T2K筑波システムの概要と利用プログラム計画

朴　泰祐
計算科学研究センター 副センター長／
システム情報工学研究科 教授


内容

- 筑波大学計算科学研究センター紹介
- 次期スーパーコンピュータ運用体制
- 次期スーパーコンピュータのコンセプト
- **T2K Open Supercomputer Alliance**
- 次期スーパーコンピュータのシステム概要
- 次期スーパーコンピュータ利用プログラム
- つくば地区特別プログラム

CCS: 筑波大学計算科学研究センター

- Center for Computational Sciences
- 広範な先進的計算科学分野と高性能計算工学分野の研究者を結集
  - アプリケーション分野
    - 素粒子物理，宇宙物理，物性物理，地球環境，生物情報
  - 計算機工学分野
    - 高性能計算システム，グリッド，大規模データベース，マルチメディア
- 応用側のニーズとシステム側のシーズの融合
- 実応用に即したHPC研究の日常的推進
- 応用分野に必要な計算機資源を自らの手で開発
  - QCDPAX, CP-PACS, PACS-CS, FIRST

CCSの計算機資源

＜計算科学全般汎用＞

PACS-CS (2006-)
14.4 TFLOPS
2006.6 #34/TOP500

＜計算宇宙物理専用＞

FIRST (2005-)
汎用：3.5 TFLOPS
専用：35 TFLOPS

＋

筑波大学
次期スーパー
コンピュータ
T2K筑波システム

筑波大学次期スーパーコンピュータの運用体制

- 筑波大スパコンはH18年度まで、学内教育用計算機／学内ネットワーク等と共に、学術情報メディアセンターが管理・運用

- H19年度より、スパコンに関する管理・運用を計算科学研究センターに移行
  - 旧VPP5000システムの運用停止（平成19年2月末）
  - 次期スパコン運用開始までのジョブを計算科学研究センターのフロントエンド計算システム（FCS-V）及びPACS-CSで賄う

- H20年度運用開始の次期スパコンについては計算科学研究センターが調達・運用を行う
  - 約1年間の準備期間を経て超並列・超高速計算システムの導入
  - 計算科学研究センターの経験を生かし、次期スパコンの仕様策定から調達を実施

次期スーパーコンピュータのコンセプト

- 益々高まる超高速・大容量コンピューティング需要への対応
  - 大規模シミュレーション
  - 大容量データ処理
- 従来のベクトル型スパコン向け応用だけでなくより幅広い応用への対応
  - 計算科学／計算工学
  - データベース／データ解析
  - 大規模並列処理／膨大な逐次処理
- 科学技術計算専用ベクトル計算機から超並列クラスタ型計算機へ
  - 旧VPP-5000⇒Linuxクラスタ
  - Parallel Vector⇒Multi-core Parallel Scalar
- 全国共同利用の大型計算機センターに劣らない超高速システムを導入

T2K Open Supercomputer Alliance（T2Kアライアンス）

- 大規模・超高速計算のニーズに応えるために（次期スパコンのコンセプト）
  - 大幅な計算性能拡大のニーズ
  - 従来の「スパコンユーザ」の枠にとらわれない、幅広いユーザ層の獲得
  - 世の中の他の計算機システムとのアプリケーション互換性
- 上記コンセプトが、東京大学・京都大学の次期スーパーコンピュータのコンセプトと合致
  ⇒**T2K Open Supercomputer Alliance**
  (Tsukuba, Tokyo & Kyoto)
- オープンアーキテクチャを想定したスパコン導入
  - システムアーキテクチャのオープン性
  - システムソフトウェアのオープン性
  - ユーザアプリケーションのオープン性

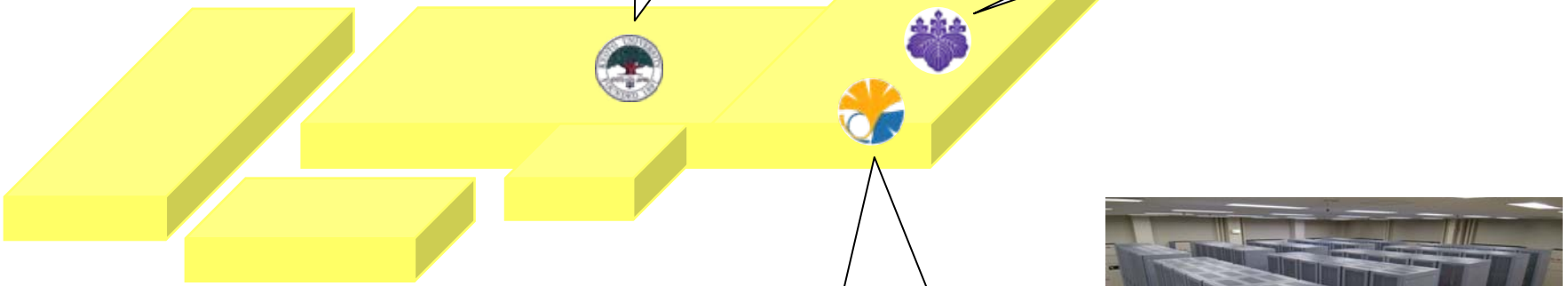
T2K Open Supercomputer Alliance（続き）

- 筑波大・東大・京大のスパコン調達が全て平成20年6月の新規運用開始で揃う
- 従来のメーカー主導スパコン調達では性能／価格比が頭打ち
  ⇒コモディティベースの超並列システムの導入
- 3大学で基本仕様を統一し調達を行う
  - ベンダーの「やる気」を促進
  - Feasibility Study結果の共有
  - アプリケーション開発・展開コストの低減
  - グリッド運用

T2K Open Supercomputer Alliance
メンバー組織

Kyoto U.
学術情報メディア
センター

U. Tsukuba
計算科学研究
センター

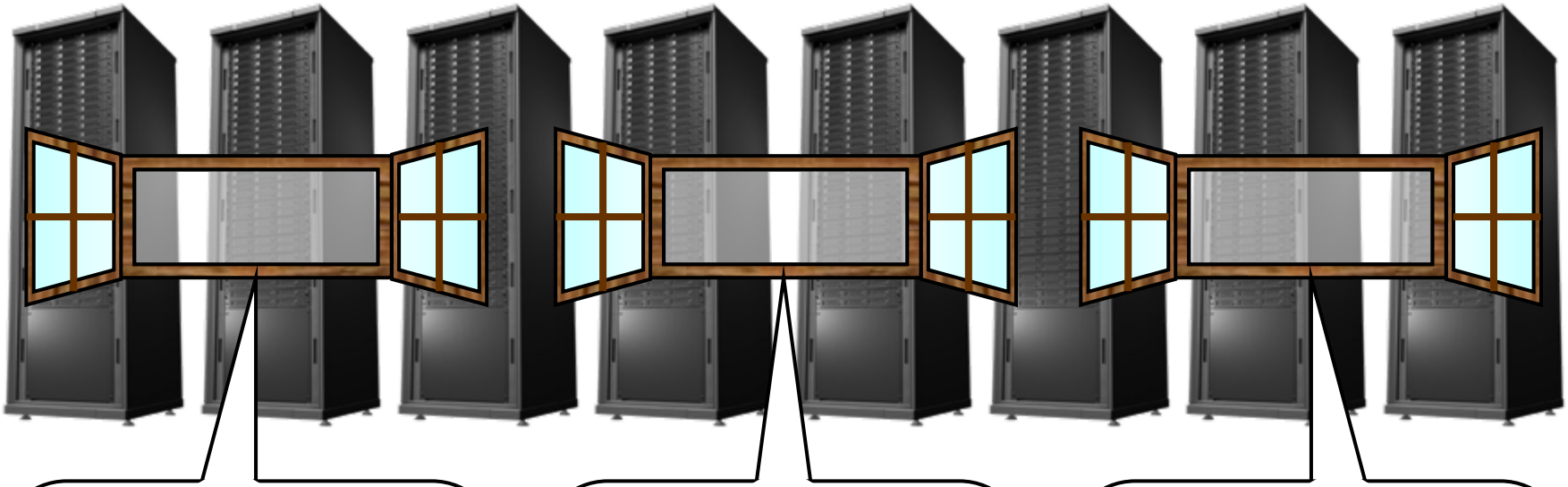


U. Tokyo
情報基盤センター

なぜallianceを結ぶか？

- スパコン調達のスタイルを変える
  - 従来のベンダー主導の受身なスタイルからの脱却
  - 既存のマーケットにある商品から選ぶのではなく
- 新しい方式で
  - 大学主導の調達
  - Technology marketを中心に
- 真の高性能を目指して
  - 最先端テクノロジの利用
  - 広範囲のアカデミアユーザを対象

何がオープンなのか？



Open Hardware Arch.

