情報基盤センターの
スパコン

東京大学情報基盤センター

- 人間の全ての行動において「情報」と無縁なものは無い
  - 学問，研究もその例外では無い
- 東京大学における様々な「情報」に関わる活動を支援する
  - 学術情報メディア
  - 図書館電子化，学術情報
  - ネットワーク
  - スーパーコンピューティング
- 大量で多様な情報：コンピュータ＋ネットワーク

スーパーコンピューティング部門（1/2）

http://www.cc.u-tokyo.ac.jp/

- スーパーコンピュータの運用，利用支援
  - 3つのシステム
    - Hitachi SR16000（Yayoi）
    - Hitachi HA8000（T2K東大）
    - Fujitsu PRIMEHPC FX10（Oakleaf-FX）
  - 合計約2,000人のユーザー（学外が半分）
    - 大学（研究，教育），研究機関，企業

東大センターのスパコン（～2011.09E）

1システム～6年，3年周期でリプレース

HITACHI SR11000 model J2

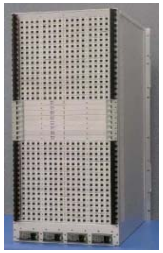
Total Peak performance	: 18.8 TFLOPS
Total number of nodes	: 128
Total memory	: 16384 GB
Peak performance per node	: 147.2 GFLOPS
Main memory per node	: 128 GB
Disk capacity	: 94.2 TB

IBM POWER5+ 2.3GHz

T2K（東大）（HA8000クラスタシステム）

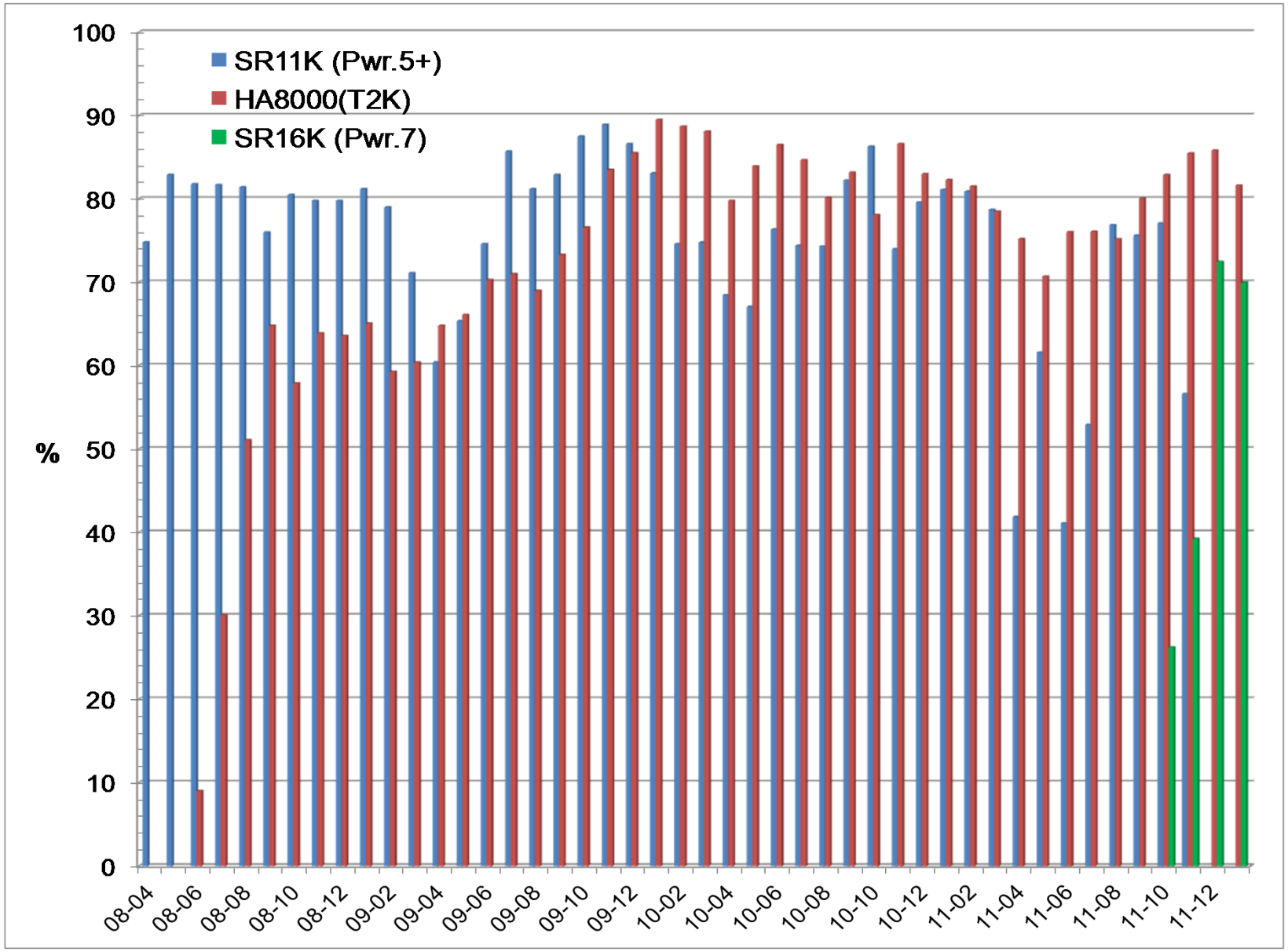
Total Peak performance	: 140 TFLOPS
Total number of nodes	: 952
Total memory	: 32000 GB
Peak performance per node	: 147.2 GFLOPS
Main memory per node	: 32 GB, 128 GB
Disk capacity	: 1 PB

