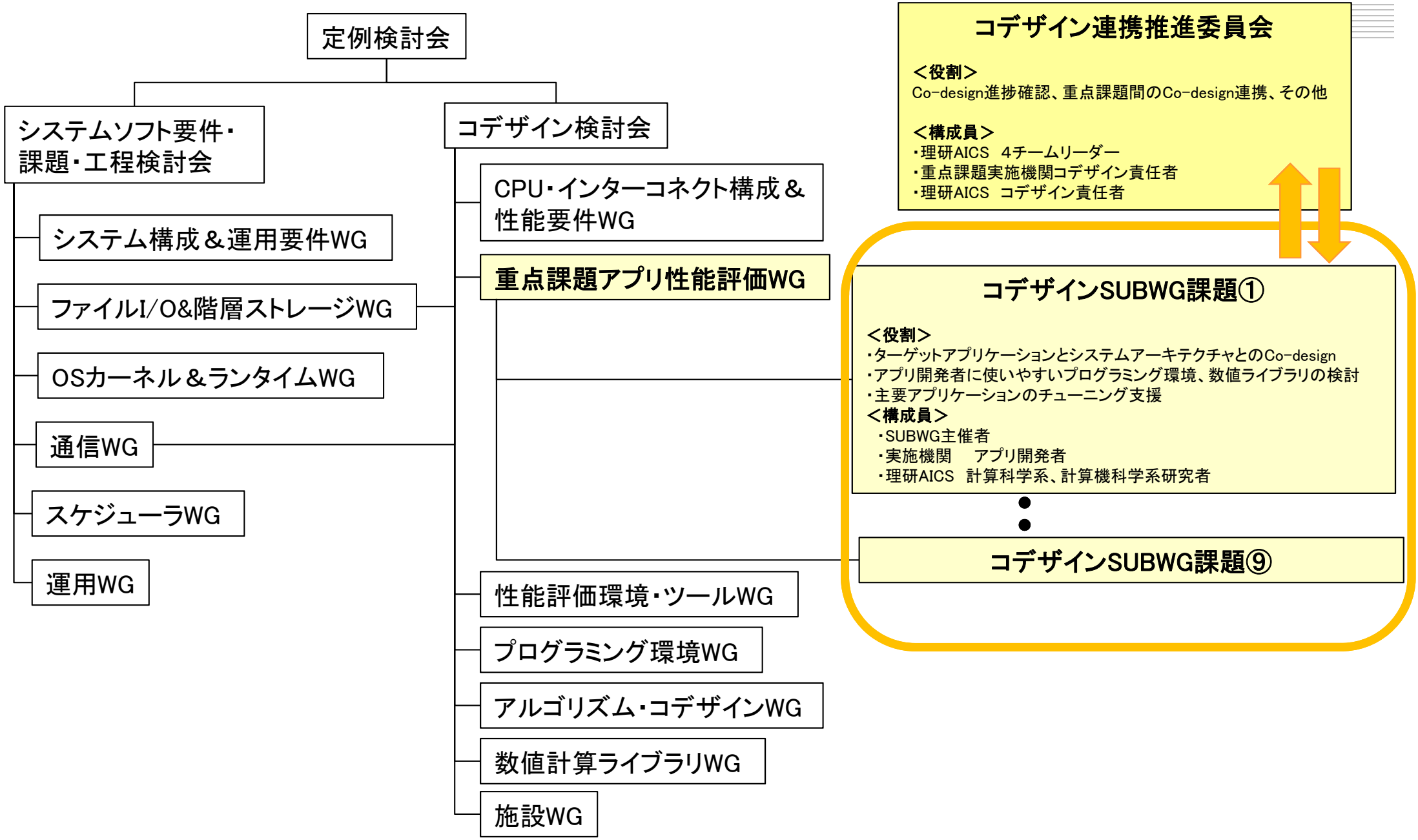
Japanese Universities
(Pre-)Standardization of API/SPI, Benchmarks, etc.
- Power Control API, FT API, etc.
- Evaluation of Architecture & Co-design

