T2K オープンスパコン
新たな運用法と新たな連携

中島　浩

（京都大学学術情報メディアセンター）

special thanks to:

佐藤三久（筑波大），朴泰祐（筑波大），石川裕（東大）
岩下武史（京大），平野彰雄（京大）

T2K Open Supercomputer Alliance

目次

- T2K オープンスパコン・アライアンス
  - メンバー／目的／経緯
- T2K オープンスパコン
  - サイトの構成
  - 京大システムの構成
- 新たな運用法@京大
  - 課金方針／耐故障運用
- 新たな T2K 連携
  - シームレス高生産・高性能プログラミング環境
  - 学際計算科学推進

T2K オープンスパコン・アライアンス
同盟メンバー

京都大学 (Kyoto U.)
学術情報メディアセンター

筑波大学 (U. Tsukuba)
計算科学研究センター

東京大学 (U. Tokyo)
情報基盤センター

T2K Open Supercomputer Alliance

T2K オープンスパコン・アライアンス

同盟の目的

- 旧来型のスパコン調達から…
  - 受動的＆ベンダー主導
  - 製品市場からの選択型
- 新たな調達スキームに転換し…
  - 能動的＆大学主導
  - 技術市場からの創造型
- 高性能計算の新たなソリューション提供
  - 最先端技術に基づく設計
  - 幅広い大学ユーザへの対応

T2K オープンスパコン・アライアンス
結成～運用開始の経緯

- 06/05: T2K 最初の打合せ
- 06/09: 記者会見＆公開シンポジウム
- 06/10: 資料招請＆調達手続開始
- 07/03: 仕様書案公表
- 07/10: 最終仕様書公表＆入札手続開始
- 07/11: 応札締切
- 07/12: 開札
- 08/04: 連携協定締結（後述）
- 08/06: 納入＆運用開始（本格@K）
- 08/10: 本格運用開始@T2

目次

- T2K オープンスパコン・アライアンス
  - メンバー／目的／経緯
- T2K オープンスパコン
  - サイトの構成
  - 京大システムの構成
- 新たな運用法@京大
  - 課金方針／耐故障運用
- 新たな T2K 連携
  - シームレス高生産・高性能プログラミング環境
  - 学際計算科学推進

T2K オープンスパコン
サイトの構成

FUJITSU

416nodes=6656cores
61.2TFlops / 13TB
Rmax = 50.5TFlops
(w/416nodes, #34)

648nodes=10368cores
95.4TFlops / 20TB
Rmax = 76.5TFLops
(w/625nodes, #20)

APPRO

HITACHI
Inspire the Next

952nodes=15232cores
140.1TFlops / 31TB
Rmax = 83.0TFlops
(w/768nodes, #16)