AMD Quad Core Opteron 2.3GHz

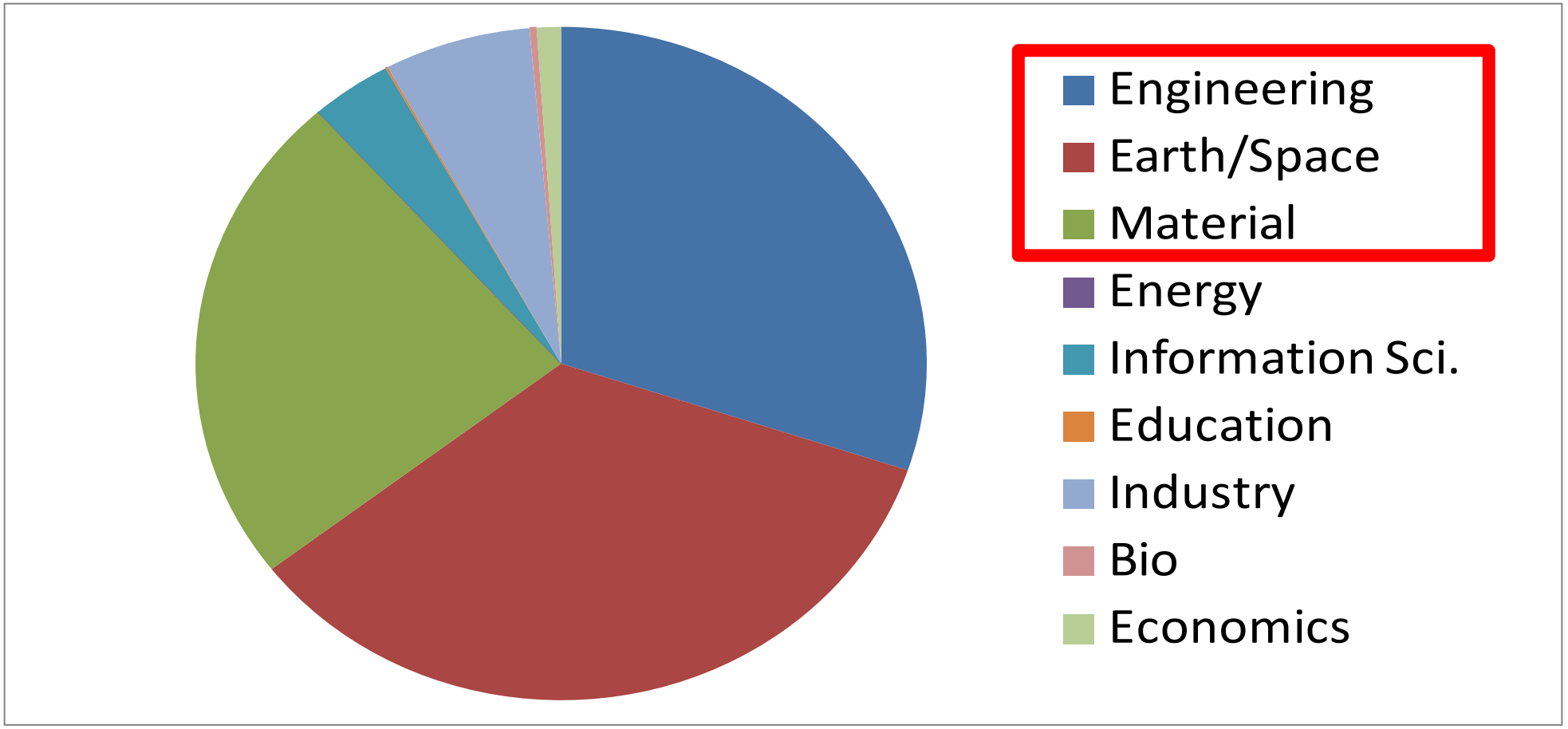


東大センターのスパコン（~2012.01E）

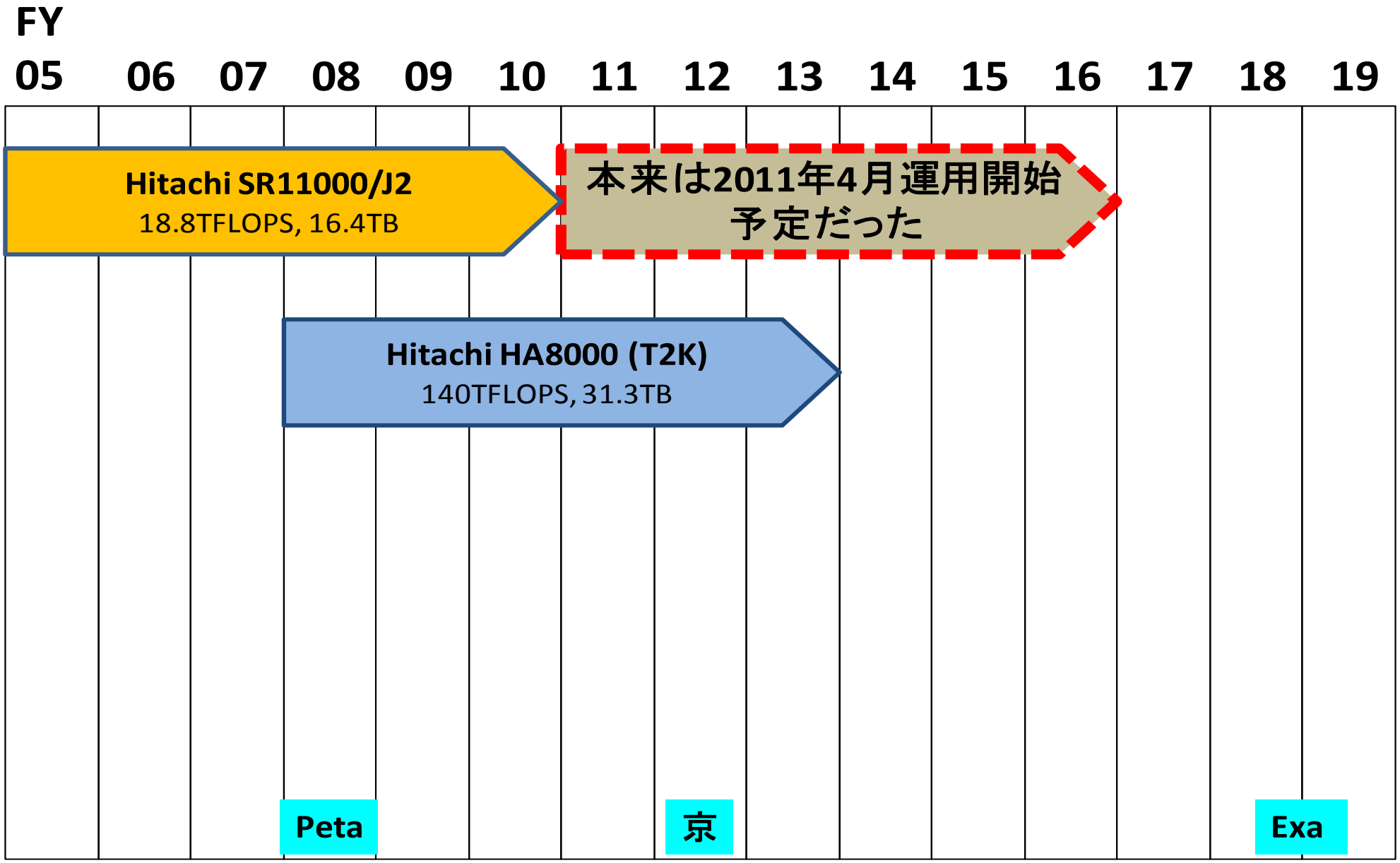
利用者：SR11K-約490名，SR16K-約360名，HA8000-約1,100名



