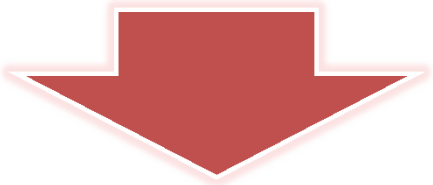
先端的計算科学研究プロジェクト

研究課題名：大規模スーパーコンピュータにおける電力資源管理システムのスケーラビリティ評価

近藤 正章, 坂本 龍一
(東京大学 情報理工学系研究科)

研究の背景

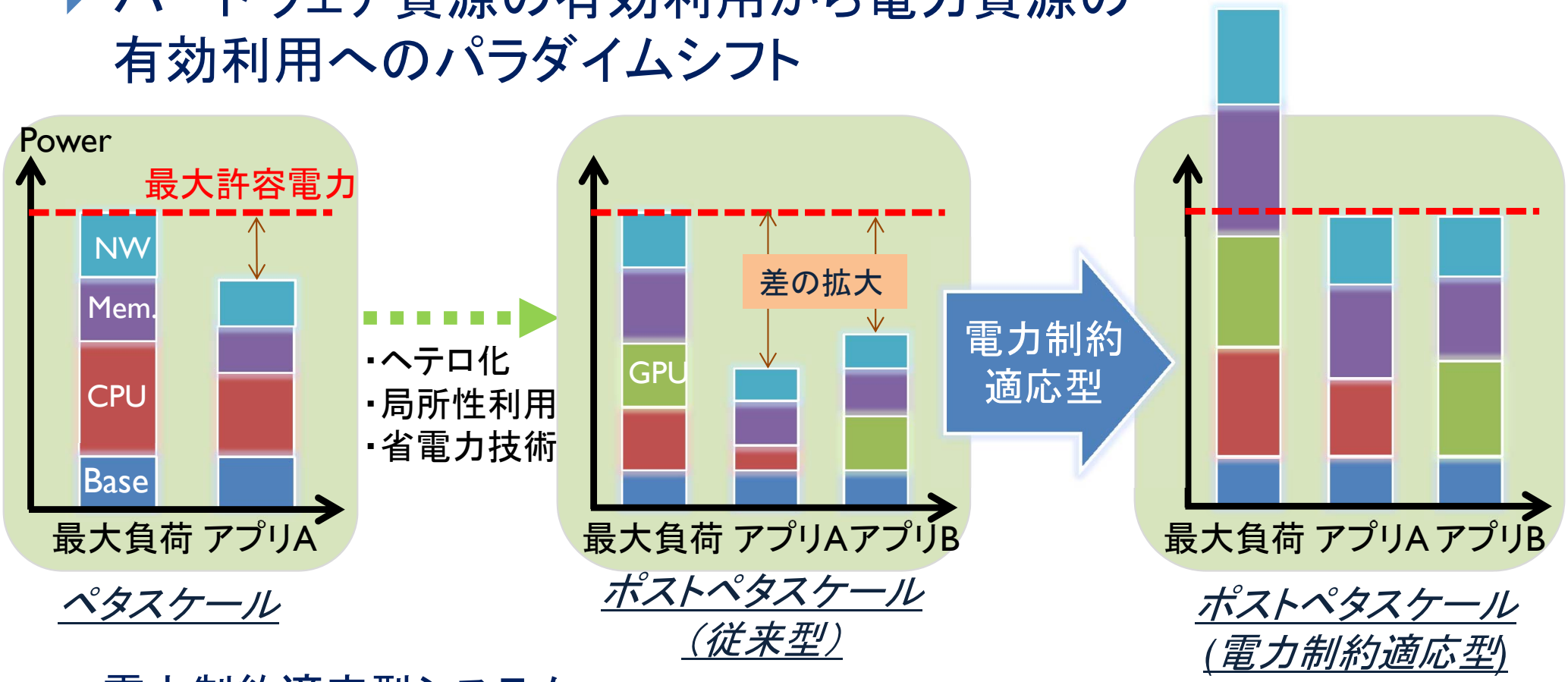
- ポストペタ時代のシステムは消費電力が最大の設計制約
  - 京コンピュータでは10PFLOPSを13MWで実現
  - 将来的にも20～40MWが電力供給の限度
  - エクサシステムでは同程度の電力で100倍の性能向上が必要
- アプリケーションのシステムへの要求の多様化
  - 計算・記憶・通信の各要素への要求が異なる
  - 電力がシステム制約となる状況下では各要素へ投入するハードウェア資源は制限せざるを得ない



運用時のピーク電力が制約を超えないことを保証する
worst case設計ではシステムをスケールさせることは難しい

ポストペタスケールシステムのあるべき姿

- ハードウェア資源の有効利用から電力資源の有効利用へのパラダイムシフト



ペタスケール

ポストペタスケール（従来型）

ポストペタスケール（電力制約適応型）

電力制約適応型システム

- 最大負荷時電力が電力制約を超過することを積極的に許容
- 電力性能ノブを制御することで実効電力を制約以下に抑制
- 電力資源を計算・記憶・通信へ適応的に配分することで実効性能向上へ

