研究開発ロードマップ検討状況 –システムソフトウェア–

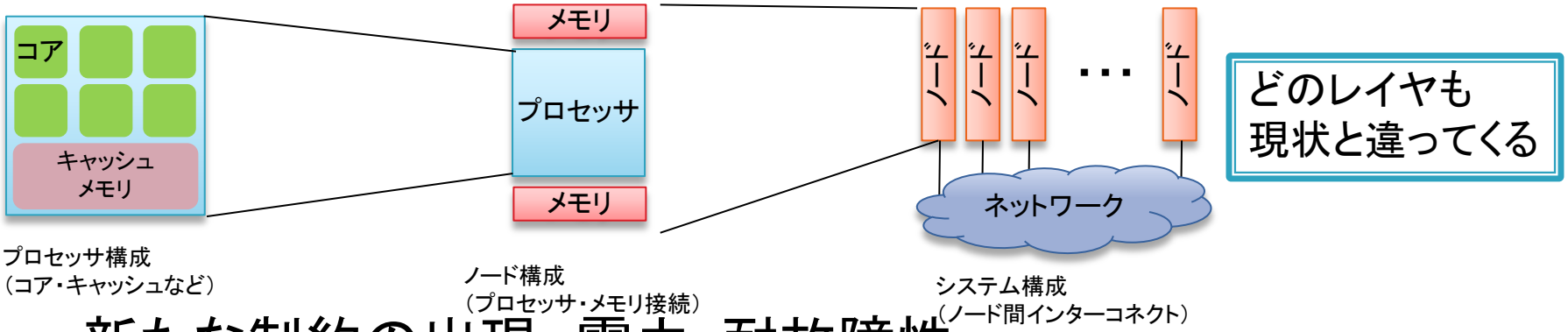
2011/12/08 @ PC Cluster Symposium

メンバー

- オペレーティングシステム
  - 清水 正明(日立), 高野 了成(産総研), 宇野 俊司(富士通), 松葉 浩也(日立)
- ランタイムシステム
  - 野村 哲弘(東工大), 今田 俊寛(理研AICS), 南里 豪志(九大)
- I/O
  - 建部 修見(筑波大), 佐藤 仁(東工大), 安井 隆(日立), 大野 善之(理研AICS)
- システム管理・外部連携
  - 中田 秀基(産総研), 竹房 あつ子(産総研), 遠藤 敏夫(東工大), 鴨志田 良和(東大), 滝澤 真一朗(東工大)
- 耐故障性
  - 實本 英之(東大)

背景

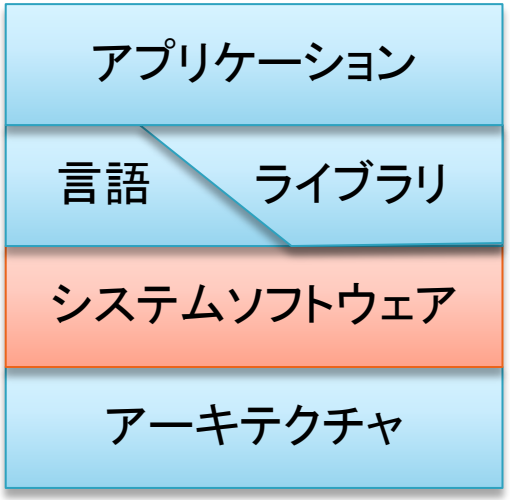
- Post-Peta/Exaへ向けてアーキテクチャが変化するのは必然
  - ヘテロ化、メモリ階層複雑化、通信のnon-uniform化
  - 電力バジェットが真に問題となる
  - コンポーネント数の増加→MTBFは5分???