利用ノード時間積による利用分野
T2K：FY.2011（2012.1月末時点）
専用キュー＋教育＋企業



東大情報基盤センターのスパコン



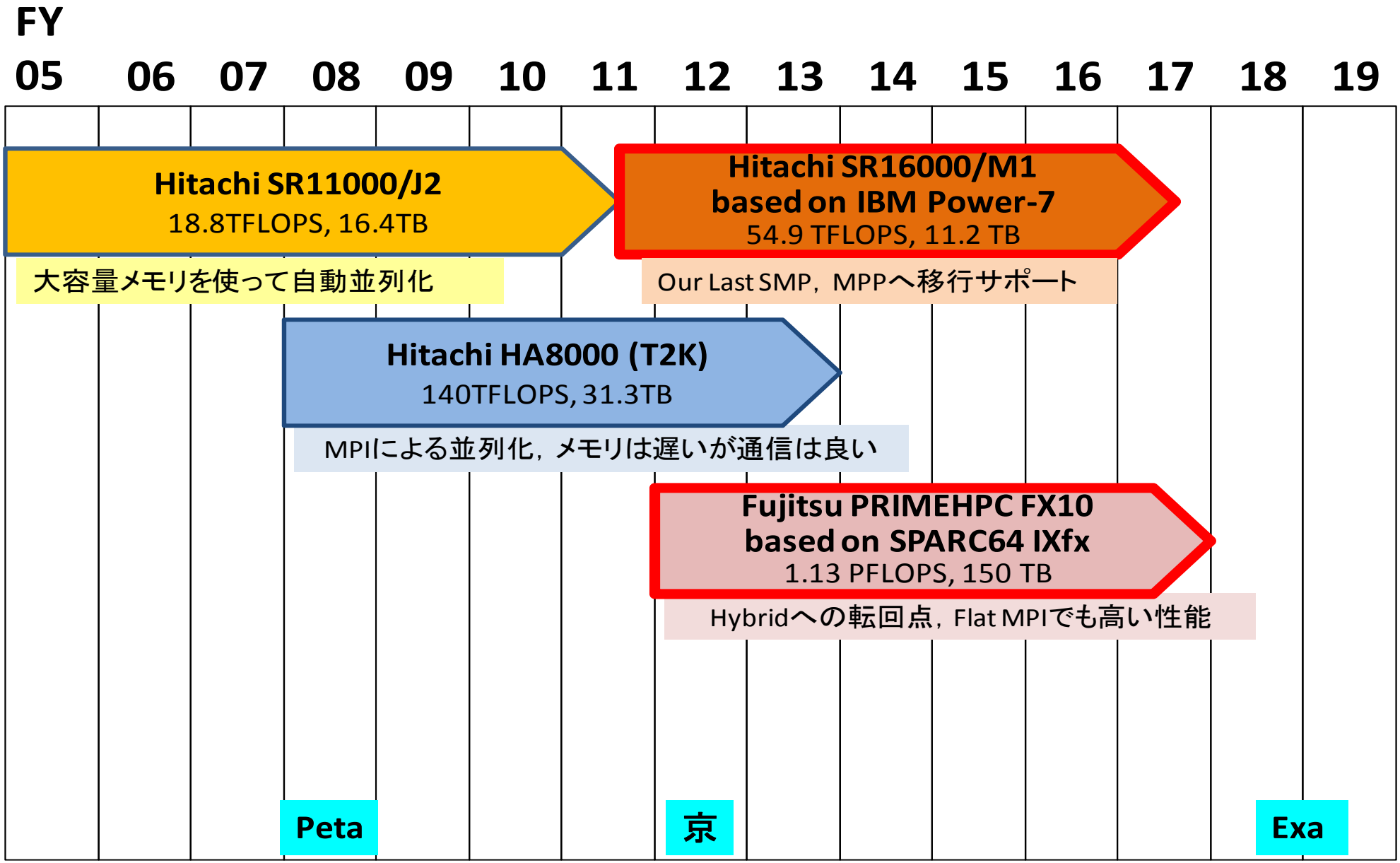
新システム導入の経緯

- 2009年4月頃から次期システムに関する検討を開始
- 2システムの導入
  - SR後継機（Power7）
  - PFLOPS級MPP，総メモリバンド幅400TB/sec以上
    - アクセラレータ，コプロセッサ無し
    - 計算性能～消費電力のバランス，コンパクト性
    - ファイルシステム性能
    - オープンソースライブラリ・アプリケーション
- 柏地区への移転
  - 電力，設置面積
- 東日本大震災
  - 調達やりなおし
    - 消費電力に配慮（空調込み2.0MW以下）
    - ピークカットを考慮し，柔軟な運用が可能となるような要求を付加

新システム

- **SMP：Hitachi SR16000/M1**
  - SR16000システム（SMP）（Yayoi）
  - ピーク性能 54.9 TFLOPS
  - 56計算ノード
    - IBM POWER 7, 32 cores/node, 200 GB/node
  - 2011年10月3日より試行運用，11月25日より本運用開始
  - 大容量メモリノードを有するタイプのシステム（SMPと呼んでいる）の導入はこれで最後（データサーバー等除く）
    - 利用者は6年以内に並列化を進め，MPP等へ移行する
      - センターも講習会，個別相談などできる限りのサポートをする
- **MPP: Fujitsu PRIMEHPC FX10**
  - FX10スーパーコンピュータシステム（Oakleaf-FX）
  - ピーク性能　1.13 PFLOPS
  - 4,800計算ノード
    - SPARC64 IXfx，16 cores/node, 32GB/node