- **コモディティテクノロジ**
  e.g. x86, IB/Myri-10G
  → **現在のITマーケットで最もcost/performanceの良いもの**
- HPC**向けの特殊ハードは対象としない**

Open Software Stack

- **オープンソース＆標準システムソフトウェア**
  e.g. Linux, MPI, Globus
  → **オープンソースなHPC向けミドルウェア＆ライブラリ**

Open to User's Needs

- Floating Point**ユーザだけでなく，**
- Integer**ユーザ（大規模データ処理等）を含めた幅広いユーザを対象に**

T2K共通仕様（主なもの）

- ノード当たり145FLOPS以上のマルチソケットx86アーキテクチャ
  - 2007～2008年に出揃うマルチコア×マルチソケットCPUを想定
- ノード当たり5GB/s以上のマルチリンクネットワーク
  - 高性能ネットワークをさらにマルチリンクで増強
- ノード当たり32GB以上の主記憶
  - 従来スパコンの多くのアプリケーションを吸収
- ノード当たり130GB以上のローカルディスク
  - ファイルサーバだけに頼らない高バンド幅・大容量分散ディスク
- ノード当たり40GB/s以上のメモリバンド幅
  - 大容量アプリケーションのサポート
- Linux, Fortran, C, C++, MPI
  - 標準的なプログラミング環境／並列化環境をサポート

筑波大学次期スーパーコンピュータのシステム

- 「T2K筑波システム（仮）」

- システム納入ベンダー
  - 計算ノード本体：米Appro International 社
  - システム構築、保守：クレイジャパンインク
  - システム納入：住商情報システム

T2K筑波システムの概要

- システム規模・緒元
  - 計算ノード数＝648台 (Appro XtremeServer-X3)
  - 計算性能＝147.2 GFLOPS/node x 648 = 95.39 TFLOPS
  - 通信総バンド幅＝ 8GB/s/node x 648 = 5.18 TB/s
  - ファイルシステム容量＝800TB (RAID6, user space)
- コモディティプロセッサとコモディティネットワークを用いた「オープンアーキテクチャ」
- 複数台の最先端のマルチコアプロセッサで計算ノードを構成し、これを多数接続
- 超高性能計算ノード間通信を支えるmulti-rail高バンド幅ネットワークリンク、数十～数百ノード単位での並列処理を可能とする高バンド幅なFat Tree相互結合網

T2K筑波システムの概要(続き)

- 全ノードから共有される大容量ファイルサーバ
  - 総ユーザ使用容量約800TB
  - 全ノードから Fat Tree Infiniband網を通してマルチリンク接続
  - クラスタファイルシステムによる高バンド幅接続(多数の並列ジョブから並列アクセス可能)
  - RAID6+Lustreファイルシステムによる多階層多重冗長構成
- SGE (SUN Grid Engine)による効率的なジョブキュー／ノード管理
- 論理クラスタ構築・管理システム(ACE)による柔軟なシステム構造管理
- 複数のログインノードによりインタラクティブ処理の負荷分散

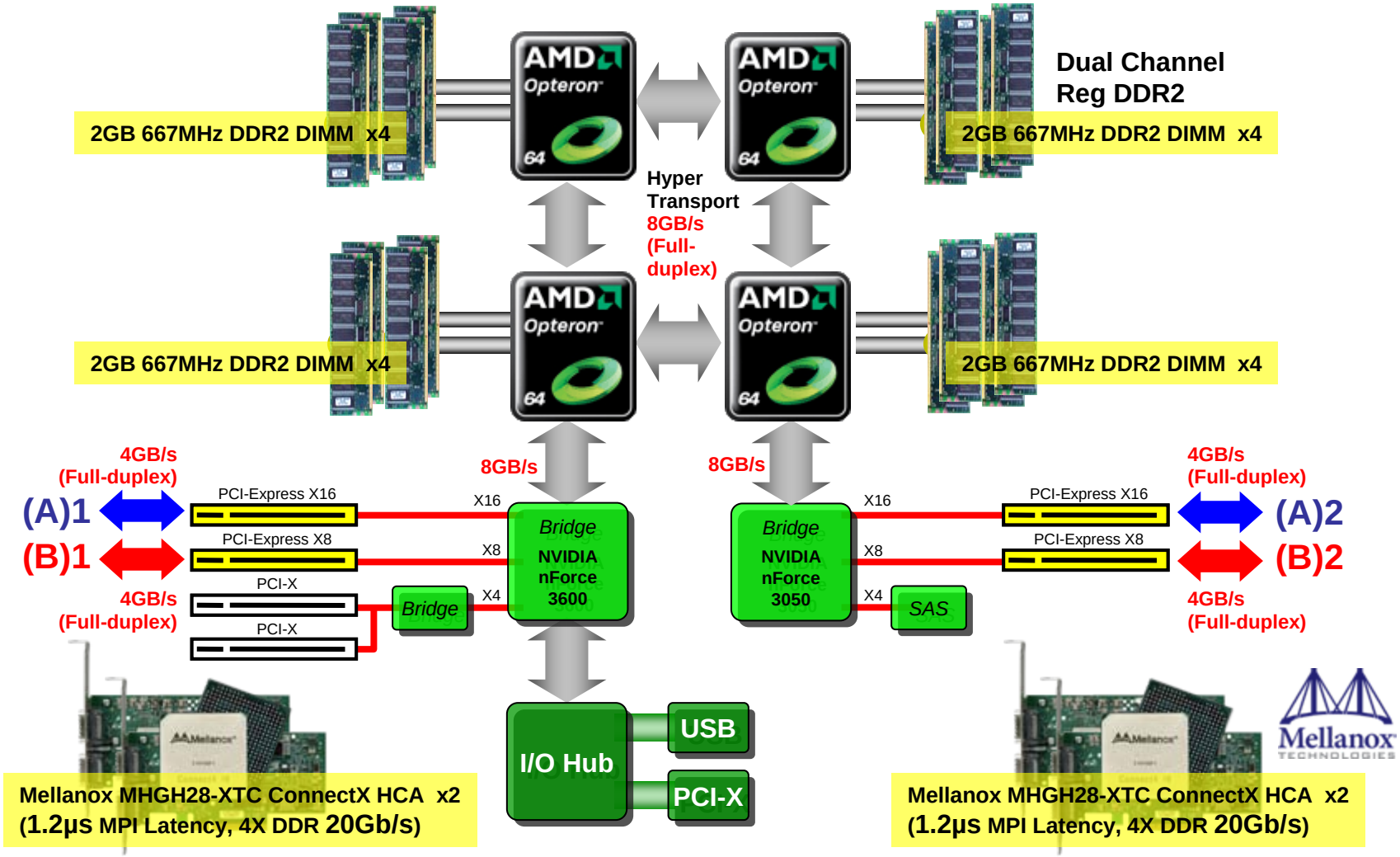
計算ノードの構成

- **Multi-core & multi-socket構成による超高性能演算**
  - AMD quad-core Opteron 8000シリーズ (Barcelona) の採用
  - 4 socket / node 構成⇒　ピーク演算性能 147GFLOPS/node
  - OpteronのNUMAアーキテクチャにより、各プロセッサのメモリバンド幅を効率的に利用⇒　メモリバンド幅 40GB/s/node
- **Multi-rail ネットワークによる超高バンド幅ネットワークリンク**
  - 4xDDR Infiniband (Melanox ConnecX) x 4 / node
  - Quad-rail Infiniband をトランキングすることにより高バンド幅ネットワークリンクを実現⇒　ピーク通信性能 8GB/s/node
- **柔軟なプロセス構成による最適化された並列処理**
  - MVAPICH (modified by Appro) の multi-rail configuration により、MPIプロセス当たりの利用rail数を制御可能
    ⇒ノード上の並列プロセス数に応じてmulti-railを使い分け可能

計算ノードの詳細

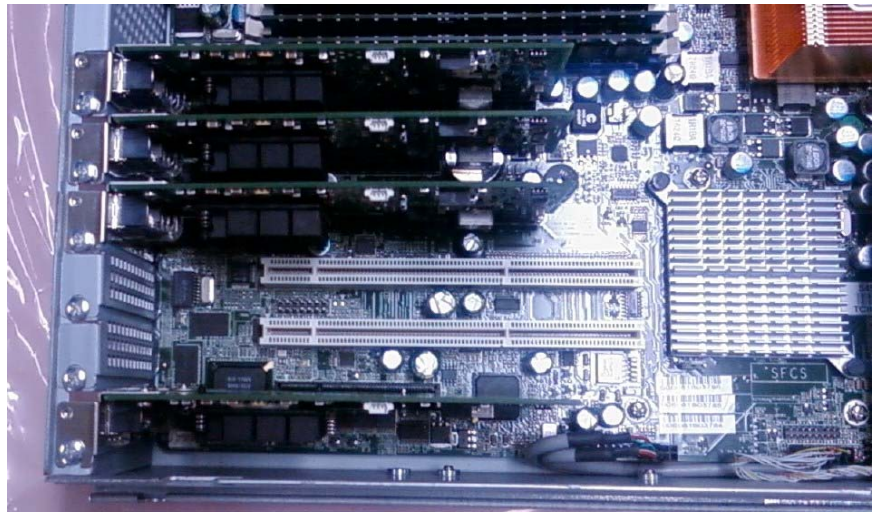
- Appro XtremeServer-X3 (x 648 node)
  - CPU: AMD Opteron Quad-core 8000 series (Barcelona)
    2.3GHz, peak perf. 36.8GFLOPS
    x4 processor (16 core)
  - Memory: 32GB (667MHz DDR2 x 16), 42.688 GB/s
  - HDD: 250GB x 4 (SATA-II, RAID1)
  - Network Interface: DDR Infiniband Melanox ConnectX x 4
    supported by PCI-ex8 (x 4)
  - GbEthernet x 3
  - 2U chassis
  - SSE命令の同時発行により最大 4FLOP/clock
    ⇒ 36.8 GFLOPS/socket
    ⇒ 147.2 GFLOPS/node

計算ノードのブロックダイアグラム



計算ノードの内部

ノードシャーシ内部



Infiniband ConnectX x 4
(各PCI-Express x 8lane)

インターコネクションネットワーク

- 4 rail / node のInfiniband ConnectXリンクを Full-bisectionバンド幅のFat Tree網によって結合
  ⇒ Full-bisectionバンド幅＝8GB/s x 648 = 5.18 TB/s