- 新たな制約の出現: 電力・耐故障性
- 変化した土台と制約に合わせたシステムソフトウェアを開発しなければならない

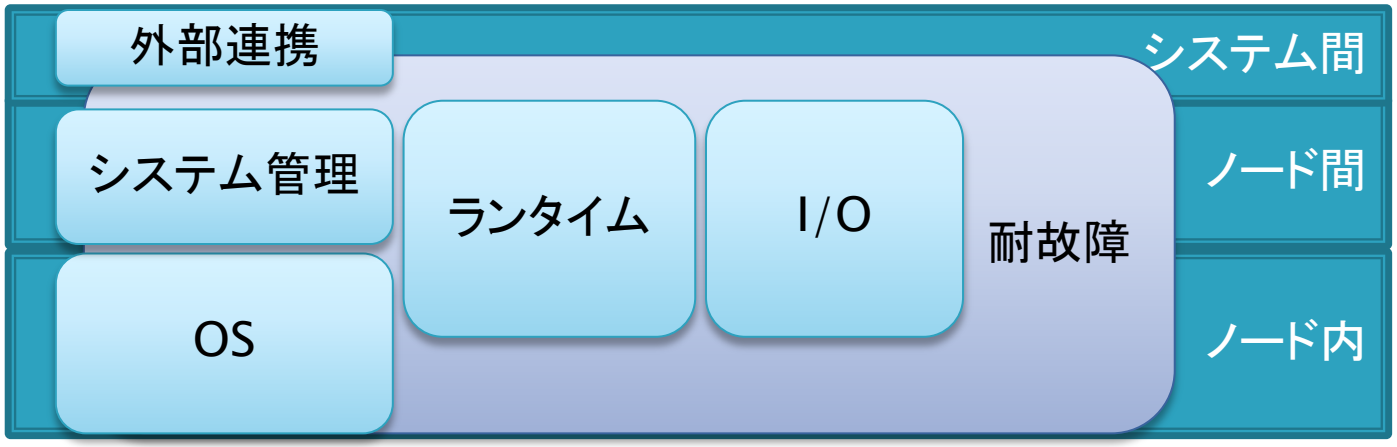
システムソフトウェアの役割

- アーキテクチャとプログラマ(言語・ライブラリ)を繋ぐ
  - ややこしいところは隠蔽する
  - 性能に影響するところは開示する
    - さらに言語やライブラリが隠蔽するかもしれない
- システムソフトウェアの進化で得られるもの
  - 新しいことが出来るようになる
    - 例: アクセラレータ透過な通信・デバッグ (電力制御)
  - 速くなる
    - 例: 通信の各種高速化手法
  - 変化しないとこのままでは動かなくなる
    - 例: 耐故障 (電力制御)
- システムソフトウェアの適用先
  - Exaスパコンだけではない



システムソフトウェアのアプローチ

- 以下のグループで研究項目を列挙



- 共通する前提
  - Scalabilityの確保
    - 例: ノード内でO(プロセス数) のメモリはもう使えない
  - 互換性・標準化
    - 効率化するために今までのものを切るときは…
      - 従来のものに対する互換レイヤを作る (生産性)
      - 独自のインタフェースを作るときはガラパゴス化しないよう国際協調して標準化
        - アプリ・ライブラリetcはどこまで許せる / ついてきてくれるのか?

オペレーティングシステムの研究課題

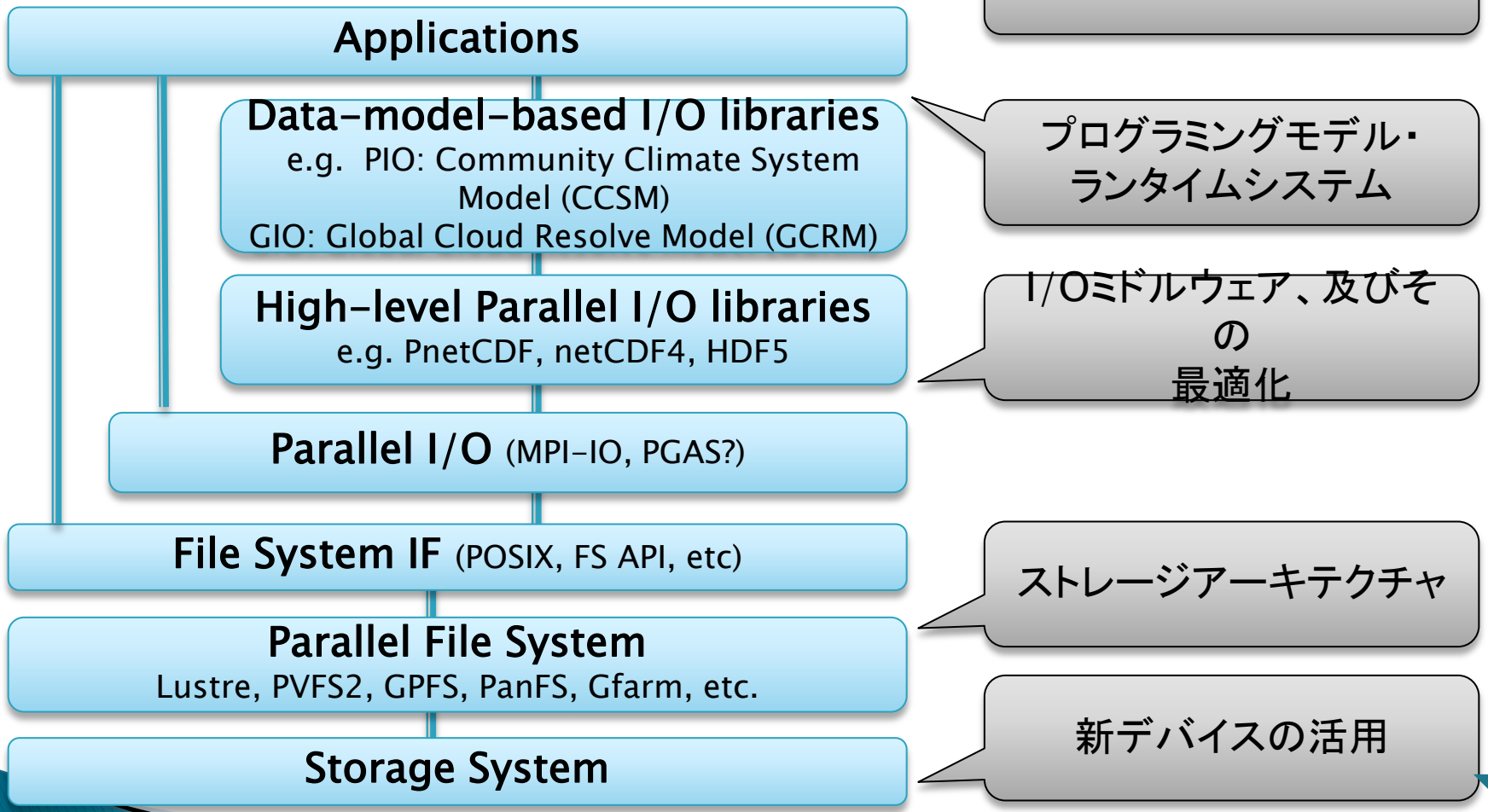
- ヘテロジニアスアーキテクチャへの対応
  - プロセス管理モデルの確立
  - 演算プロセッサ上のLightWeight OSの検討
    - POSIX+glibc対応をするか これらの制限を緩めて効率化と高スケーラビリティを果たすか (特にI/O ダイナミックリンク pthread…)
    - 演算プロセッサ向け簡素な静的メモリ管理
- 大規模並列: ~512プロセッサ/ノードへの対応
  - 割り込みによる外乱阻止、スレッド管理の効率化
  - DVFS、パワーゲーティング
- メモリアーキテクチャの変化への対応
  - メモリ階層の変化・深化に対応したAPIの構築