T2K Open Supercomputer Alliance

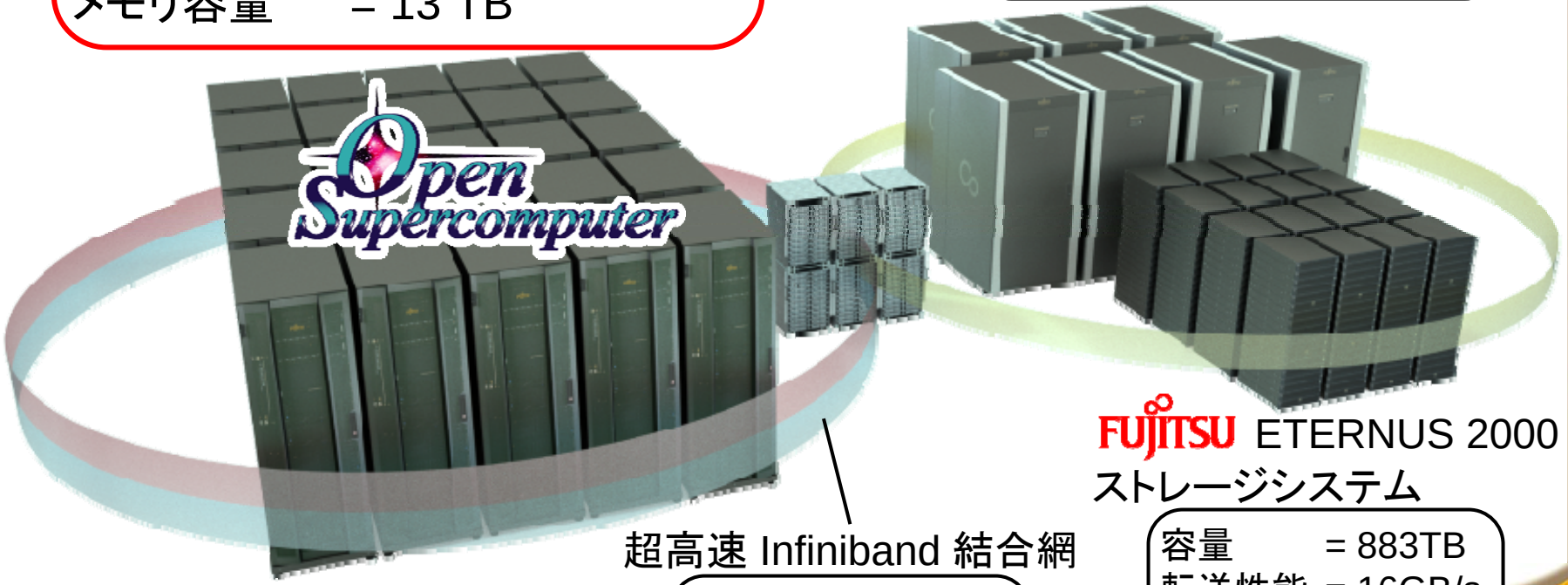
T2K オープンスパコン@京大
システム構成

FUJITSU HX600 クラスタ

ノード数　　= 416
コア数　　= 16 x 416 = 6656
ピーク性能　= 61.2 TFlops
Linpack 性能 = 50.5 TFlops(#34)
メモリ容量　= 13 TB