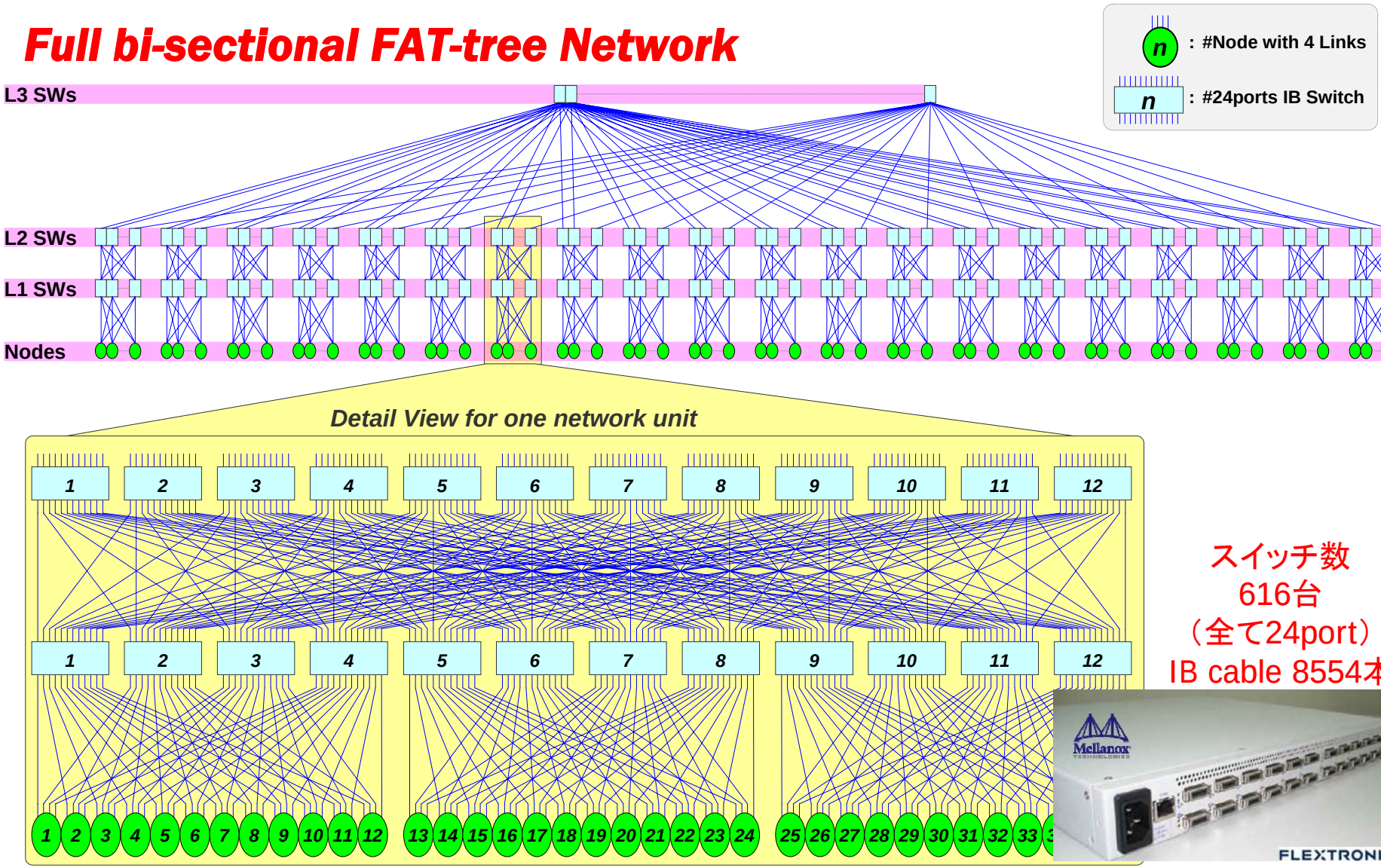
- ノードからのリンク間通信に制約がないため、4本のリンクの柔軟な利用が可能

- Fat Tree Networkは全て24 portの Infiniband switchで構成
  ⇒3段の Fat Tree
  (Infiniband Switch: Flextronix社製)

- 総スイッチ数：616台、総ケーブル数：8554本

Infiniband Fat Tree Network

Full bi-sectional FAT-tree Network

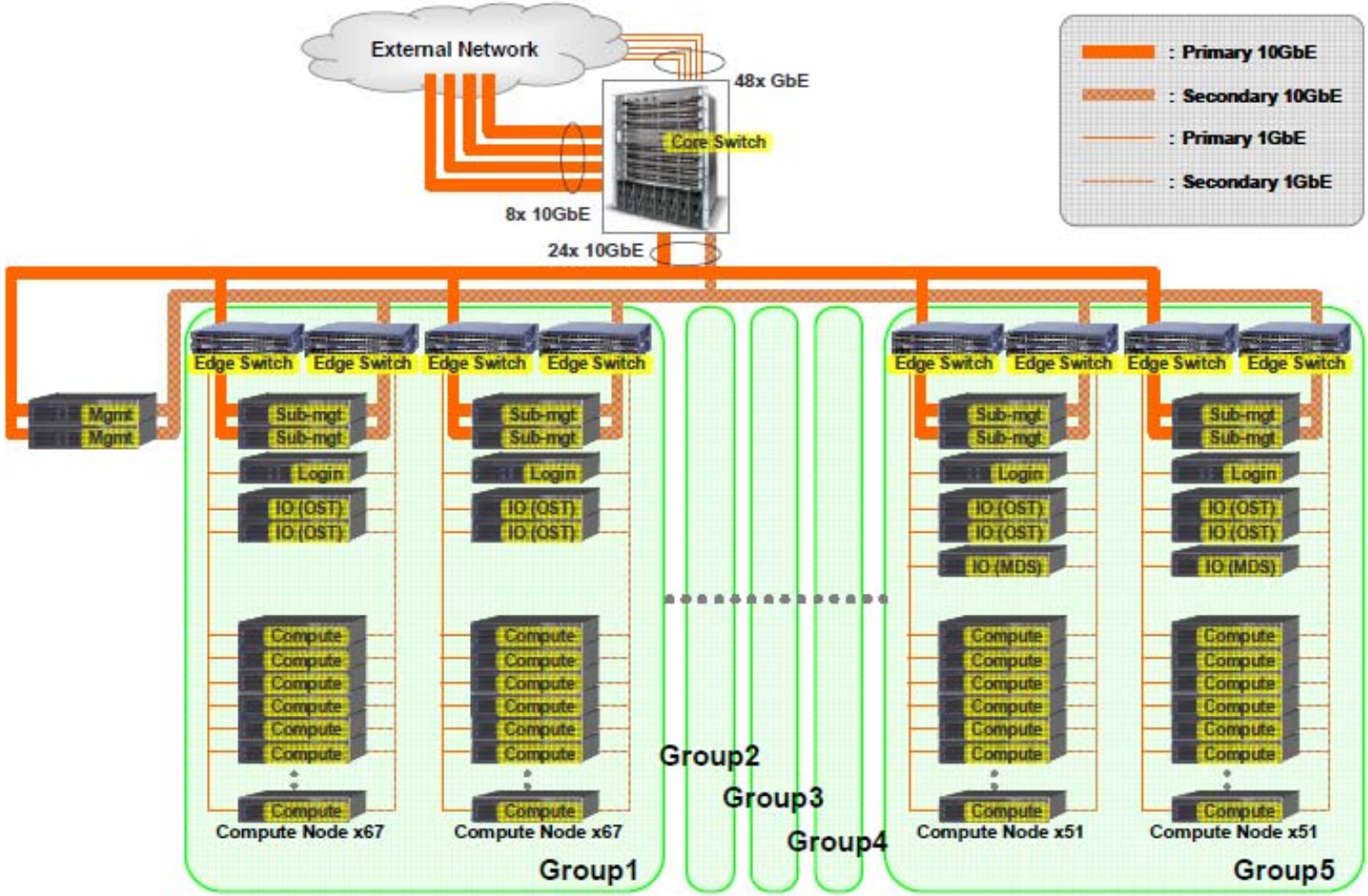


スイッチ数
616台
（全て24port）
IB cable 8554本

システム運用ネットワーク

- 外部接続 Ethernet switch (x 1)
  - Force-10 C-series
  - 10Gb-LR x 8
  - 10Gb-SR x 24
  - 1Gb-T x 48
- 内部管理用 Ethernet switch (x 20)
  - Netgear L3 stackable
  - 10Gb-SR x 3
  - 1Gb-T x 96