ランタイムの研究課題

- 大規模並列化・ヘテロ化への対応(通信フレームワーク)
  - 割り込みdriven通信、非同期通信バッファ
  - 通信専用コアの確保、アクセラレータ間直接通信
  - ネットワークトポロジの変化に対応した通信アルゴリズムの動的最適化
- 電力管理の強化
  - 細粒度電力制御API、電力データ取得APIの定義 →Archとの協業
  - 電力最適化ポリシーの構築
- 生産性の向上のために統一的なAPIを提供
  - 演算プロセッサ主体の通信手段(含エミュレーション)
  - アクセラレータ上でのデバッグ機(デバッガ・scalableコアダンプ?)
  - POSIX互換をOSが捨てた時の対応(互換性)
    - Emulation
    - 新たなインタフェースを作る&標準化する

I/Oの研究課題

IO Software Stack



全ての研究項目において、性能、信頼性、省電力を考慮

システム管理の研究課題

- 大規模並列化・ヘテロ化に対応したジョブ・タスク管理
  - Scalableなスケジューリング
    - 資源記述記法 ギャング ネットワーク データアフィニティ
      Bag of Tasks/密結合ジョブ型の区別
  - ワークフロー
- ロギング・モニタリング
  - 低負荷な情報収集

外部資源連携の研究課題

- I/O: データ配置の問題への対処
  - 外部ストレージ連携
    - オンデマンドネットワーク経路・帯域予約
    - 広域ワークフローエンジン

耐故障性に関する研究課題

- Fault Resilience Framework
  - 故障検知APIの規格化・情報伝播機構の効率化 → Archとの協業
  - Fault Resilientミドルウェア
    - システムソフトウェア自体の耐故障
      - RAIDおよびファイルのレプリケーション
  - 耐故障基礎技術
    - ヘテロ対応
    - モニタリング・ロギング手法（システム(HWやOS)とアプリ両面から）
      - 情報量爆発への対処
      - 低負荷な収集機構の開発
  - System-level(隠蔽型)耐故障の大規模化と効率化

まとめ

- Exascaleに向けたシステムソフトウェアに関する課題
  - 電力
  - 故障頻度(部品数)
  - ヘテロ化 …
- システムソフトウェアに求められる役割
- システムソフトウェアにおける研究項目
  - オペレーティングシステム
  - ランタイム
  - I/O
  - システム管理
  - 外部資源連携
  - 耐故障技術