FUJITSU SPARC Enterprise M9000 fat node サブシステム

ノード数　　= 7
コア数　　= 128 x 7 = 896
ピーク性能　= 8.96 TFlops
メモリ容量　= 1TB x 7 = 7 TB



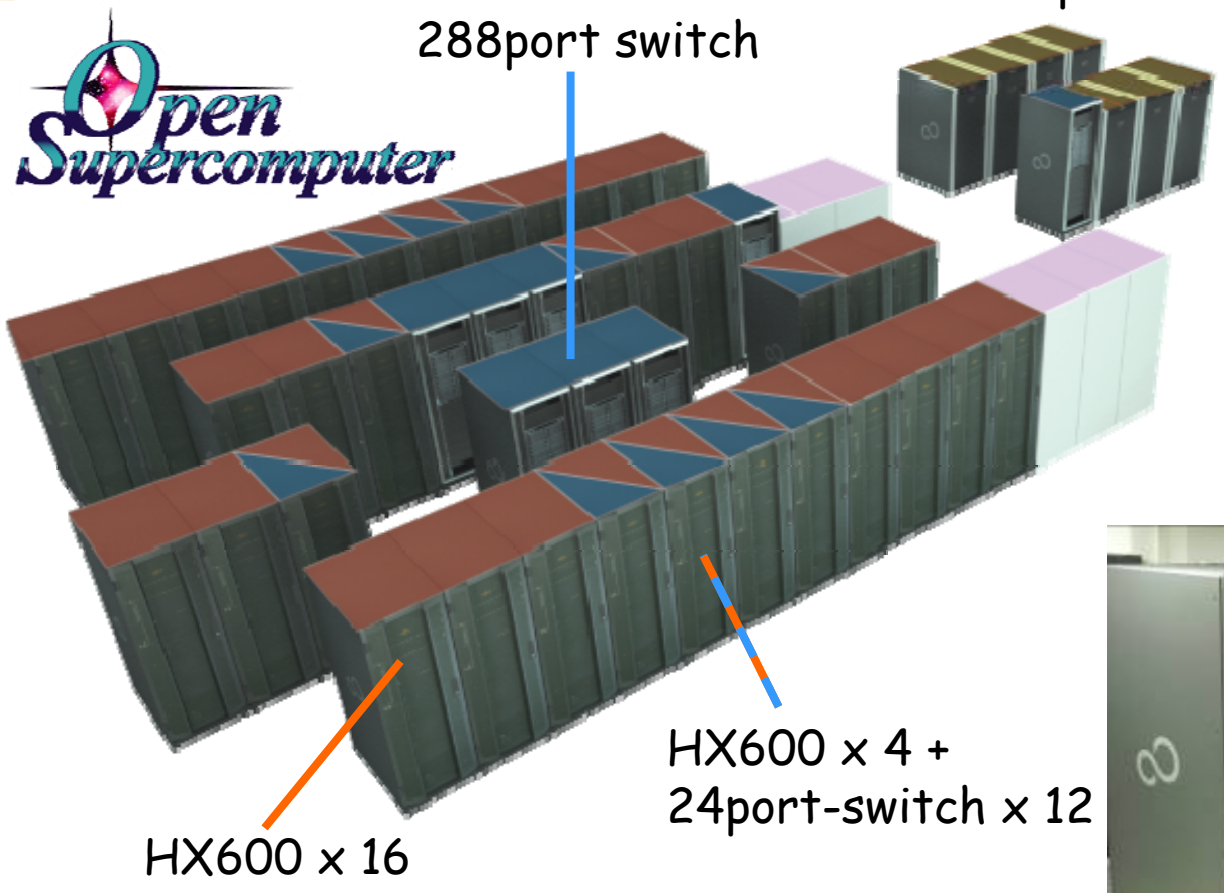
超高速 Infiniband 結合網
通信性能 = 3.3TB/s

FUJITSU ETERNUS 2000 ストレージシステム

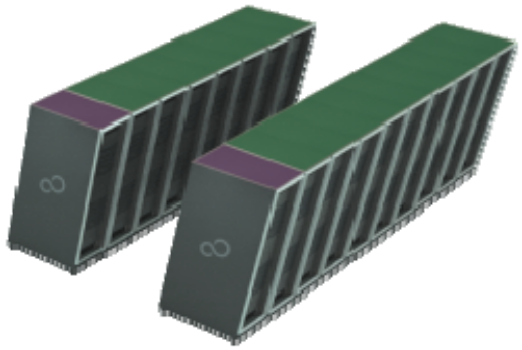
容量　　= 883TB
転送性能 = 16GB/s

T2K Open Supercomputer Alliance

T2K オープンスパコン@京大
システムレイアウト



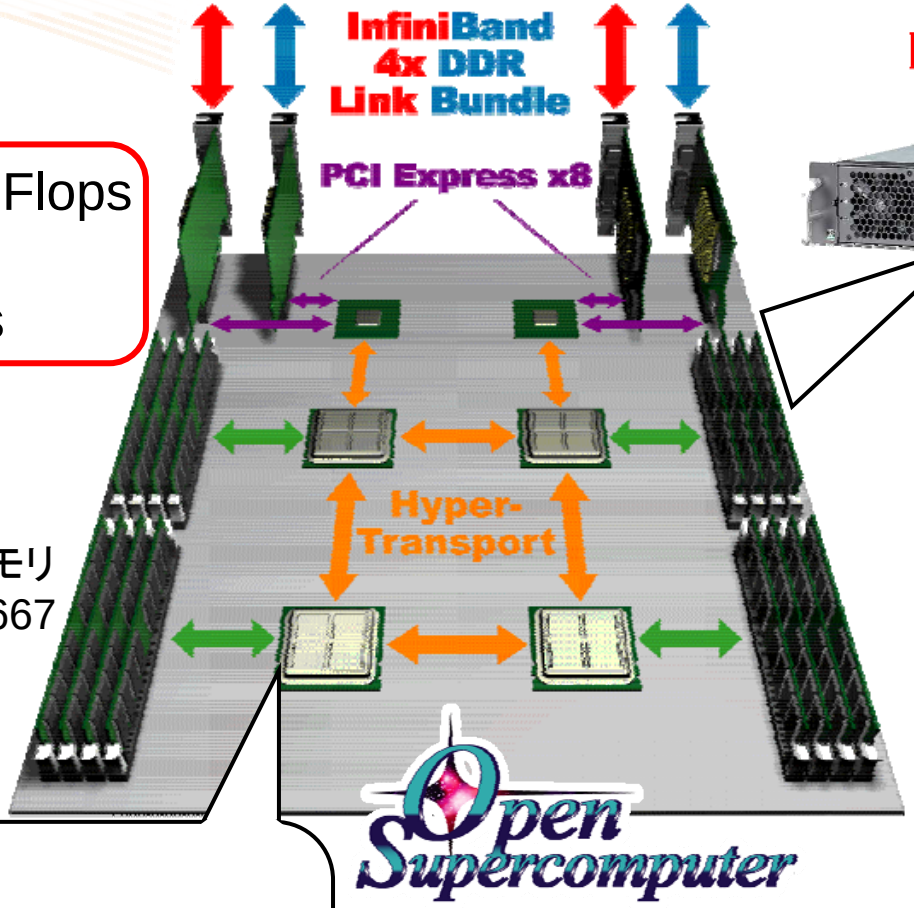
ETERNUS 2000



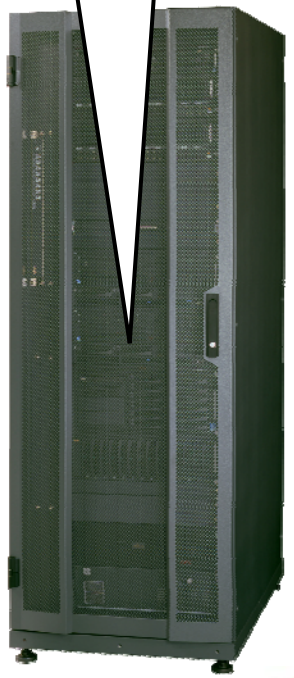
T2K Open Supercomputer Alliance

T2K オープンスパコン@京大
ノード構成