電力資源配分の最適化

- ジョブの特徴に応じた電力資源配分最適化に必要な技術

ジョブ内の電力資源配分

- プロセッサ・DRAM間
- プロセッサ・ネットワーク間
- ジョブ内のノード(MPIランク)間

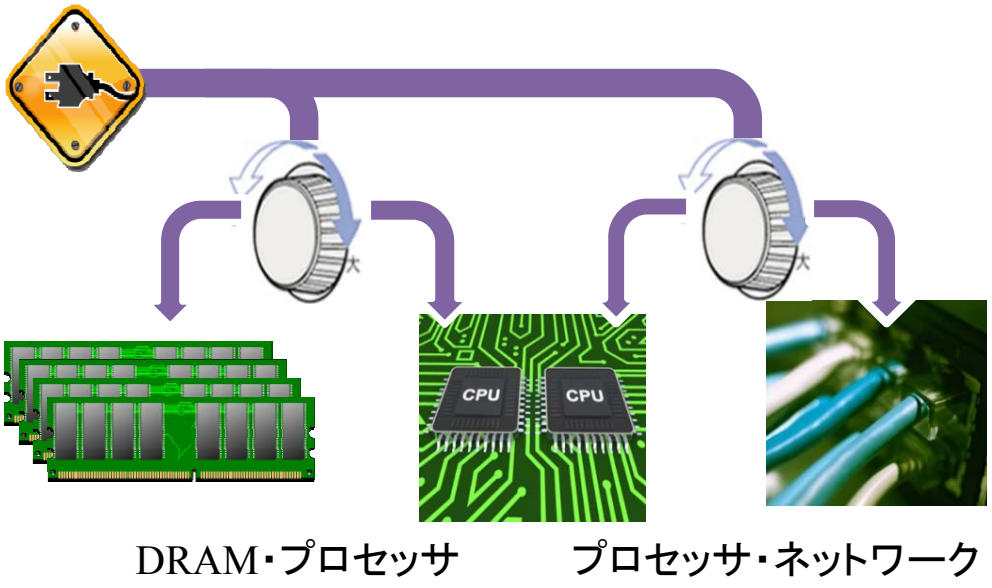
→ 最適化アルゴリズムとその自動化

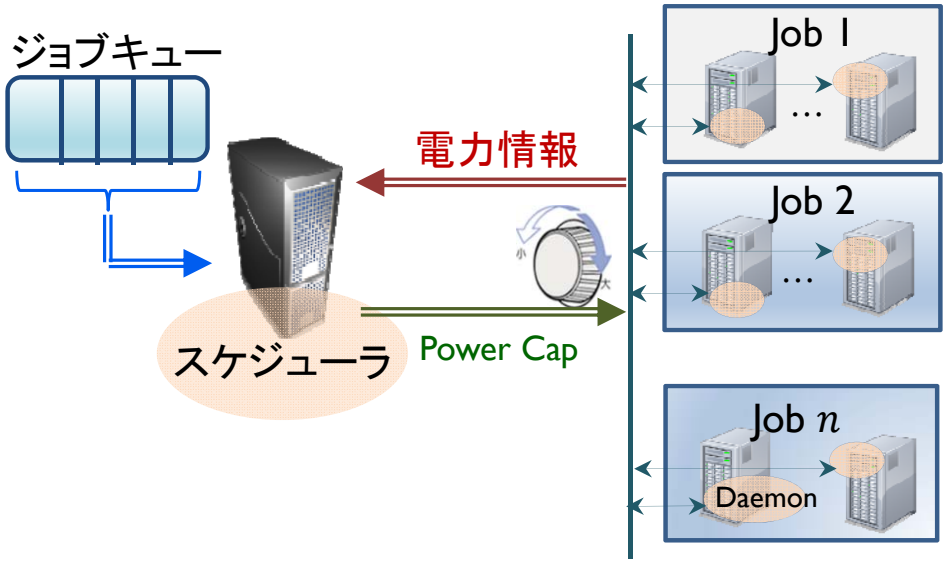
ジョブ間の電力資源配分

- 各ジョブへの電力バジェット配分
- キューからの最適実行ジョブ選択
- 各ジョブの動的な電力監視と制御

→ 電力を考慮したスケジューリング

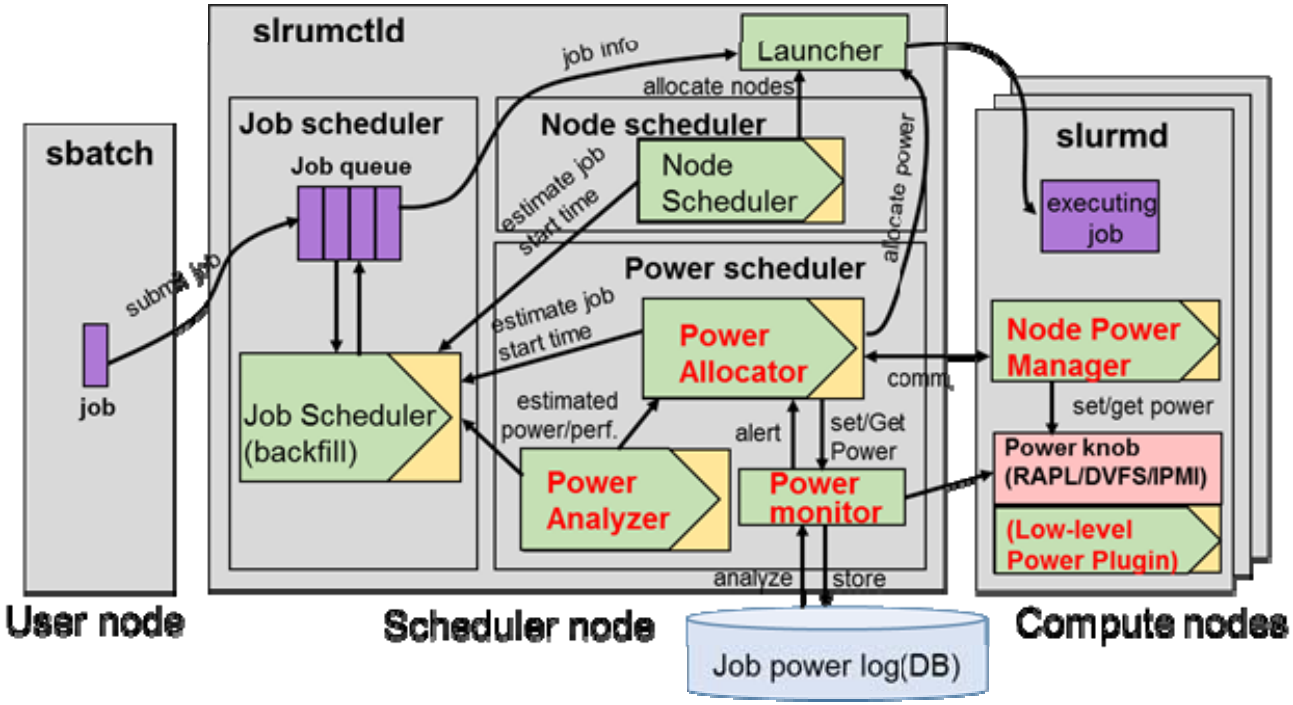
→ ジョブ特徴に応じた電力配分最適化





電力制約適応型資源管理ソフトウェアの開発(※)