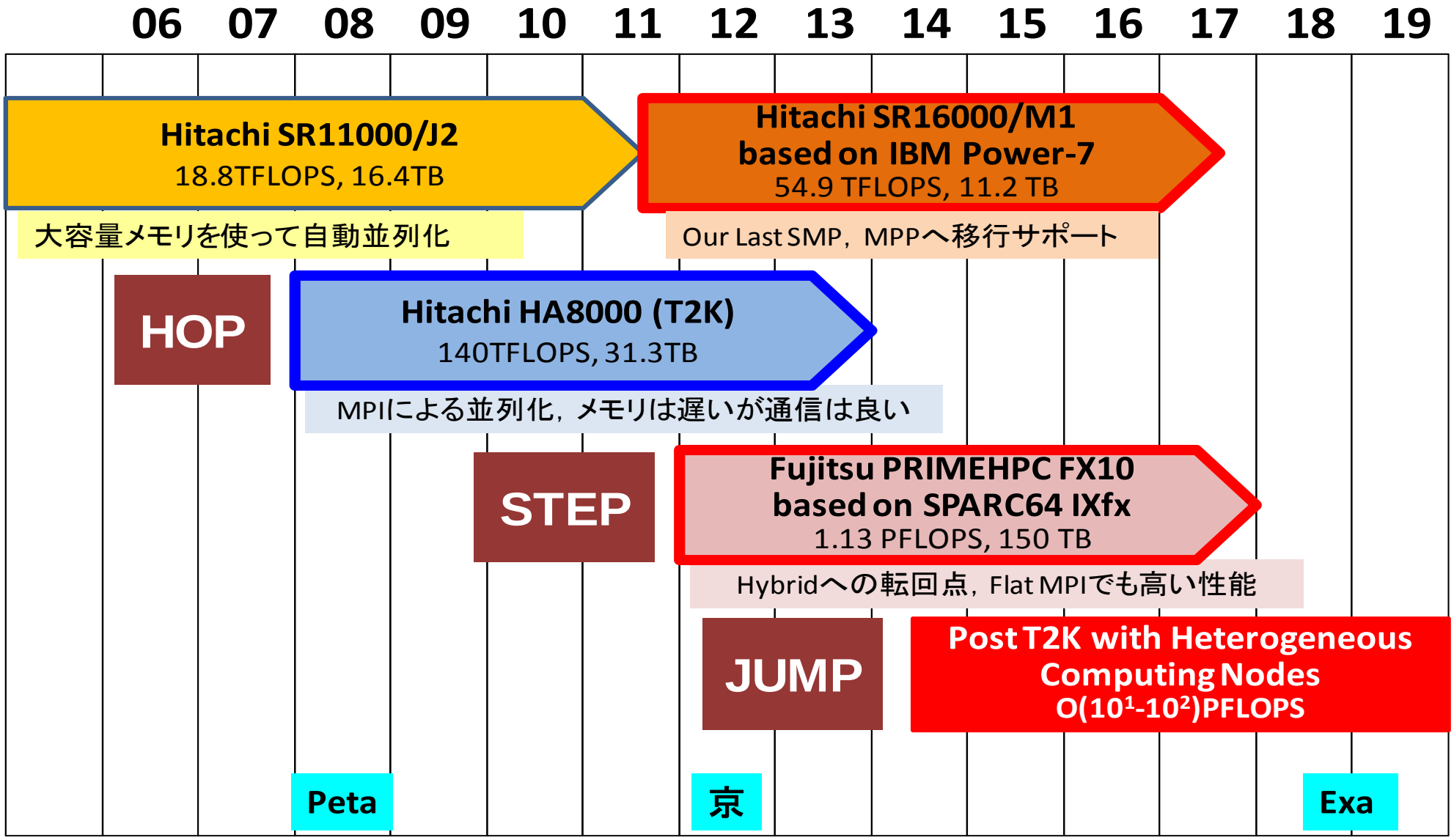
東大情報基盤センターのスパコン



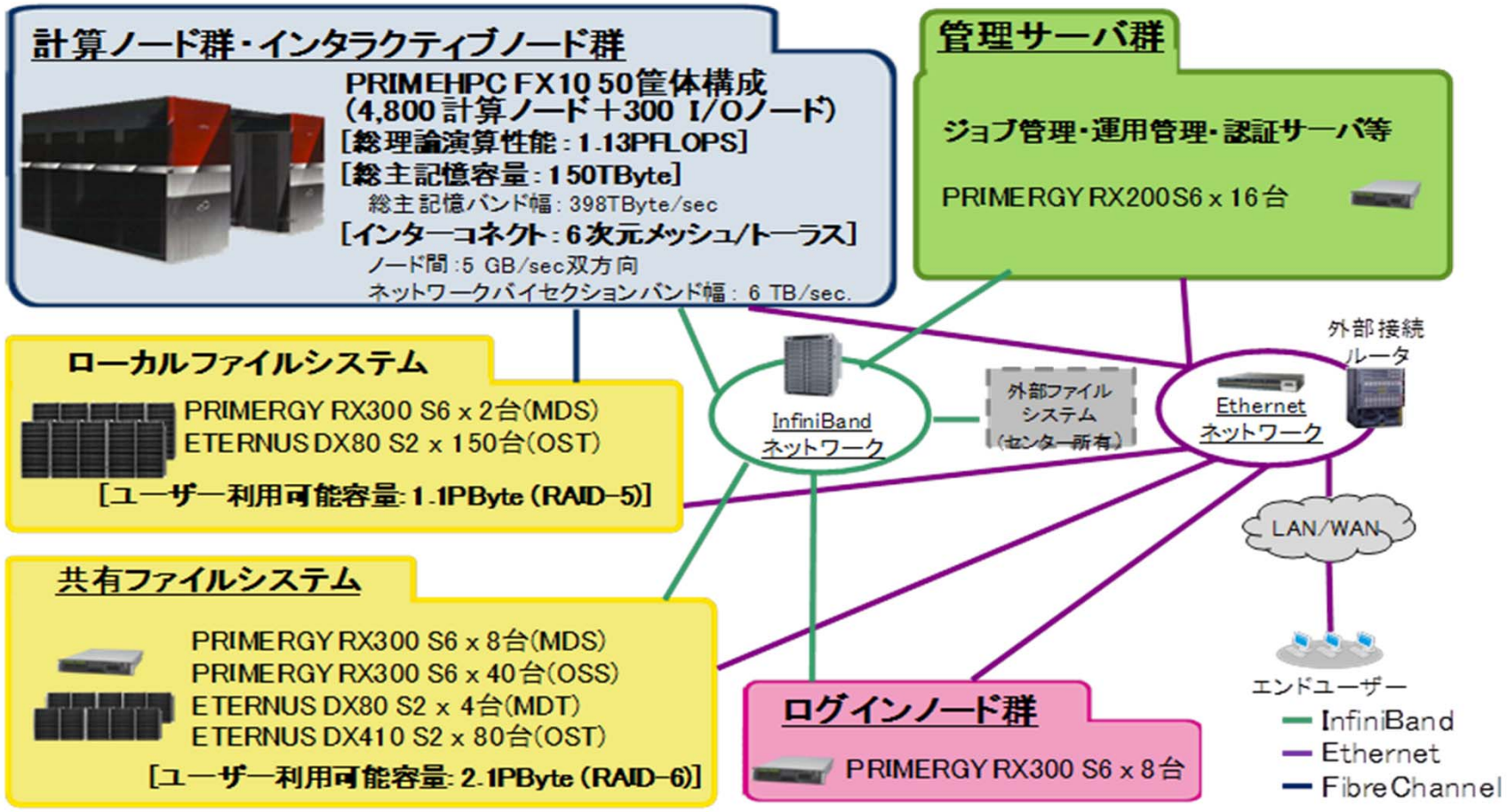
新MPPの位置づけ：三段跳びの「Step」

- Hop
  - HA8000（T2K），Homogeneous Compute Nodes
  - O(10^{-1}) PFLOPS
  - Flat MPI
- Step
  - PRIMEHPC FX10, Homogeneous
  - O(10^{0}) PFLOPS
  - MPI + OpenMP，但しFlat MPIも充分速くなければ使えない
- Jump
  - Post T2K, Heterogeneous
    - 省電力，メモリバンド幅：Heterogeneousな計算ノード
  - O(10^{1}-10^{2}) PFLOPS
  - MPI + X（OpenMP, CUDA, OpenCL … OpenACC）
- その先にExaがあるはず

東大情報基盤センターのスパコン



FX10 System (Oakleaf-FX)



- Aggregate memory bandwidth: 398 TB/sec.
- Local file system for staging with 1.1 PB of capacity and 131 GB/sec of aggregate I/O performance (for staging)
- Shared file system for storing data with 2.1 PB and 136 GB/sec.
- External file system: 3.6 PB