ピーク性能 = 147 GFlops
メモリ容量 = 32GB
通信性能 = 8GB/s

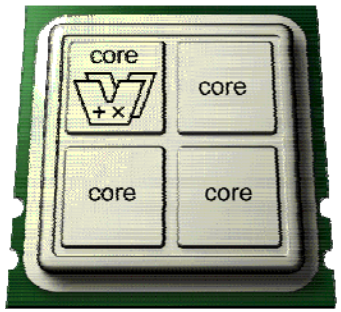


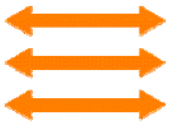
FUJITSU HX600



8GB メモリ
DDR2-667

AMD
Opteron Barcelona
36.8 GFlops



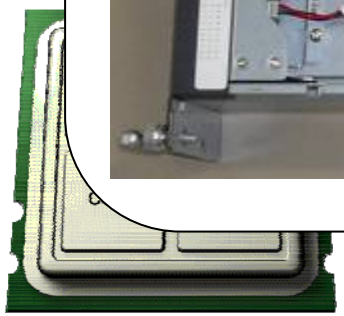


T2K Open Supercomputer Alliance

T2K オープンスパコン@京大
ノード構成

ピー
メモ
通信

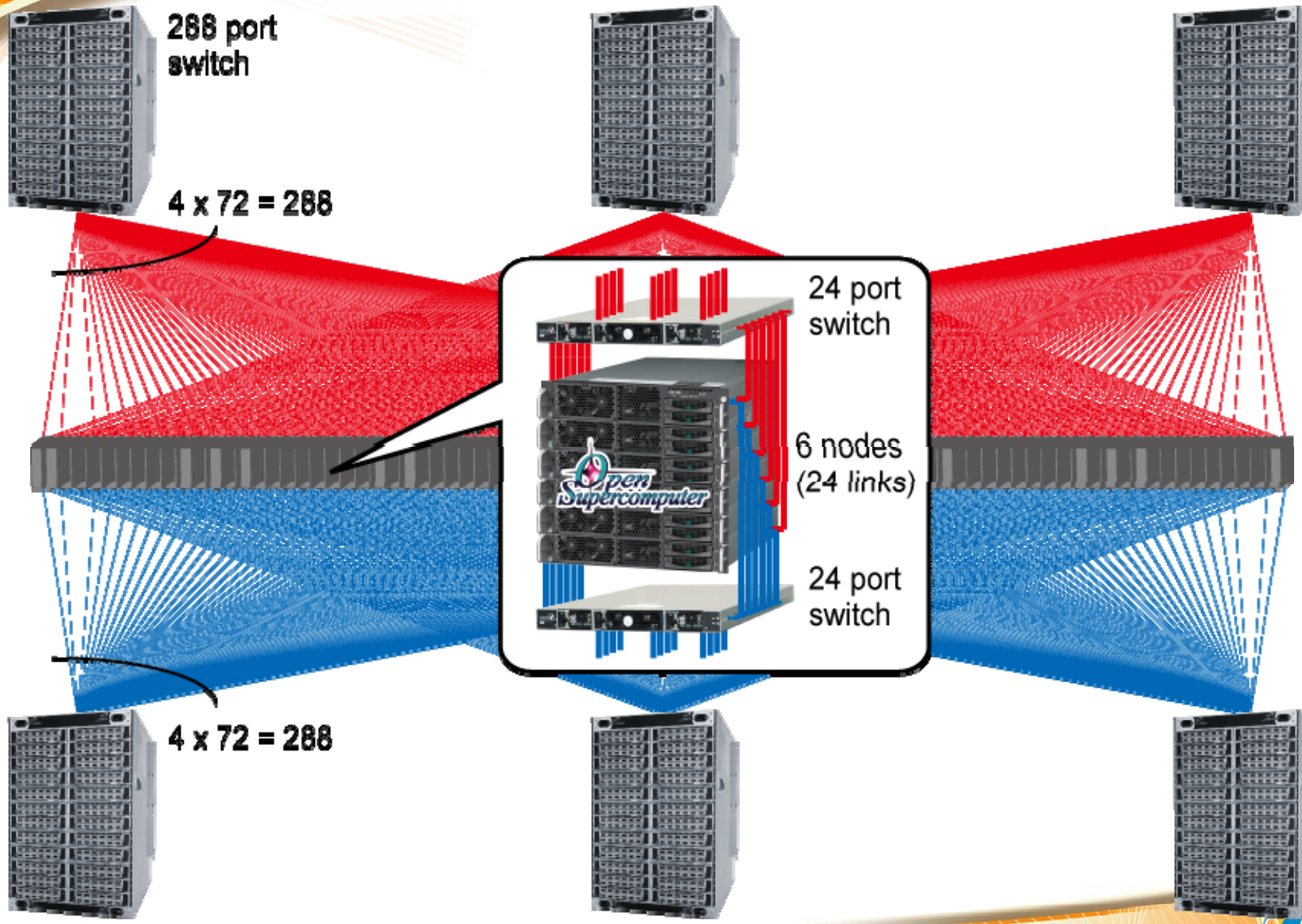
AM
Opte
36.8



Memory I/F

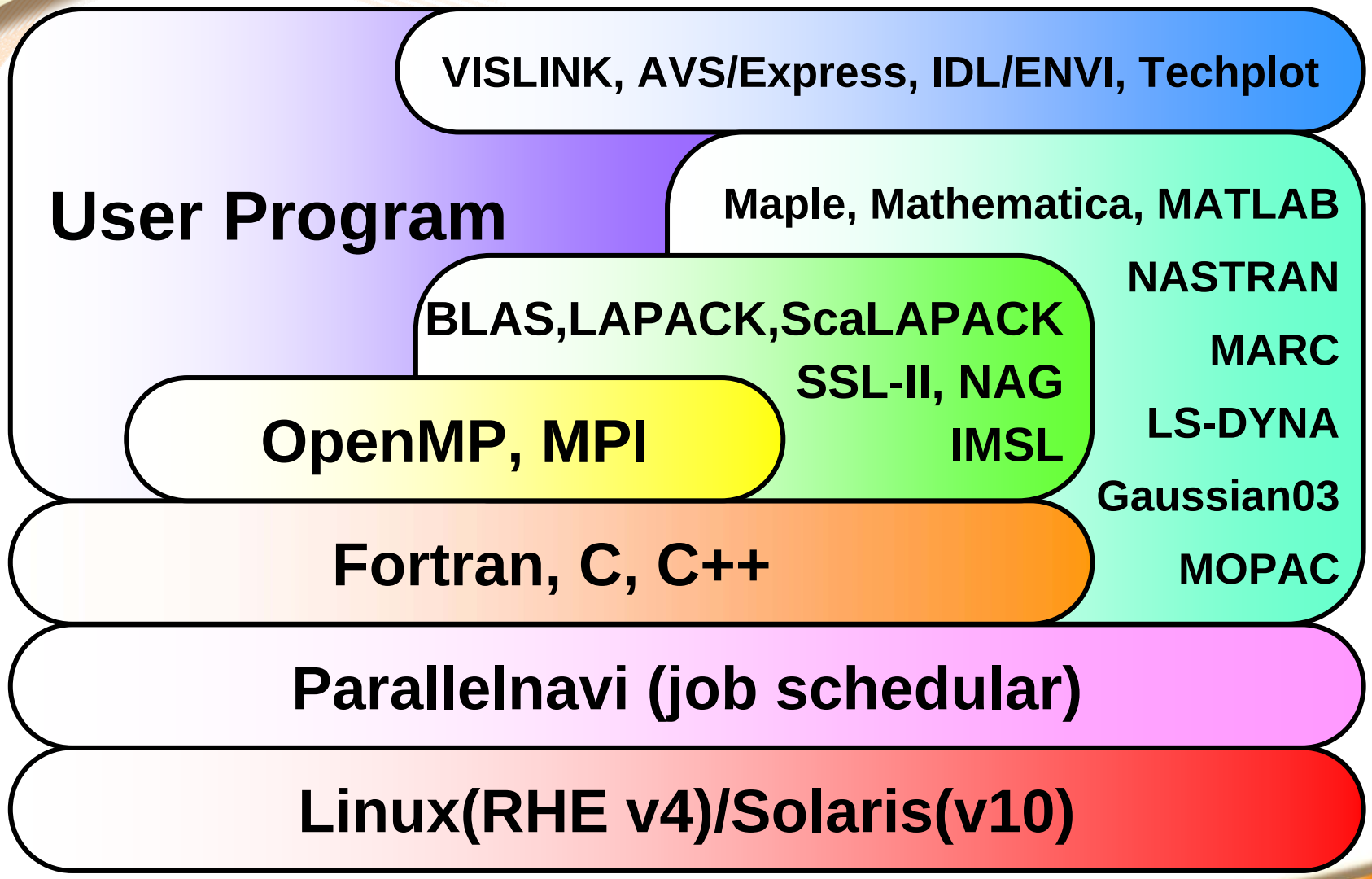
T2K Open Supercomputer Alliance

T2K オープンスパコン@京大
Infiniband 結合網



T2K Open Supercomputer Alliance

T2K オープンスパコン@京大
ソフトウェアスタック



T2K Open Supercomputer Alliance

目次

- T2K オープンスパコン・アライアンス
  - メンバー／目的／経緯
- T2K オープンスパコン
  - サイトの構成
  - 京大システムの構成
- **新たな運用法@京大**
  - **課金方針／耐故障運用**
- **新たな T2K 連携**
  - **シームレス高生産・高性能プログラミング環境**
  - **学際計算科学推進**

新たな運用法
課金方針 (1)

- 従来＝従量制
  - ￥0.1 × CPU時間(秒) × 並列係数(#CPU/4~16)
  - ￥10 × ディスク容量(GB) × 日
- 問題点
  - 仕事量 (≠計算量) に対する費用が読めない
    - 予算ショートを恐れて conservative に計算
    - aggressive に計算して結局予算ショート
  - 繁忙期 (11~2月) の進捗が読めない
    - ジョブ待ち時間は「他人の動向」が決定
    - 「全員が満足」する scheduling は存在しない

新たな運用法
課金方針 (2)

