# System Softwareサブグループ分担

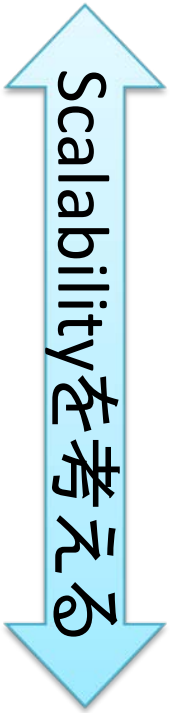
- Operating Systems
  - 清水、宇野、高野
- Runtime Systems
  - 實本、今田、野村、南里、鈴村、三浦
- I/O Systems
  - 佐藤、山本、大野、建部、安井
- Systems management、External environments
  - 遠藤、鴨志田、滝澤、竹房

# Operating Systems

- HPCに必要なOSとは？
  - 抽象化の範囲
    - 単一 or 分散OS
  - Right-weight OS
    - メモリフットプリント（OSの軽量化？）
  - 新サービス、API
- OSのスケーラビリティ
- OSジッタ
  - コアの占有化、ハードウェアマルチスレッディング
- ヘテロOS
- 耐故障性
  - VM、micro reboot、checkpoint
- 省電力
  - パワーゲーティングなど、ハードウェア・コンパイラとの連携
- 新しいメモリ階層（不揮発メモリ、フラッシュストレージ）
- メモリバイパス通信

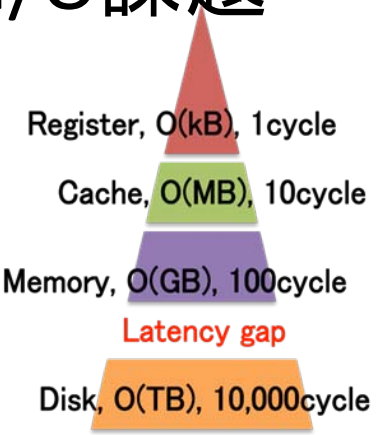
# Runtime System

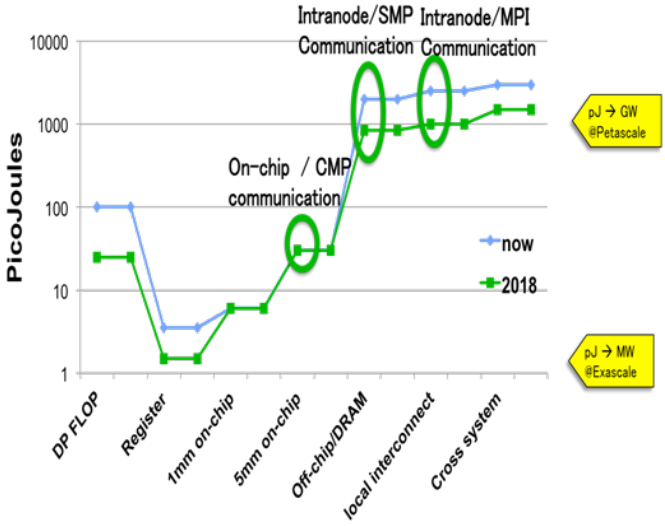
- Survey: IESPのロードマップの項目+αが分類としてはよさそう
  - Asynchrony/overlap (今田 野村 三浦)
    - User or MPIに見える部分
    - 足回りとしての部分
  - Heterogeneity (GPU/IntelのMany core) (同上)
  - Runtime Optimization
    - たとえばCollective, 通信以外も (南里)
  - Fault Tolerance/Resilience (實本+外から巻き込む)
  - 通信以外の要素(Profiling, Process management)
  - MPIコアの実装におけるScalability or MPI以外の方法
    - PGASの例, BlueGeneの例 (鈴村)
- 上下のレイヤが出てこないと動けない
  - Co-designにこちらから巻き込みに行く必要がある