運用ネットワーク構成（Ethernet, 10G + 1G）

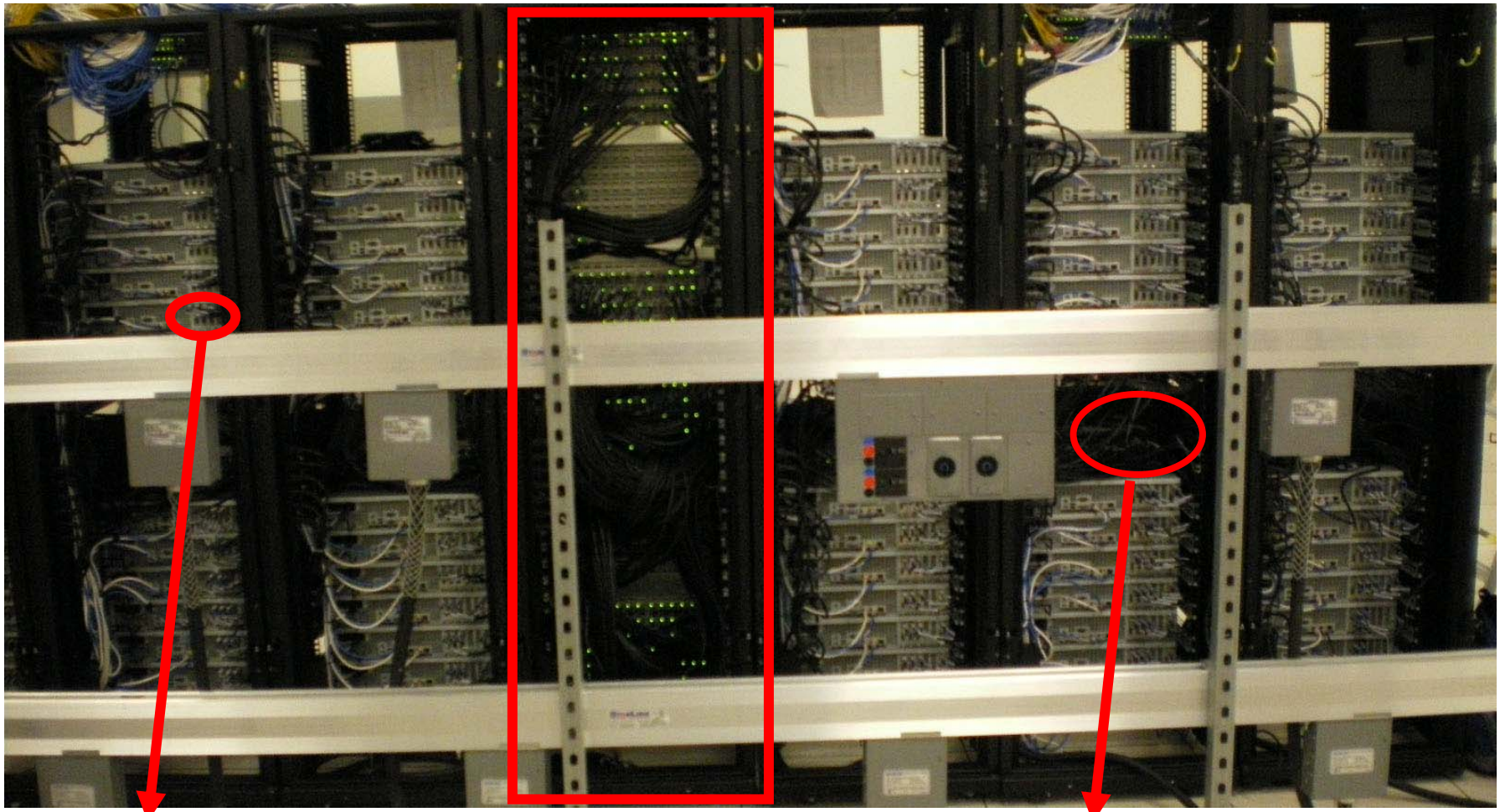


ラック構成の様子

- SU (System Unit) = 72 node, 6 node rack + 1 switch rack

ノードラック　スイッチラック　ノードラック

ラック構成（背面）

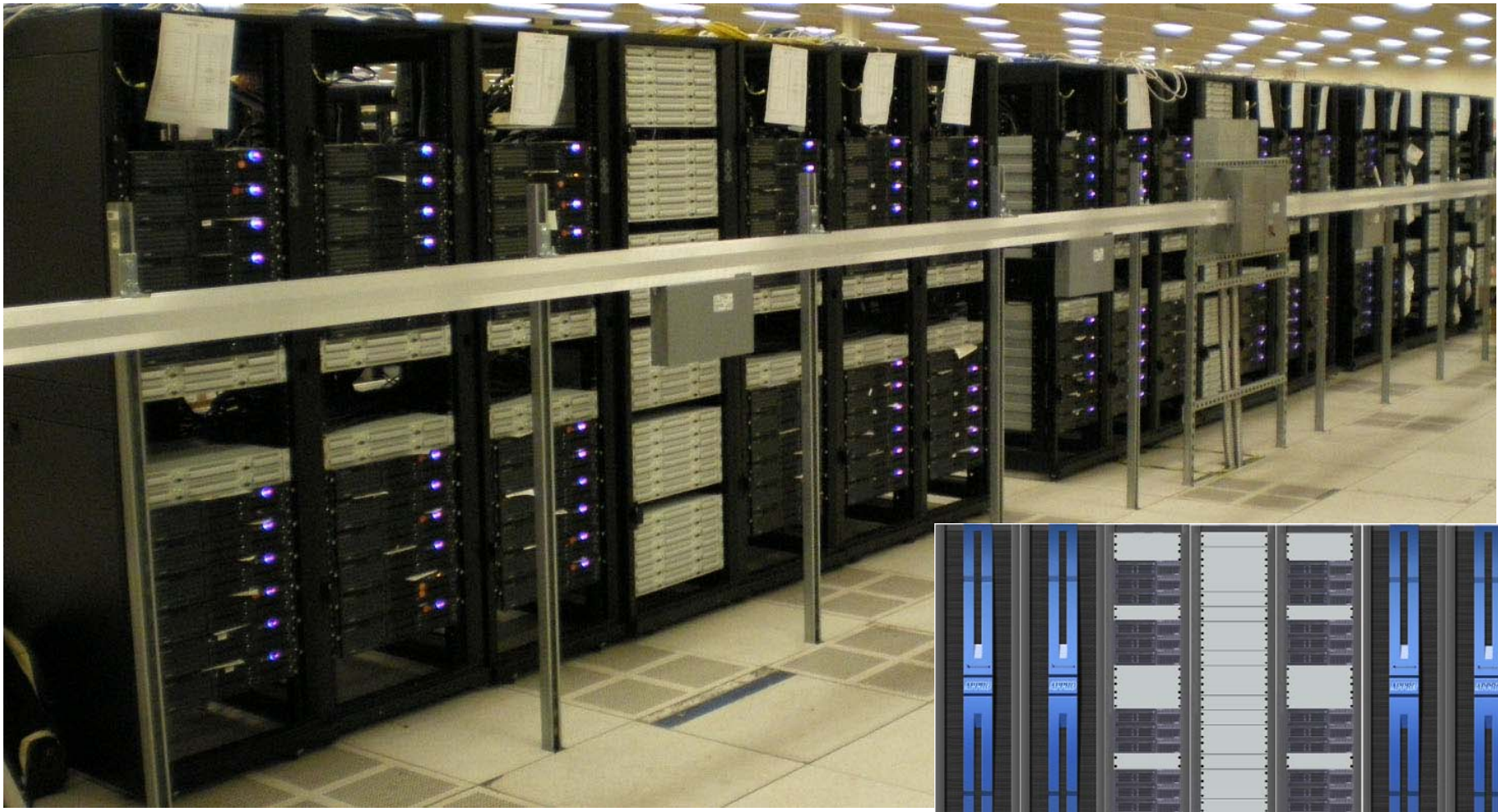


4 rail Infinibandの配線

スイッチラック

Infinibandケーブル収容スペース

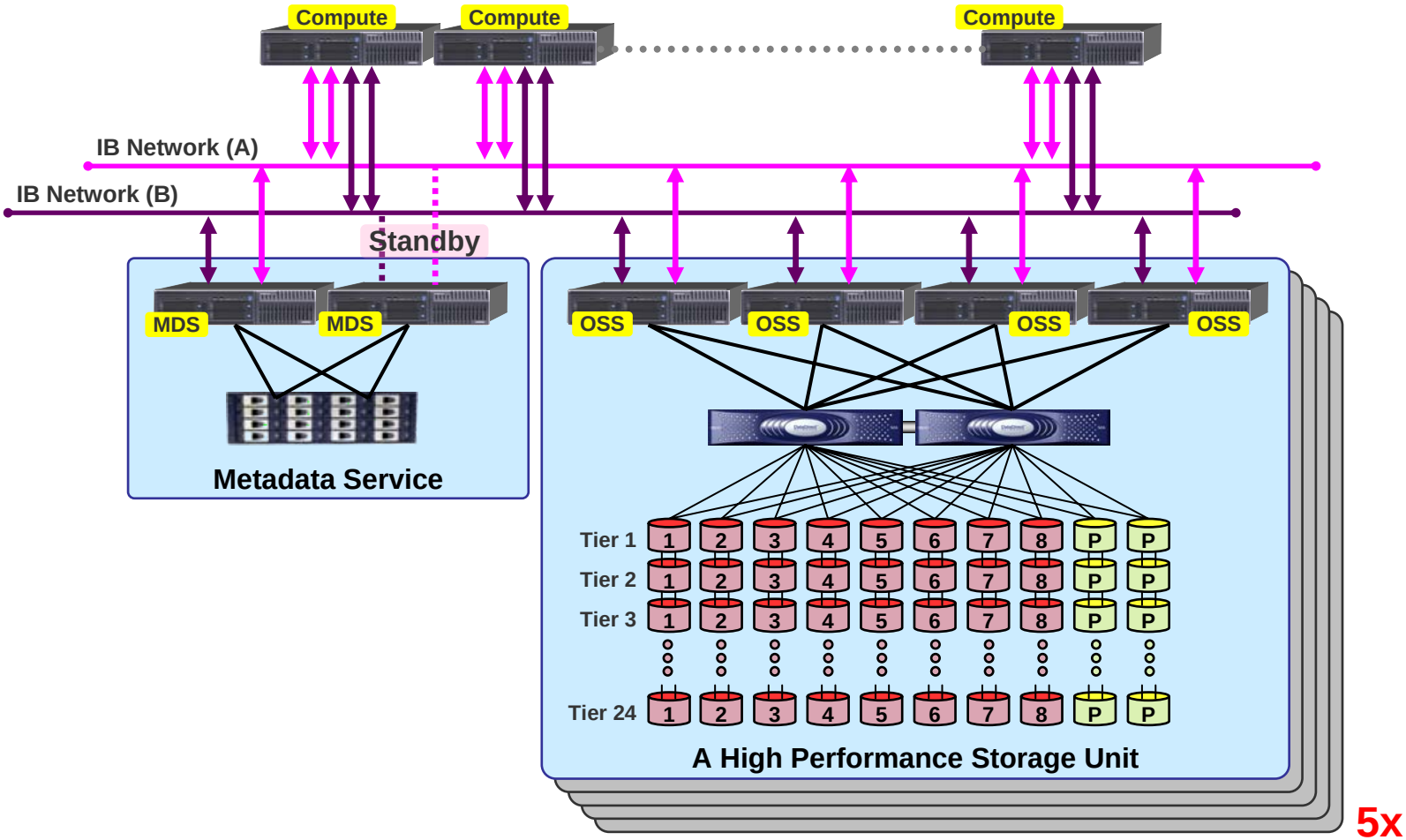
ラック構成（3 SU分）



共有ファイルシステム

- 並列処理インターコネクションのInfinibandを、ストレージネットワークとしても併用
  ⇒ Lustre cluster file system over Infiniband
- 1TB（一部 750GB）HDDを多数用い、RAID-6の冗長性ファイルシステムを構築、これをベースに Lustre 分散ファイルシステムを構築
- Lustreを構成するファイルサーバ及びMDS (Meta-Data Server)、さらにこれらを結ぶInfinibandは全て冗長化構成
  ⇒非常に高い信頼性・耐故障構成
- ユーザ利用可能領域 800TB の大規模ファイルサーバ