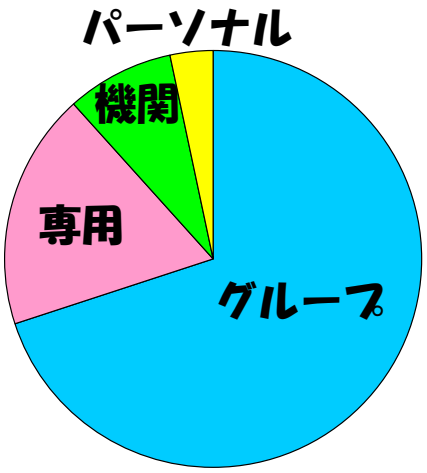
- 新方針＝資源買取制
  - 100万円/年＝8nodes＋8TB＋24users
  - グループメンバーが利用可能な資源量で調整
    - ≠物理的なノード割当
    - 通常稼動時 (incl. 繁忙期) は 8node を (ほぼ) 保証
    - 非通常時 (大規模ジョブ実行時) も 4node を保証
- 利点
  - 費用対効果 (利用可能資源) が読める
  - 繁忙期の計算進捗が読める
  - 「全員が納得する」scheduling を実現

新たな運用法
課金方針 (3)

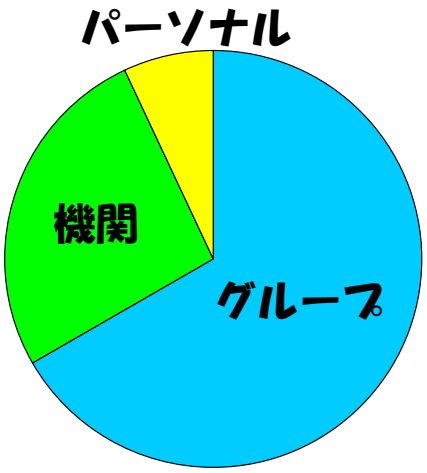
- メニュー
  - グループコース (標準的コース)
    - HX600：￥125K/node・年
    - M9000：￥100K/socket(4core)・年
  - 専用クラスタ (物理的占有)＝グループ x 1.5
  - 機関・部局定額 (ユーザ多数)＝グループ x 1.5
  - パーソナルコース＝￥100K 定額
    - n 人で ￥100K x n の資源を共有 (w/ fare-share)
  - 大規模ジョブ：週単位の大規模並列実行
    - HX600：￥6K/node・週
    - M9000：￥5K/socket・週

新たな運用法
課金方針 (4)

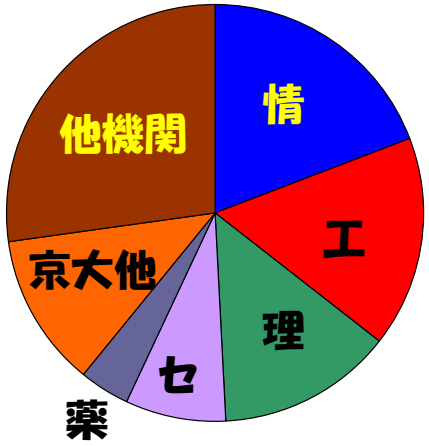
- 売れ行き



HX600



M9000



機関・部局別

新たな運用法
耐故障運用

- HX600: 1~2ノード/週の故障を前提
  - 代替待機ノード＝8 (4 hot ＋ 4 cold)
  - 故障発見時 (自発報告/生存監視) に自動交替
  - 自動チェックポイント＆復旧 (実施予定)
- IB 結合網
  - リンク故障　　：自動縮退
  - スイッチ故障 ：自動迂回
- ストレージ：RAID 6 + Backup
  - safe：shadow：risky
    ＝3：3：0/2：2：2/1：1：4/0：0：6

目次

- T2K オープンスパコン・アライアンス
  - メンバー／目的／経緯
- T2K オープンスパコン
  - サイトの構成
  - 京大システムの構成
- 新たな運用法@京大
  - 課金方針／耐故障運用
- **新たな T2K 連携**
  - **シームレス高生産・高性能プログラミング環境**
  - **学際計算科学推進**