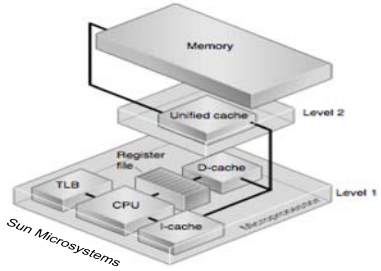
# エクサスケールスパコンに向けて解決すべきI/O課題

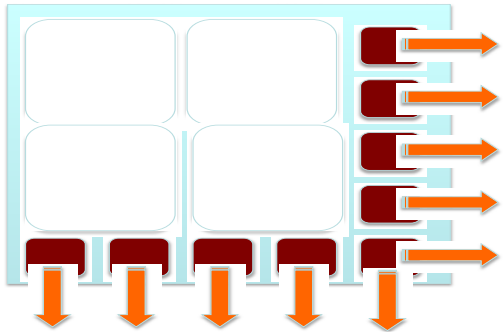
- ストレージ階層の深化
  - HDD : スピンドル制限
    - 多数デバイスを集約して利用
  - NVRAM/SSD : 高速だが高コスト
    - 容量増は見込めない
  - 共有FS, 分散FS
    - POSIXのコンシステンシモデルが困難
    - 細粒度，ランダムメタデータアクセス集中問題




- データの移動
  - 性能の低下，消費電力の増大
    - 特に，On-chip, Off-chip間のデータ転送が致命的
  - 局所性
    - ストレージの階層性(垂直)
    - 多数スレッドによるコンカレントなアクセス (水平)
    - ノードのスクラッチストレージの活用



# サーベイ分担

- 分散・共有ファイルシステム
  - 実装方式一般・既存FSの問題点
    - 山本, 建部
  - データアーカイブ
    - 建部
- データ処理ミドルウェア
  - I/O最適化, インターフェース(net CDF、HDF5、MPI IOなどのファイルAPI)、処理系
    - 大野,
- I/Oワークロード
  - 細粒度I/O、ランダムアクセス、メタデータ操作
    - 佐藤

# System Management + External System

- Resource management + Scheduling
  - 100万ノードクラスのscalability
  - VMスケジュールを含む
  - サイト間スケジュールを含む
  - ネットワーク経路・パス割当てを含む

竹房 + 遠藤

- Monitoring + logging
  - 100万ノードクラスのscalability
  - 電力、故障、性能異常
  - 解析、及びResource Managementへのフィードバック

遠藤 + 鴨志田

- Remote system integration (Data + Comp.)
  - 認証・ネットワークシステムアーキテクチャ
  - 広域ネットワークにおけるパス・帯域確保技術
  - プロトコル（標準化？）

滝澤 + 竹房

- Security（主にスケーラビリティ?）

不在

# サブグループ横断項目

- スケーラビリティ
- フォールトトレランス
- 省電力
- ・・・

# 次回までのスケジュール

- ML、Wiki作成（遠藤）
- 2週間単位でサブグループ毎にML報告
- 6月末：全体打ち合わせ
  - ざっくりした技術俯瞰、ロードマップ
  - サブグループ横断の検討項目
- 7月中旬：全体打ち合わせ
  - パワポベースのまとめ作成

- サーベイ項目洗い出し
- 論文リスト
- 技術俯瞰
- ロードマップ

- 仮定するアーキテクチャは？
  - 並列度（1～10億のオーダ）
    - コア数の増大 e.g. BlueGene
    - ファットノード
      - 電力的に厳しい？
  - 階層化並列
    - MPI、OpenMP、SSE
    - フラットMPIでどこまで行けるのか？
- OSがカバーすべきなのは単一ノード
- ゴール

- No MPIアプローチは？

# Runtime Systems

- MPI
  - 京で8万 rank
- MPIの大規模化の問題点
  - EuroMPIなどの議論を参照
    - Communicatorの巨大化（メモリ利用）
    - Comm_splitなどglobal operationの負荷

# Runtime systems

- Survey: IESPのロードマップの項目+αが分類としてはよさそう
  - Asynchrony/overlap (今田 野村 三浦)
    - User or MPIに見える部分
    - 足回りとしての部分
  - Heterogeneity (GPU/Intelのやつ) (同上)
  - Runtime Optimization(たとえばCollective) (南里)
  - Fault Tolerant (實本+外から巻き込む)
  - 足りないところは?
    - 通信以外の項目
  - Scalability (鈴村)
- 上下のレイヤが出てこないと動けない
  - Co-designに巻き込みに行く必要がある