Lustre共有ファイルシステム



共有ファイルシステムのファイルサーバ

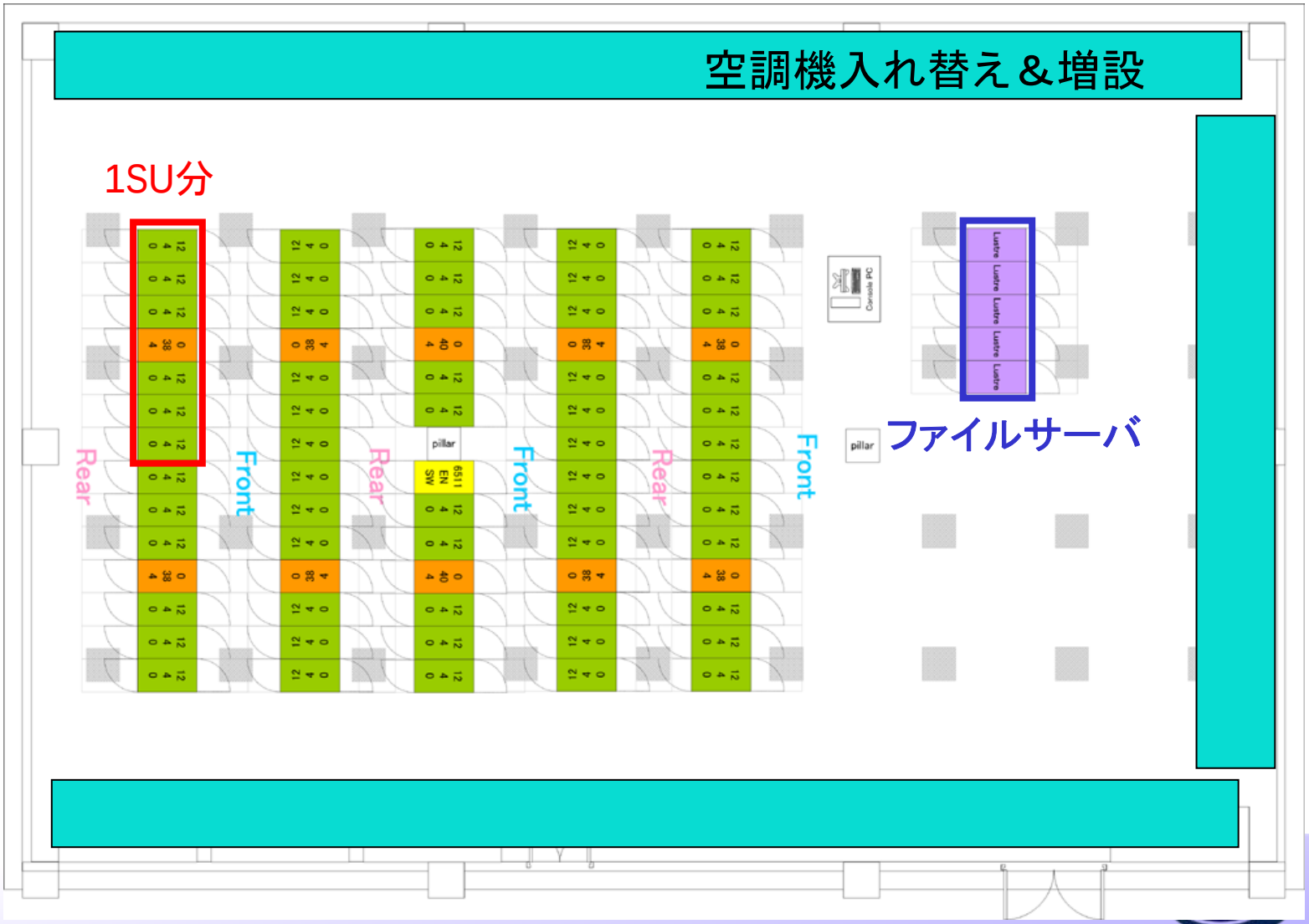
- **DDN S2A9550**

システム規模

- 全74ラック
  - 69ラック：ノード、スイッチ、他
  - 5ラック：ファイルサーバ
- 総電源容量：745kVA
  - 管理ノード、ファイルサーバ類はUPSバックアップ
  - 計算ノードは非UPS
- 電源電圧：200V単相
- 床面積：147m^2

設置場所＆フロアプラン

学術情報メディアセンター・スパコン棟１階



計算機室の現状

現在、フリーアクセスの支柱入れ替え完了

今週後半～　全ラック搬入

４月下旬　ノード／ネットワークスイッチ組み込み

T2K筑波システムのプログラミング

- ソフトウェア環境
  - OS: Red Hat Enterprise Linux v.4 WS (Linux kernel 2.6)
  - 使用可能言語：F90, C, C++, Java
  - コンパイラ：PGI (Portrand Group), Intel
  - MPIライブラリ：MVAPICH（Approによる修正版）
  - 数値計算ライブラリ：IMSL（一部ノード群）, ACML, SCALAPACK
  - チューニング環境：PGPROFR, PAPI

- プログラミング
  - 逐次プログラミング：コンパイラによる最適化
  - 並列プログラミング
    - コンパイラによる共有メモリ並列化：ノード内
    - ユーザによるOpenMP並列化：ノード内
    - ユーザによるMPI並列化：ノード間、ノード内

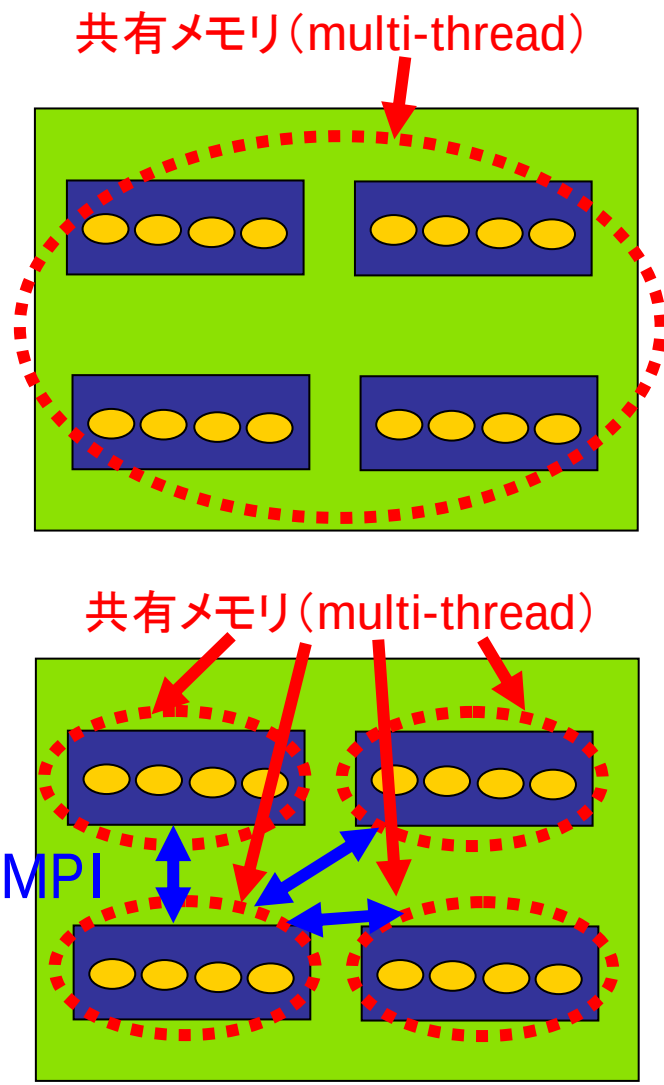
様々な並列化

- ノード内自動並列化
  - 最大16スレッド（16コア）による共有メモリ並列化
  - コンパイラによる解析に依存するため、単純ループ構造が対象
- ノード内OpenMP並列化
  - 最大16スレッド（16コア）による共有メモリ並列化
  - 標準的なOpenMP directiveによる並列化指示
- ノード間MPI並列化
  - 標準的なMPI (Message Passing Interface)による明示的な並列プログラミング
- ハイブリッド並列化
  - ノード内の共有メモリ並列化（コンパイラ自動 or OpenMP）とMPI並列化を組み合わせる