T2K Open Supercomputer Alliance

新たな T2K 連携
T2K 連携協定

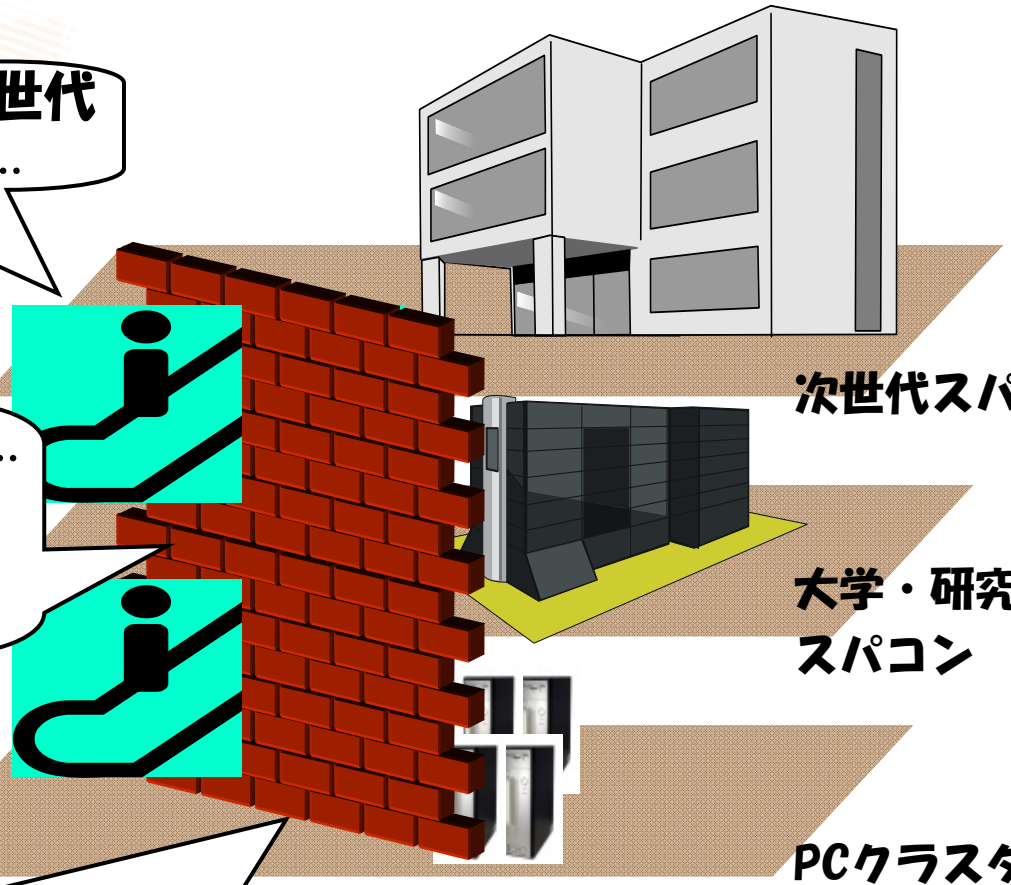
2008/4/1 締結

→計算＆計算機科学推進のための連携協力

- **正称**：筑波大学計算科学研究センター、東京大学情報基盤センター及び京都大学学術情報メディアセンターの間における連携・協力の推進に関する協定書
- **基盤技術共同研究**；e-science S/W 研究開発
  シームレス高生産・高性能プログラミング環境
  - 高性能並列プログラミング言語（筑波大）
  - 高生産並列スクリプト言語（京大）
  - 高効率・高可搬性ライブラリ（東大）
- 学際共同研究推進
- 人材育成・教育
- 研究支援体制整備

新たな T2K 連携
シームレス高生産・高性能…

研究室→センター→次世代
と step up したくても…



環境の違いがもたらす…
- 動作互換性の壁
- 性能互換性の壁
- プログラム難度の壁

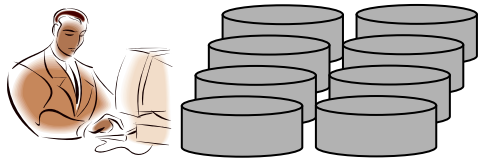
壁を打破するシームレスなプログラミング環境
- MPI から脱却する指示文ベース高性能並列言語
- お手軽並列処理を真にお手軽にするスクリプト言語
- 環境によらず使える高性能数値＆通信ライブラリ