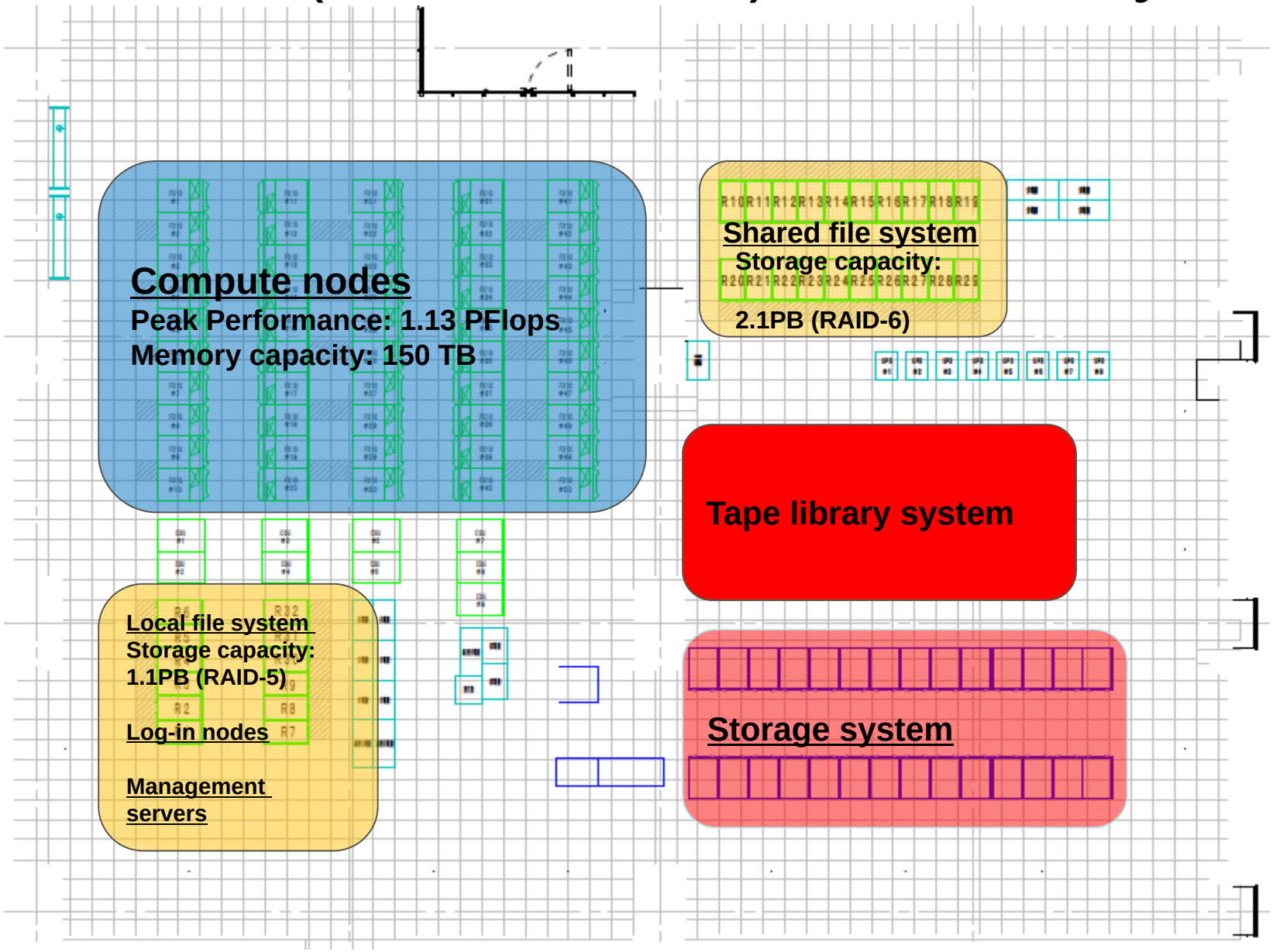
39th TOP 500 List (June 2012) (1/2)

	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	LLNL United States	**Sequoia** BlueGene/Q, 2011 IBM	1572864	16324.75	20132.66	7890.0
2	RIKEN AICS Japan	**K computer**, SPARC64 VIIIfx , 2011 Fujitsu	705024	10510.00	11280.38	12659.9
3	Argonne United States	**Mira** BlueGene/Q, 2012 IBM	786432	8162.38	10066.33	3945.0
4	Leibniz Rechenzentrum Germany	**SuperMUC** iDataPlex/Xeon E5-2680 2012 IBM	147456	2897.00	3185.05	3422.7
5	NSCS in Tianjin China	**Tianhe-1A** Heterogeneous Node 2010 NUDT	186368	2566.00	4701.00	4040.0
6	ORNL United States	**Jaguar**, Cray XK6 （一部 Heterogeneous）, 2009 Cray Inc.	298592	1941.00	2627.61	5142.0
7	CINECA Italy	**Fermi** BlueGene/Q, 2012 IBM	163840	1725.49	2097.15	821.9
8	Forschungszentrum Juelich (FZJ) Germany	**JuQUEEN** BlueGene/Q, 2012 IBM	131072	1380.39	1677.72	657.5
9	CEA/TGCC-GENCI France	**Curie thin nodes** Xeon E5-2680, 2012 Bull	77184	1359.00	1667.17	2251.0
10	NSCS in Shenzhen China	**Nebulae**, Heterogeneous Node 2010 Dawning	120640	1271.00	2984.30	2580.0

39th TOP 500 List (June 2012) (2/2)

	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
11	NASA Ames United States	**Pleiades** SGI Altix ICE, 2011 SGI	125980	1243.00	1731.84	3987.0
12	IFRC, EU-Japan Japan	**Helios** Xeon E5-2680, 2011 Bull	70560	1237.00	1524.10	2200.0
13	Daresbury Lab. United Kingdom	**Blue Joule** BlueGene/Q , 2012 IBM	114688	1207.84	1468.01	575.3
14	GSIC – Tokyo Tech Japan	**TSUBAME 2.0** Heterogeneous Node 2010 NEC/HP	73278	1192.00	2287.63	1398.6
15	LANL/SNL United States	**Cielo** Cray XE6, 2011 Cray Inc.	142272	1110.00	1365.81	3980.0
16	LBNL United States	**Hopper** Cray XE6, 2010 Cray Inc.	153408	1054.00	1288.63	2910.0
17	CEA France	**Tera-100** Xeon X7560, 2010 Bull	138368	1050.00	1254.55	4590.0
18	ITC/U. Tokyo Japan	**Oakleaf-FX**, SPARC64 IXfx, 2012 Fujitsu	76800	1043.00	1135.41	1176.8
19	LANL United States	**Roadrunner** Heterogeneous Node 2009 IBM	122400	1042.00	1375.78	2345.0
20	U. Edinburgh United Kingdom	**DiRAC** BlueGene/Q, 2012 IBM	98304	1035.30	1258.29	493.1