高度なハイブリッド並列化

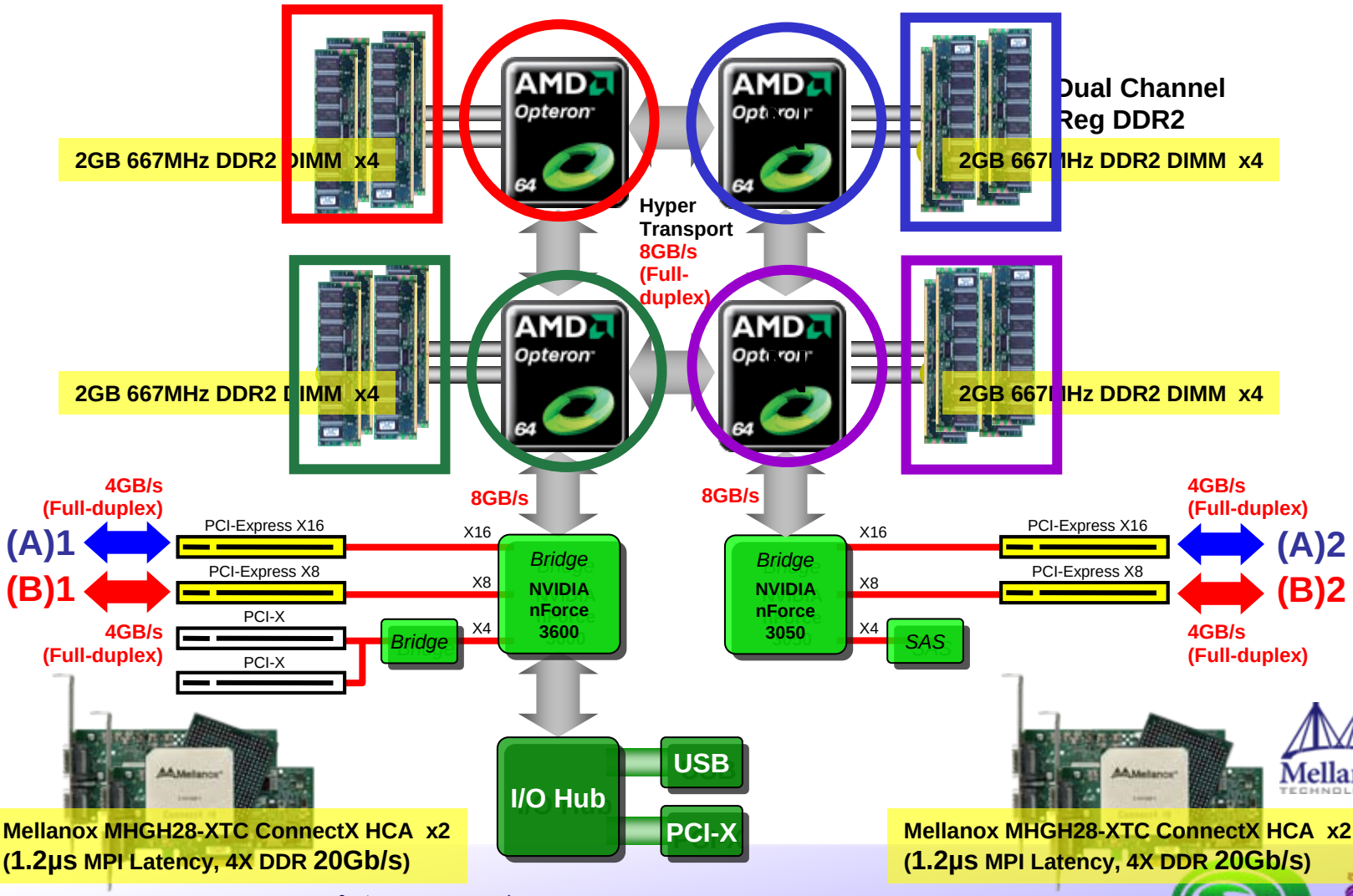
- 共有メモリ並列（スレッド並列：自動またはOpenMP）の対象はノード内全コアとは限らない
  - 各CPUコアから見た他のコアとのメモリ共有レベル
    - L1及びL2キャッシュはコア独立
    - L3キャッシュは（同一ソケット内）コア間共有
    - オフチップメモリはソケット間・コア間共有

    ⇒全16コアによる共有メモリプログラミングは性能を最大限に生かせない可能性がある
  - ソケット（4コア）内で共有メモリ並列化を行い、ノード内といえどもソケット間はMPIによるメッセージ通信という使い方も可能
    - プログラムのループ構造やデータサイズ（ワーキングセット）により最適化が可能