T2K Open Supercomputer Alliance

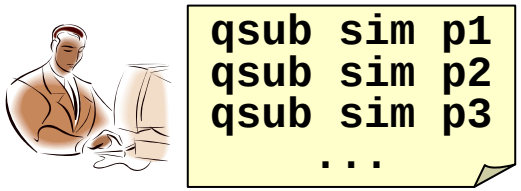
新たな T2K 連携
シームレス高生産・高性能…

- 高生産スクリプト言語@京大
  →PDCA cycle の生産性向上

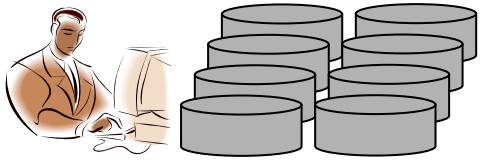
P: create huge size of input data



D: submit huge number of jobs



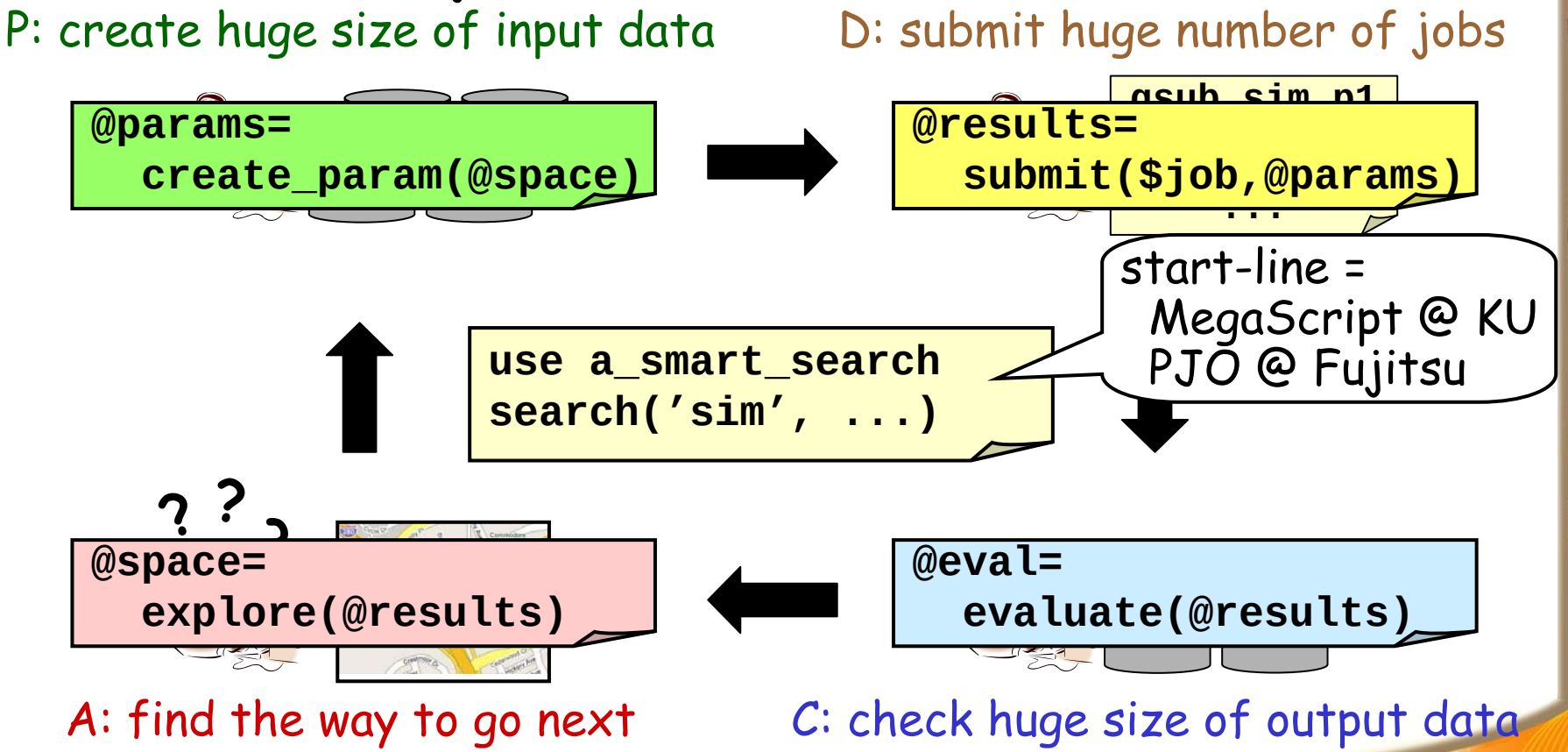
C: check huge size of output data



A: find the way to go next

新たな T2K 連携
シームレス高生産・高性能...

- 高生産スクリプト言語@京大
  →PDCA cycle の生産性向上



新たな T2K 連携
計算科学との連携・融合

T2K Open Supercomputer Alliance

学際計算科学推進委員会

研究推進

人材育成

筑波大学
計算科学研究センター

東京大学
情報基盤センター

ネットワーク連携
中核組織

京都大学
学術情報メディアセンター

支援体制整備

計算科学・工学

大学間コミュニティ
（気象、海洋、…）

大学＆研究所間
コミュニティ
（宇宙、素粒子、…）

X大学グループ
（計算物理）

Y大学グループ
（計算化学）

学際的
共同研究
推進・支援

学際的
計算科学
教育
プログラム

グリッド基盤

A大学センタ

B大学センタ

P大学グループ
（計算機科学）

数値解析
コミュニティ

スパコン計算基盤

新たな T2K 連携
計算科学との連携・融合

プログラム高度化支援事業
＝あなたのプログラムをタダで高度化・高速化

- H20 事業（経費≒3000万円）
  - 応募課題 = 海洋物理／地震予知／MD／プラズマ物理
  - チューニング／アルゴリズム改良／アルゴリズム開発
  - 内作＋外注で 8 課題を対象に実施
- 計算科学への貢献
  - 高性能プログラミングのツボを伝授
  - 自力（または自腹）での高度化・高速化のステップ
- 計算機科学への貢献
  - 研究開発成果のリアルな適用対象
  - 新たな研究ネタの発掘

まとめ

- 2.5 年の T2K 連携で得られたもの
  - 新たなコンセプトによるオープンスパコン仕様
  - 国内最高峰のスパコン×3
  - “Yes we could do it” の実績＆自信
  - “So we can go further” に基づく計画
- 今後の展開
  - やると決めたことの着実な実行
  - 次世代スパコン連携への展開
  - 次々世代(≠爺世代)への飛躍