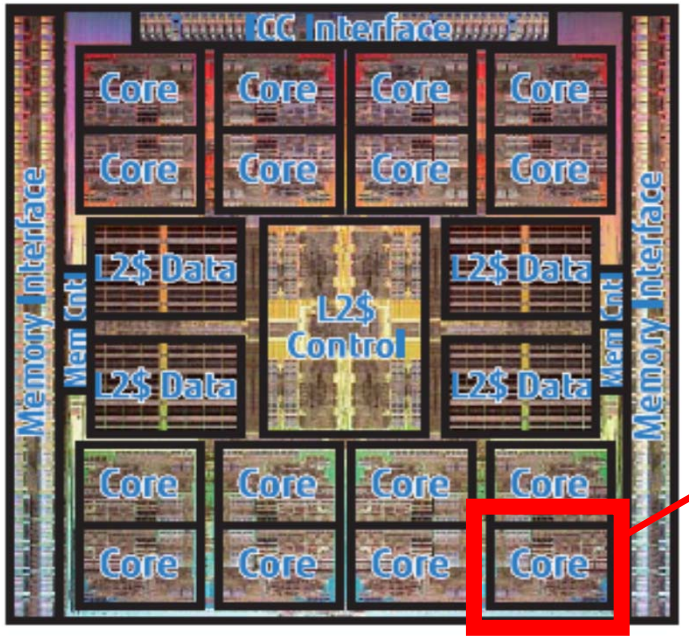
FX10 (Oakleaf-FX) Room Layout

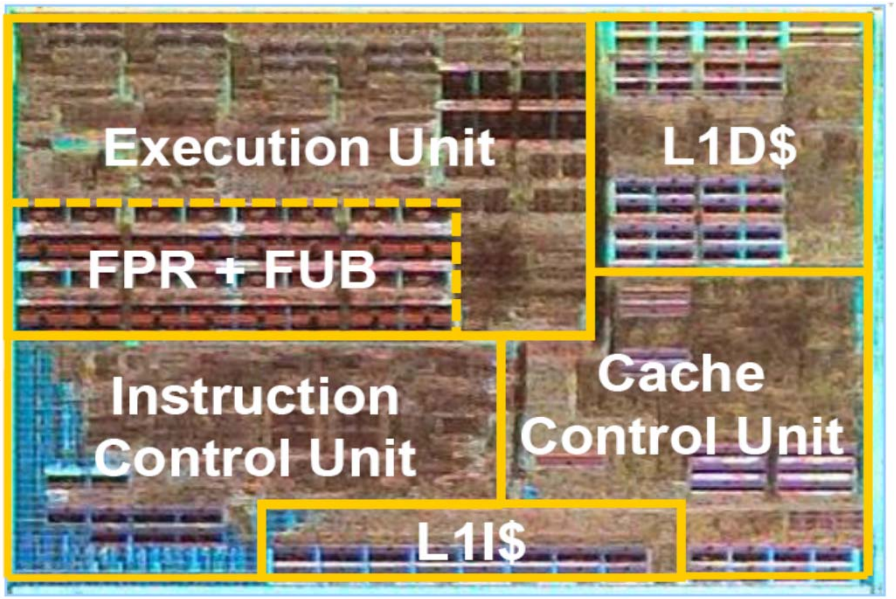


FX10（Oakleaf-FX）の概要

- ピーク性能1.13PFLOPS
  - 総メモリバンド幅：398 TB/sec.
- 周辺装置込み最大消費電力<1.40MW（Linpack最大時）
  - 空調込み2.00MW未満，1.043 PFLOPS, 1.177 MW
- SPARC64™ IXfx（16コア）
- 6次元メッシュ／トーラスネットワーク
  - Tofuインターコネクト
  - リンク当りバンド幅：5GB/sec×2，Bi-Sectionバンド幅：6 TB/sec
- 高性能ファイルシステム
  - FEFS（Fujitsu Exabyte File System）（Lustreベース）
- 通常運転～省電力運転の柔軟な切り替え
- 「京」との互換性
- 多様なオープンソースライブラリ・アプリケーション
- Flat-MPI，Hybrid共に高い計算性能

SPARC64™ IXfx

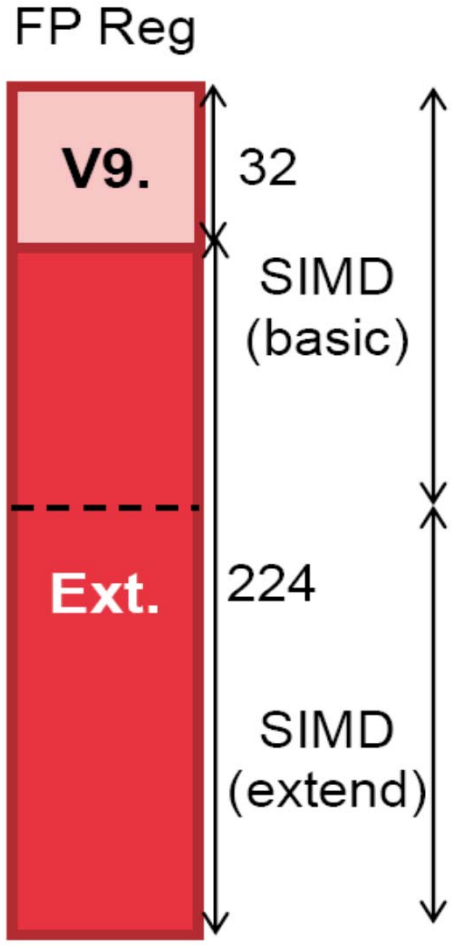


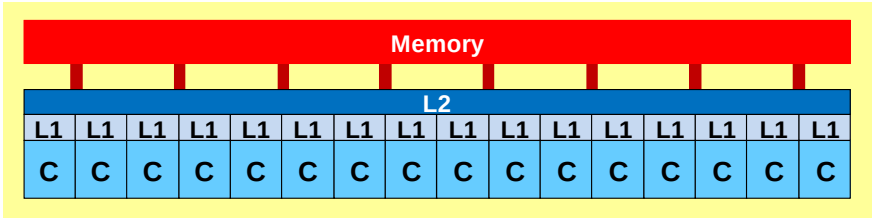


CPU	SPARC64™ IXfx 1.848 GHz	SPARC64™ VIIIfx 2.000 GHz
Number of Cores/Node	16	8
Size of L2 Cache/Node	12 MB	6 MB
Peak Performance/Node	236.5 GFLOPS	128.0 GFLOPS
Memory/Node	32 GB	16 GB
Memory Bandwidth/Node	85 GB/sec (DDR3-1333)	64 GB/sec (DDR3-1000)