メモリマップとプロセスマップ

プロセス（コア）と参照データを近接メモリにマッピング可能（*numactl* 機能）

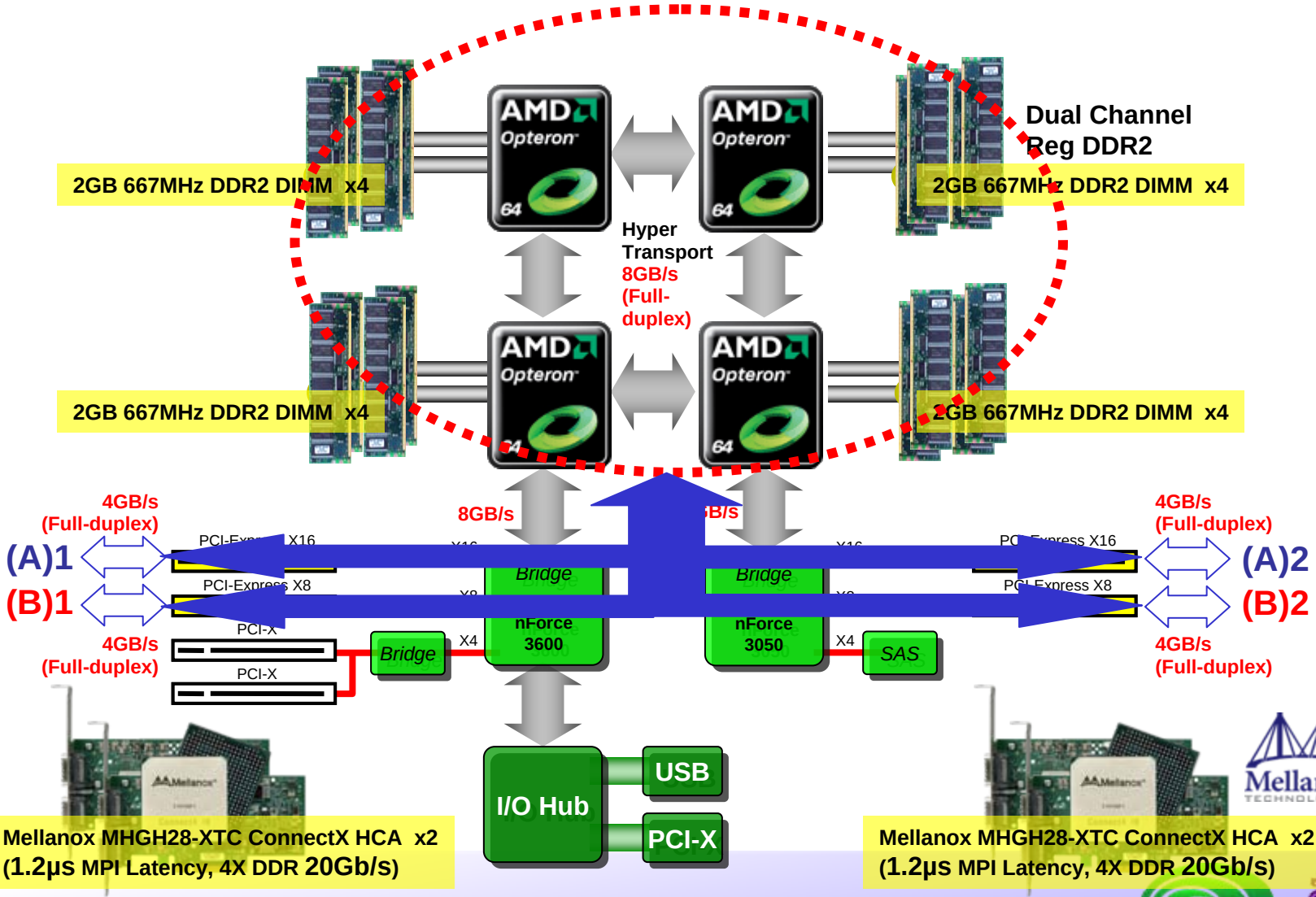


MPI通信におけるマルチリンクの利用

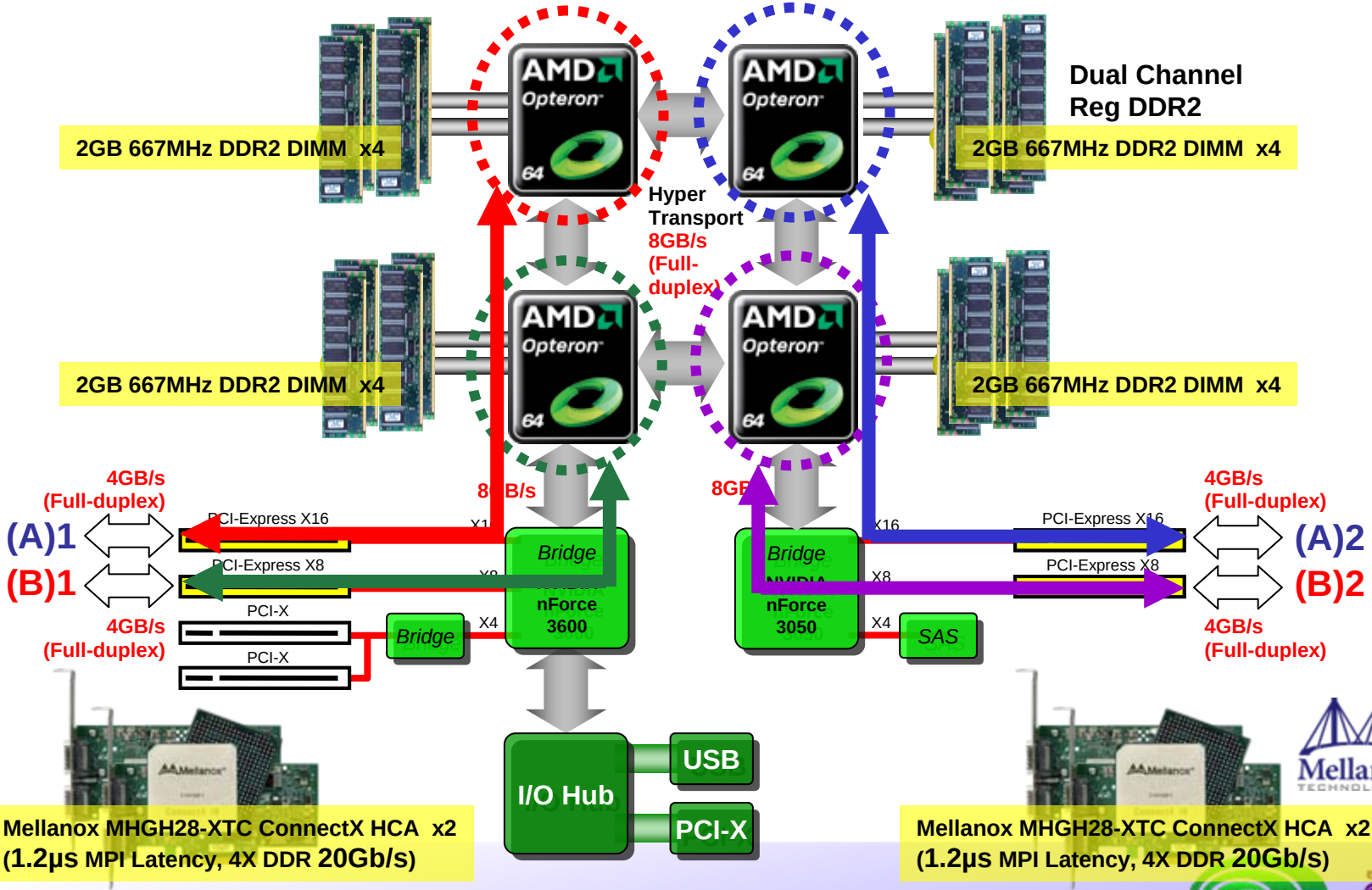
- Multi-rail の利用方法⇒様々な構成が可能
  例：
  - 4 rail / 1 MPI process (x 16 thread)
  - 4 rail / 4 MPI process (x 4 thread)
  - 1 rail / 1 MPI process (x 4 thread)
  - 4 rail / 16 MPI process (x 1 thread)

- Multi-rail Infinibandは耐故障機能を持つ
  - 平常時は全リンクを均等利用
  - リンクに故障が生じるとそのリンクを除く他のリンクで通信を継続（fail-over機能）
  - リンク間のロードバランス／耐故障性はInfinibandのSubnet Management機能によって提供

4 link / 1 MPI process (x 16 thread) の例



1 link / 1 MPI process (x4 thread) の例



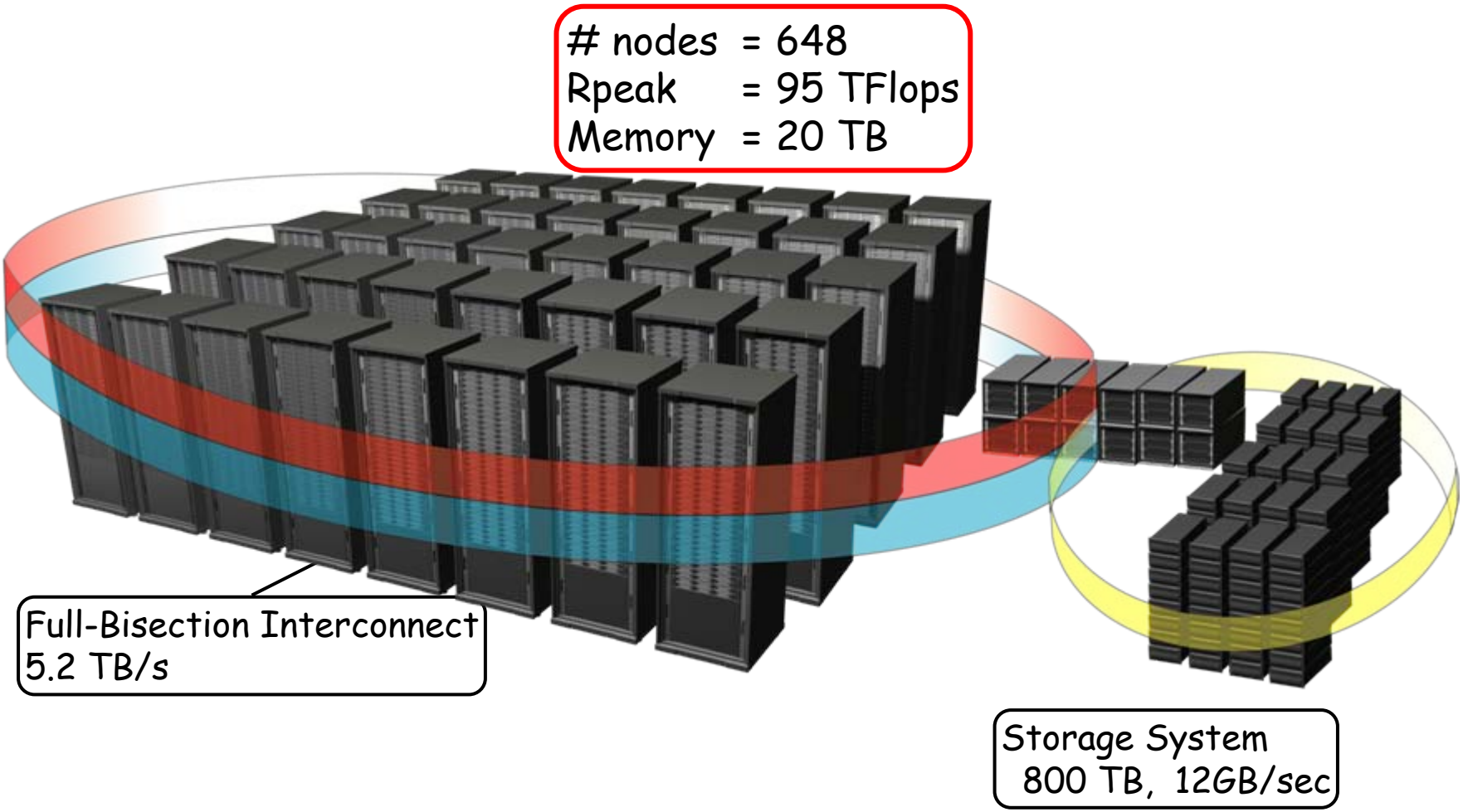
システム運用支援システム (ACE)

- ACE: Appro Cluster Engine
- 大規模クラスタの物理ノードを論理ノード群（複数の論理クラスタ）として管理・運用可能
  - PDU を用いたリモート電源管理
  - PXEを用いたネットワークブート
    ⇒OSイメージを論理クラスタ単位で変更可能
    （別バージョンOS、特定ライブラリ利用、...）
  - IP address の自動割り当て
- システム構成のダイナミックな変更に素早く対応可能
- Livermore, Sandia, Los Alamosを始めとする米国国研で利用（LLNL cluster, Tri-Lab Cluster）