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

Co-design推進体制



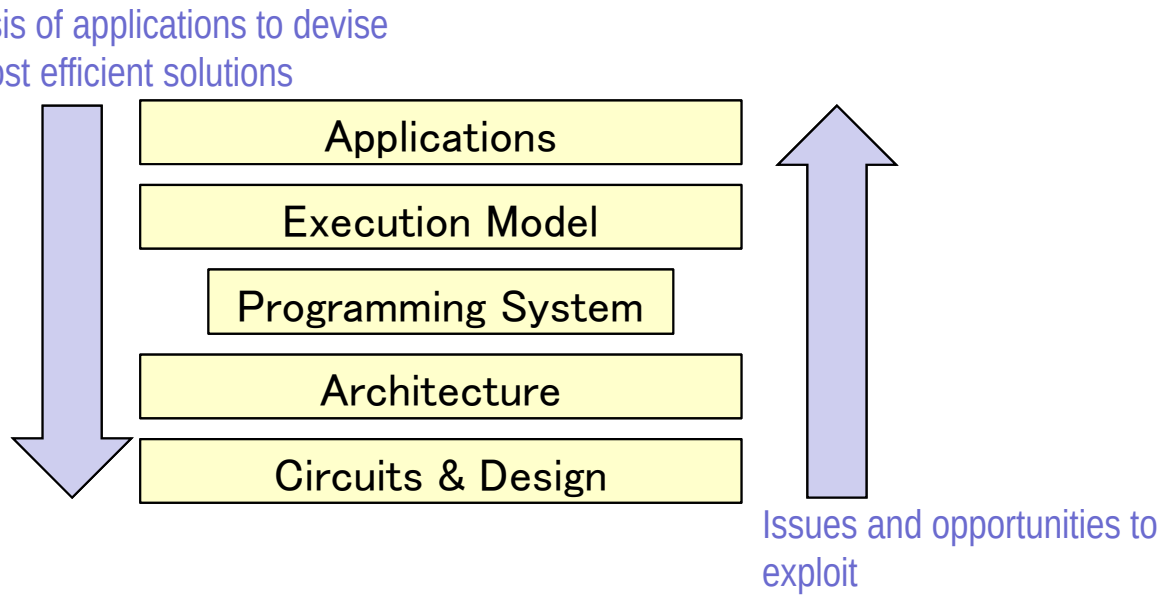
コデザイン手法

基本設計時

- ハードあるいはソフトの設計段階を提示し、それに間に合うフィードバックが得られるように評価の優先順位・スケジュールについて合意した上で、コデザインを進めた
- 重点アプリ性能評価から
  - 性能を制限する要因の分類、対応の検討。対応した場合の性能予測

今後

- アーキテクチャの特徴（メニーコアなど）を生かしたアプリ、プログラミングモデル、アルゴリズムの開発
- 電力を考慮したアプリ開発



出展：Richard F. BARRETT, et.al. “On the Role of Co-design in High Performance Computing”, *Transition of HPC Towards Exascale Computing, IOS press, 2013.*

Post K Computer

Node and Storage Architecture

- Manycore architecture

  From Fujitsu's view point, post K will be a successor (successor's successor ?) of FX100

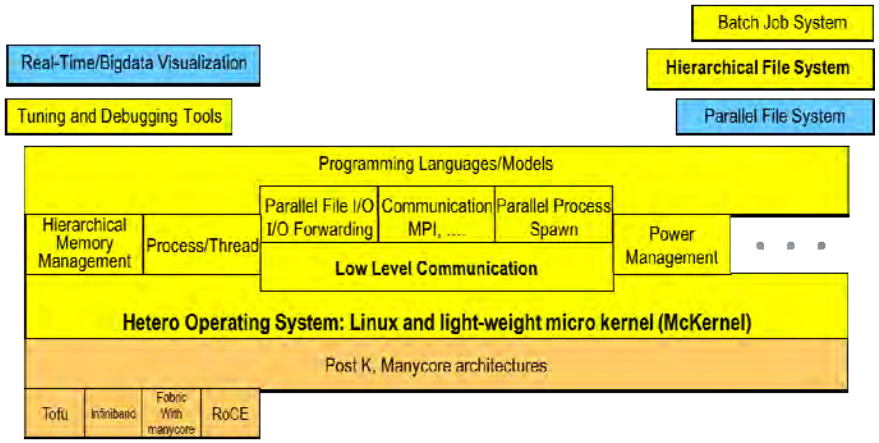
- 3 level hierarchical storage system
  - Silicon Disk
  - Magnetic Disk
  - Storage for archive

FX100 Announced in 2014, and now delivery

Architecture	SPARCV9 + HPC-ACE2
No. of cores	32 compute cores + 2 assistant cores
Peak performance	1+ TF
Memory	32 GB (HMC) read: 240 GB/s, write: 240 GB/s
Interconnect	Tofu2: 12.5 GB/s x 2 (bidirection) x 10 link

http://www.fujitsu.com/global/Images/primehpc-fx100-datasheet-en.pdf

System Software Architecture



✓The system should be designed to maximize the performance of applications in 9 social and scientific priority issues.
"Co-design" is a keyword!