SPARC64™ IXfx

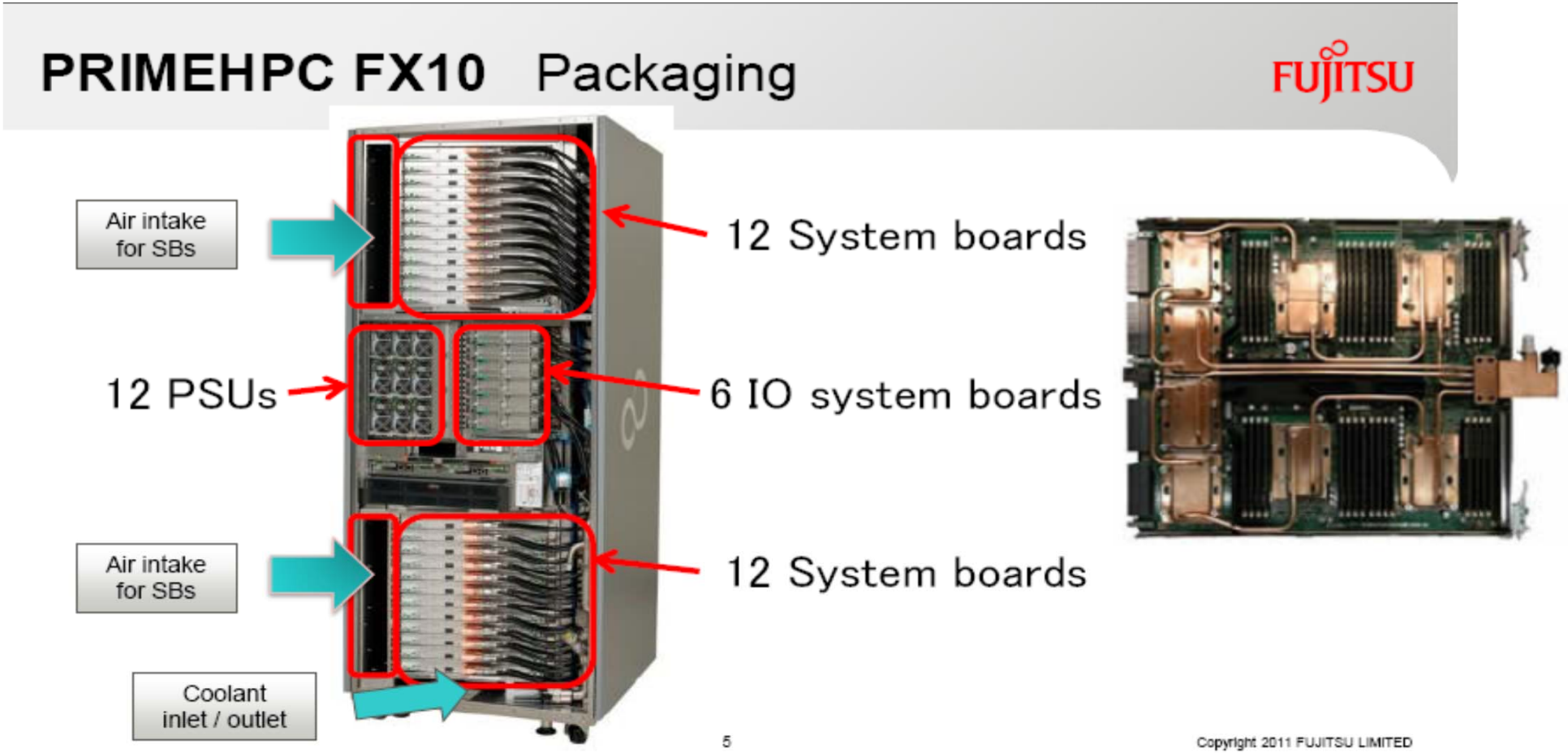
- HPC-ACE (High Performance Computing – Arithmetic Computational Extensions)
  - Enhanced instruction set for the SPARC-V9 instruction set arch.
    - High-Performance & Power-Aware
  - Extended number of registers
    - FP Registers: 32→256
    - Software Pipelining is useful
  - S/W controllable “sector” cache
- UMA, not NUMA
- H/W barrier for high-speed synchronization of on-chip cores





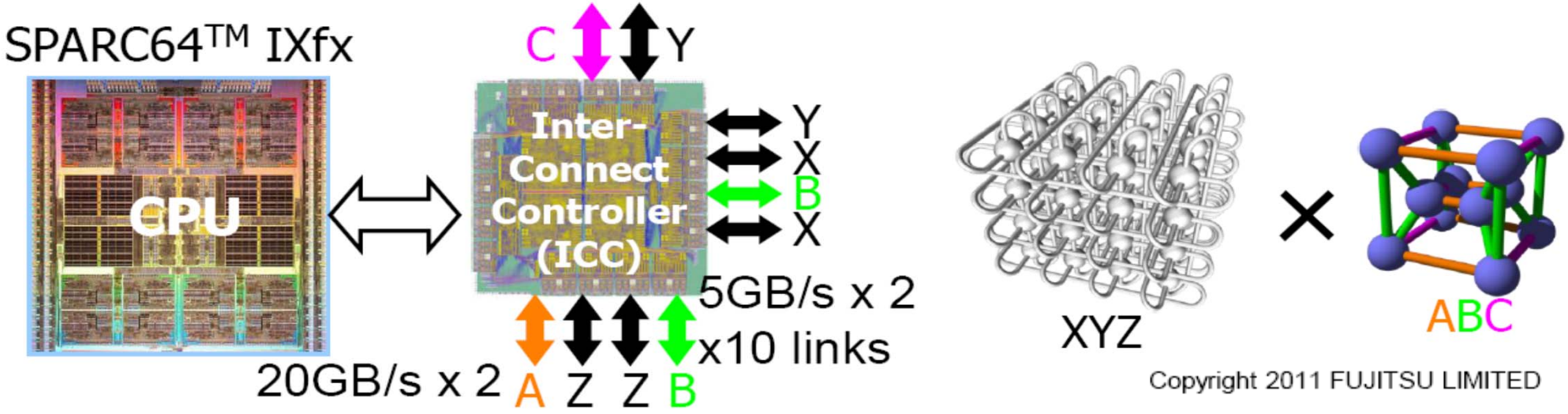
ラック構成

- システムボード：4ノード
- 1ラック：24システムボード，96ノード
- 50ラック，4,800ノード，76,800コア



Tofuインターコネクト

- ノードグループ
  - 12ノード
  - A軸・C軸：システムボード内4ノード結合，B軸：3ボード結合
- 6D：（X,Y,Z,A,B,C）
  - ABC 3D Mesh：ノードグループの12ノードを結合：　2×2×3
  - XYZ 3D Mesh：ABC 3D Meshグループを結合：　10×5×8
- ネットワークトポロジーを指定したJob Submission可能
  - 実行されたXYZは知ることができる



様々なサービス

- HA8000（T2K）における様々なサービスをFX10へ移行
- 教育利用
  - 学部・大学院講義（学外も含む）：無料
  - 試行アカウント付講習会（企業ユーザーも参加可能）
- 若手支援
  - 45歳以下の若手：無料
  - 科研費，学際大規模情報基盤共同利用共同研究拠点（8センター）公募型研究への進展が期待される
- 企業利用
  - 大規模計算普及，社会貢献，年4回募集
  - 通常有償利用：3社
  - トライアルユース（有償・無償）：5社（＋1社）
- 大規模HPCチャレンジ

大規模HPCチャレンジ

- http://www.cc.u-tokyo.ac.jp/service/4800hpc/
- 月1回1日（24時間），4,800ノード（全計算ノード）を1グループで占有して実行できる，公募制，無料。
- FX10ユーザー以外も応募可能である。
- 成果公開を義務づける
  - センター広報誌への寄稿
  - センター主催各種催しでの発表，各種外部発表への情報提供
  - 速報結果の査読付国際会議への投稿等による迅速，国際的な成果公開が望ましい。
- 企業からの申し込みも受け付ける（成果公開を義務づけ）
- 自作プログラム，オープンソースプログラム利用に限定
- 試験運転期間中は月2回（合計6回），1回48時間占有

大規模HPCチャレンジ（試験運転期間）

- 採択課題

課題名	代表者（所属）
急減圧液体における気泡分布関数の数値的解析	渡辺 宙志 東京大学物性研究所
電磁流体コードによる惑星磁気圏シミュレーション性能測定	深沢 圭一郎 九州大学情報基盤研究開発センター
2次元フラストレート系の計算科学的研究	中野 博生 兵庫県立大学大学院物質理学研究科
超並列重力多体問題シミュレーションコードの性能測定	石山 智明 筑波大学計算科学研究センター神戸分室
大規模グラフ処理ベンチマーク Graph500のスケーラブルな探索手法による性能評価	**鈴村 豊太郎 東京工業大学**
100億超格子を用いた自動車の大規模流体解析への挑戦	小野 謙二 東京大学生産技術研究所
ポストペタスケール環境における大規模疎行列解法のための数値計算・通信ライブラリに関する研究	林 雅江 東京大学情報基盤センター

4th Graph 500 List (June 2012)