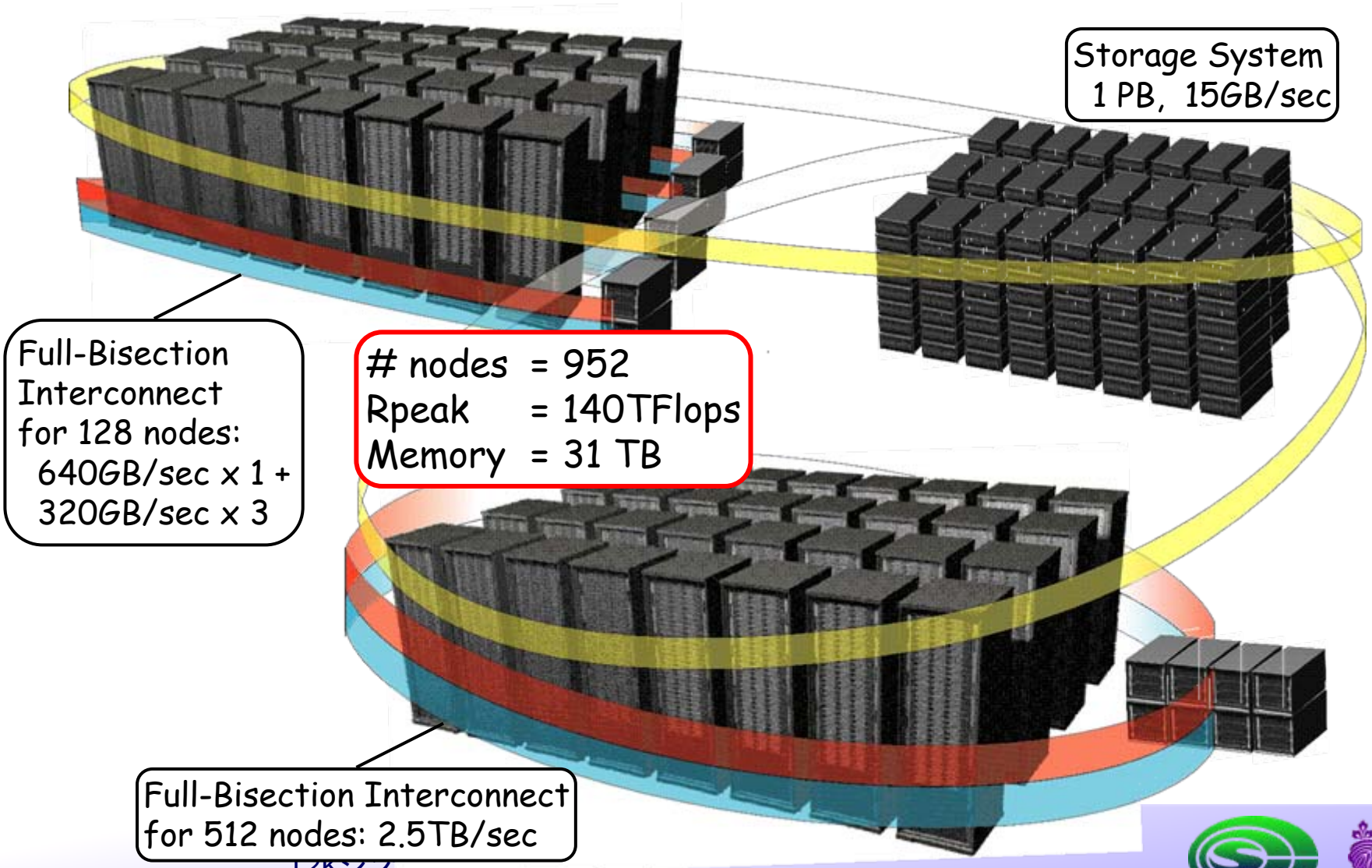
T2K Alliance下での計算機システム利用

- 3大学のシステムの一部を相互乗り入れ運用
  - 各大学の利用プログラムの共通する部分に適用
- グリッド技術を用いた single sign-on 利用
  - 1サイトでのログインにより他のサイトに自動ログイン／ジョブ投入可能
- 広域分散ファイルシステム Gfarm を用いた共有ファイルシステム
  - どのサイトでジョブを実行しても共通のファイル参照が可能
- T2K共通仕様によるオープンアーキテクチャ
  - これらの相互運用が自然な形で可能
  - 複数システム間の可搬性
    - プログラム可搬性：同一ソースの利用（再コンパイルは必要）
    - 性能可搬性：ほぼ同一の性能が達成可能

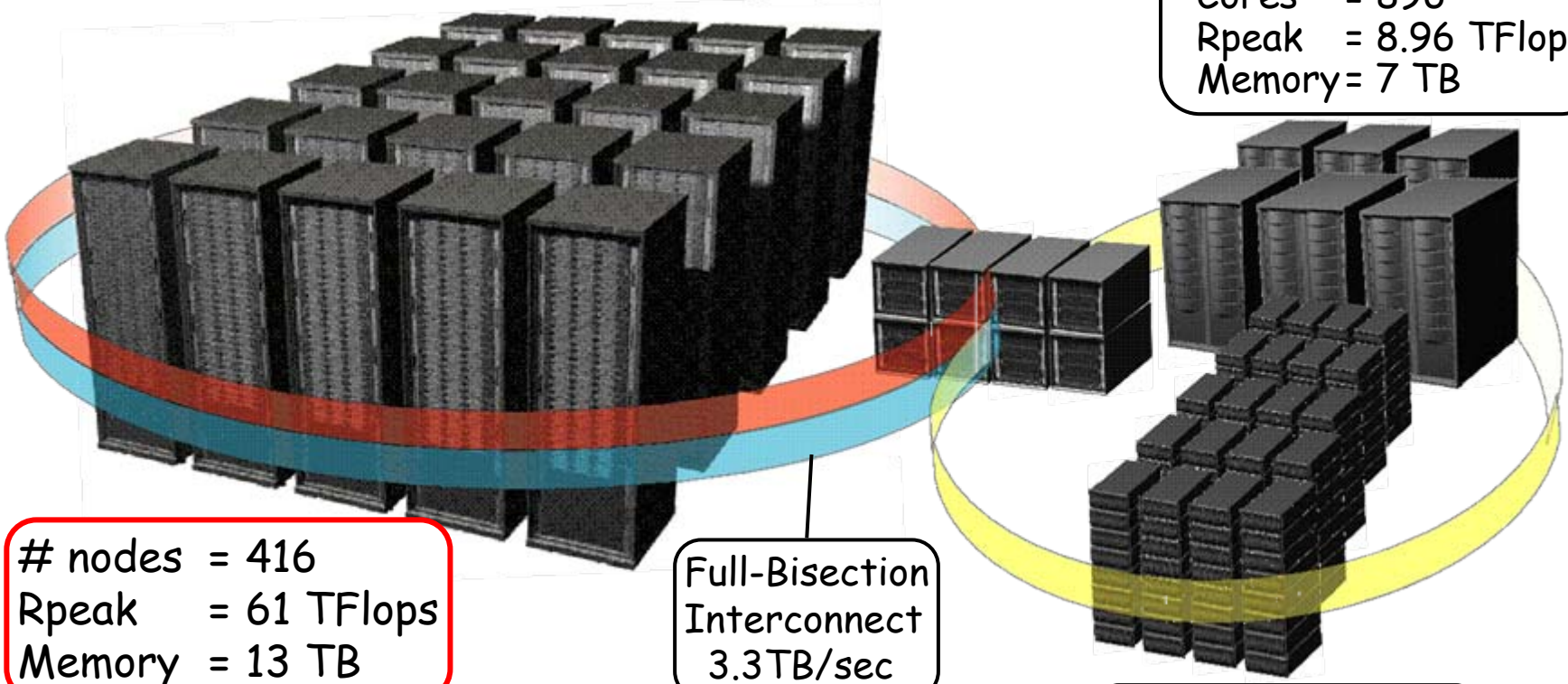
筑波大学構成 (Appro + Cray Japan)



T2K Open Supercomputer
東京大学構成（日立）



京都大学構成 （富士通）


Fat-node Subsystem
# nodes= 7
Cores = 896
Rpeak = 8.96 TFlops
Memory= 7 TB
# nodes = 416
Rpeak = 61 TFlops
Memory = 13 TB
Full-Bisection
Interconnect
3.3TB/sec
Storage System
883 TB, 16GB/sec

T2K筑波システムのスケジュール

- これまでの経過
  - H18.11月：導入説明・資料招請
  - H19. 3月：仕様書案説明・意見招請
  - H19.11月：入札説明
  - H19.12月25日：開札
- 今後のスケジュール
  - 現在、テストノードによるシステム評価、アプリケーション評価実施中
  - H20.4月：システム導入、Linpack等のテスト
  - H20.5月：システム調整、最終テスト、検収
  - H20.6月2日：システム運用開始

T2K筑波システムの利用プログラム

- 従来の筑波大スパコンは学内専用
- 次期スパコンからは計算科学研究センター(全国共同利用施設)が運用:現在はPACS-CSを共同利用
  ⇒次期スパコンについても全国共同利用を展開
- 複数の利用プラン(プログラム)を用意
  - 一般利用
  - 優先利用
  - 学際共同利用

T2K筑波システムの利用プログラム

- システム利用のモデルプラン（案）
  - 小規模・一般利用
    - 10万円／年程度で1/16 node～32 node程度を時間無制限（ただし他のユーザとの共有）により利用可能
  - 優先ユーザ
    - 100万円／8 node／年程度（大学価格）で、最優先度利用
  - 占有ユーザ
    - 160万円／8 node／年程度で、システムの一部を完全独占利用（OS入れ替え、追加アプリケーション導入も可）
  - 学際共同利用
    - 100万円／年程度で、大規模システム（数TFLOPS～数十TFLOPS相当）を利用
      ⇒計算科学推進プログラムとして、審査を経てプロジェクト採択

計算科学研究センター学際共同利用について

- 筑波大学計算科学研究センターにおいて実施
- これからのハイエンドコンピューティングにおいては
  - システム提供:「システムを作ったので適当に使って下さい」
  - アプリケーション:「私のアプリケーションが速く動くシステムを作って下さい」

という姿勢では駄目！

- これからは、システム側とアプリケーション側の研究者の共同作業が必要
  - 並列化、ベクトル化、cache-aware programming、アルゴリズム、数値解析、etc.
- さらに、異なる分野のアプリケーション同士の知識やノウハウの共有が必要になってくる
- これらの共同作業を実現する場としての枠組みを提供

つくばWANを介したT2K筑波システムの利用

- つくばWAN回線による高バンド幅通信
  - 筑波大次期スパコンをつくばWANの10Gbps回線に接続
  - ジョブ実行前後のファイル転送を超高速で実現
  - つくば地域ならではの共同利用を展開

- つくば地区特別プログラム
  - 先の価格プランは、「大学等教育機関価格」
    研究所／企業等の場合は、教育機関価格の倍程度の価格設定の予定

    ⇒つくば地区機関については、特別に教育機関並みの価格で提供

    ⇒つくば地区研究機関からの利用を推奨

おわりに

- T2K筑波システム
  - 95TFLOPSの超並列クラスタ型スーパーコンピュータ
  - 強力な計算ノードとノード間接続ネットワークにより幅広い階層・分野のアプリケーションを吸収
  - H20.6月より稼動開始
  - 全国共同利用、特につくば地区からの利用を推進
  - 学際共同利用を含む各種運用プログラム

- T2K Open Supercomputer Alliance
  - H20.6月稼動開始の筑波大・東大・京大のスーパーコンピュータのオープン仕様を共同策定
  - 計算機科学・工学の知識と経験を生かし、計算科学を始めとする広範囲のユーザに供するシステムを構築
  - グリッド技術に支えられた連携・相互運用