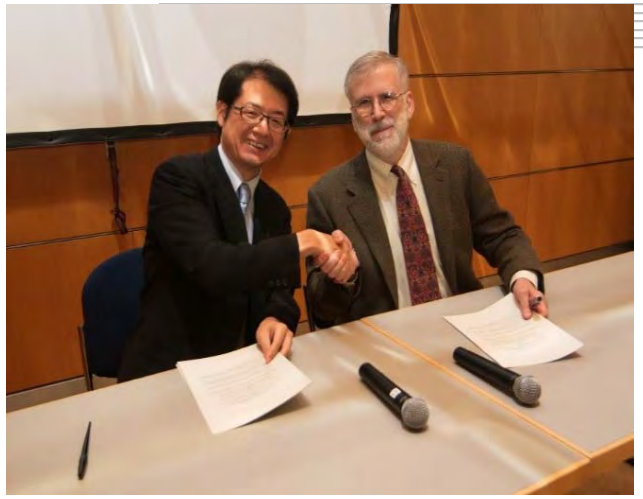
従来マシン

	京	FX10	FX100	ポスト京
導入年	2010-2011	2012	2015	2018 - 2019
微細加工技術	45 nm	40 nm	20 nm	10 nm
Peak FLOPS /node	128	236.48	1,100	
SIMD演算	128 bit (2DP) SIMD x 2 x FMA, 8 flop/cycle	128 bit (2DP) SIMD x 2 x FMA, 8 flop/cycle	256 bit (4DP) SIMD x 2 x FMA, 16 flop/cycle	
Core数	8	16	32	
メモリ階層	L1: 32KB/core, L2: 6 MB 16 GB	L1: 32KB/core, L2: 12 MB 32 GB	L1:32KB/core, L2:24MB 32GB	
メモリバンド幅、技術	64GB/s, DDR3	85GB/s, DDR3	480GB/s, HMC	
高IPC化 (対 京&FX10)	-	-	整数パイプ増 Out-of-order資源増	
Interconnect	5 GB/sec x 10, Tofu	5 GB/sec x 10, Tofu	12.5GB/sec x 10, Tofu2	
Linpack Node level GF/W	0.8	0.9	3	

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

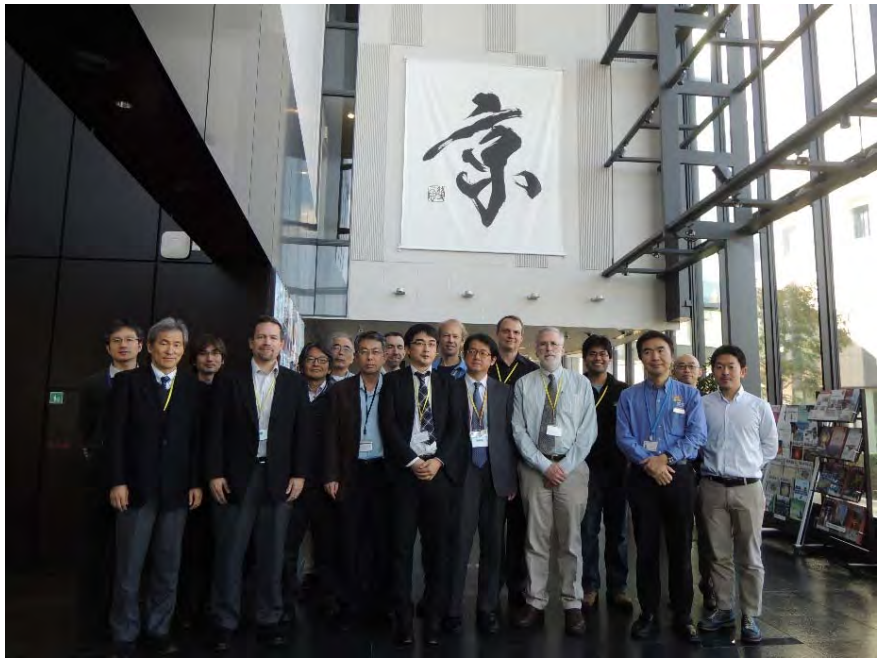
- Joint Activities
  - Pre-standardization interface coordination
  - Collection and publication of open data
  - Collaborative development of open source software
  - Evaluation and analysis of benchmarks and architectures
  - Standardization of mature technologies

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

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, 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

ビッグデータアプリ対応ファイルI/O強化

- ステージング廃止
  - ユーザがステージングするファイル名を記述：プレステージングを可能にし、計算ノード資源の有効活用が目的
  - 不要なファイルまでステージングする場合があり、高システム負荷が発生
    - ユーザはできる限り簡単な記述(e.g. XXフォルダ配下の全てのファイルというような指定)をしたときに、誤って大量の不要ファイルがステージングされたことが原因

- ファイルI/O性能強化
  - 階層ストレージを導入
    - 第一階層
      - 半導体ディスクを使用し、京のローカルファイルシステムに比べて最大100倍性能を目標にする
    - 第2階層
      - ハードディスクを用いたグローバルファイルシステム
    - 第3階層
      - テープアーカイブによる大容量ストレージ
  - 性能および容量は詳細設計で検討

- ファイルステージングとは？
  - ステージイン
    グローバルファイルシステム上にあるジョブが必要とするファイルをジョブ実行前に、高速アクセス可能なローカルファイルシステムにコピーする
  - ステージアウトはその逆