	Installation Site	Machine	Number of nodes	Number of cores	Problem scale	GTEPS
1	**ANL**	**Mira/BlueGene/Q**	**32768**	**524288**	**38**	**3541.00**
1	**LLNL**	**Sequoia/BlueGene/Q**	**32768**	**524288**	**38**	**3541.00**
2	DARPA	Power 775, POWER7	1024	32768	35	508.05
3	**ITC, U.Tokyo**	**Oakleaf-FX**	**4800**	**76800**	**38**	**358.10**
4	GSIC, Tokyo Tech	TSUBAME 2.0	1366	16392	35	317.09
5	Brookhaven National Laboratory	BlueGene/Q	1024	16384	34	294.29
6	ANL	Vesta/BlueGene/Q	1024	16384	34	292.36
7	NASA-Ames	Pleiades - SGI ICE-X	1024	16384	34	270.33
8	NERSC/LBNL	Hopper/Cray XE6	4817	115600	35	254.07
9	NNSA/IBM T.J. Watson	Blue Gene/Q Prototype II	4096	65536	32	236.00
10	STE Lab, Nagoya U.	PowerEdge R815 Opteron 6174	4	192	22	116.23

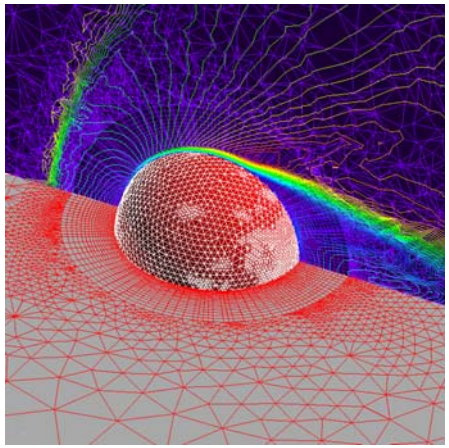
Oakleaf-FXの成果は大規模HPCチャレンジ（東工大鈴村准教授らのグループ）による。
November 2011の一位は253.4 GTEPS（BlueGene/Q Prototype II，4,096ノード（4ラック），32，IBM T.J. Watson）

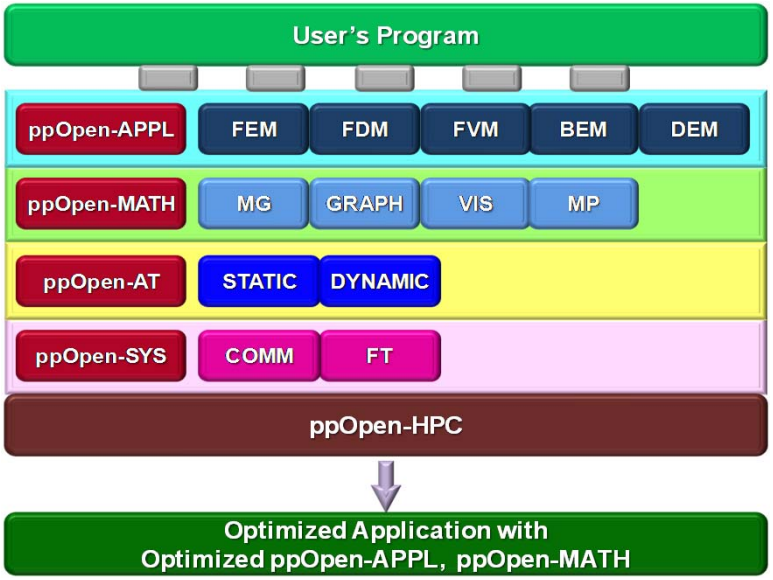
スーパーコンピューティング部門(2/2)

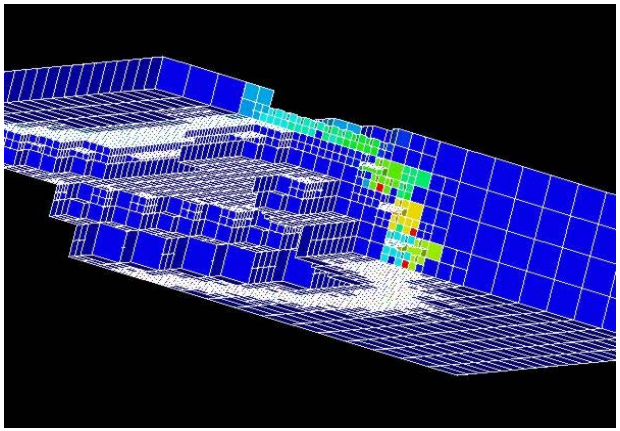
- 研究
  - 独自研究
    - コンピュータシステム，ソフトウェア，数値解法
  - 利用者(科学，工学分野)との共同研究
- 普及・人材育成
  - 学際計算科学・工学　人材育成プログラム
    - 全学的なHPC(High-Performance Computing)教育
    - http://nkl.cc.u-tokyo.ac.jp/CSEedu/
  - お試しアカウント付き講習会(Oakleaf-FX)
  - RIKEN AICS Summer School
    - http://www.aics.riken.jp/jp/library/event/2012-summer-school.html
- 広報活動
  - スーパーコンピューティングニュース(年6回＋特集号)

研究例
並列計算機による新しい科学の開拓

- 並列計算機による連立一次方程式解法等の大規模シミュレーションを支える数理的基盤の研究を，物理，モデリング，計算機ハードウェア等様々な観点から実施しています。
  - T2Kオープンスパコン等のスーパーコンピュータを駆使した研究を実施し，大規模シミュレーションによる新しい科学の開拓に貢献しています。







お試しアカウント付き講習会（2012年度）

http://www.cc.u-tokyo.ac.jp/publication/kosyu/schedule_kosyu.html

名称	期間	時期（予定）	内容
MPI基礎	1日半 ～2日	2012年7月2･3日 2012年9月3･4日 2013年3月4･5日	• MPIによる並列プログラミングの基礎に関する講習，実習 並列化の基礎知識 • MPIのAPI説明 • 行列積の並列化実習 • makeを使った分割コンパイルと並列処理 • Oakleaf-FX（東大）による実習
MPI応用	1日半	2012年10月中旬 2013年　1月中旬	• MPIを使用した並列アプリケーション開発手法に関する講習，実習 有限体積法によるポアソン方程式ソルバーの概要 • 並列データ構造の考え方 • 領域分割手法 • 並列化手法 • Oakleaf-FX（東大）による実習
OpenMP (基礎＋応用)	1日半 ～2日	2012年12月上旬 2013年　2月中旬	• OpenMPによるマルチコアプロセッサ向け並列プログラミング，最適化手法に関する，実アプリケーションに基づく講習，実習 有限体積法によるポアソン方程式ソルバー，ICCG法の概要 • OpenMPの基礎 • リオーダリングによる並列化，最適化 • Oakleaf-FX（東大）による実習
ライブラリ利用	2日	2012年12月19･20日 2013年　2月　4･5日	• 密行列ライブラリBLAS, LAPACK, ScaLAPACK、および、疎行列ライブラリPETsc, Lisの利用法に関する講習，実習 数値解法の原理と特徴の説明 • 数理的モデリング，離散化，データ格納 • ブロック化、データ分散の考え方 • Oakleaf-FX（東大）による実習