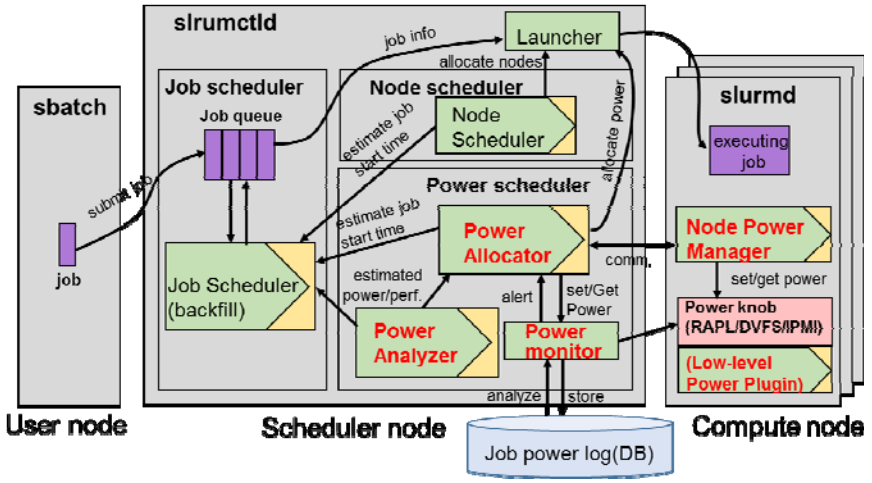
- 合計電力資源に基づき各ジョブの消費電力を管理
  - リソースマネージャからの電力制約設定と管理
  - 合計電力制約を基にしたジョブスケジューリング
  - 計算ノードでの電力モニタリングとキャップ設定
  - Slurm Workload Managerをベースに開発



※JST CREST「ポストペタスケールシステムのための電力マネージメントフレームワークの開発」で実施

開発資源管理ソフトウェアの機能と特徴

- 機能と特徴
  - 電力管理とジョブスケジューリング
    - 電力制約内でのスケジューリングとbackfilling
    - ノード内での動的電力管理・スケジューラ連携
  - Pluginベースのインタフェース(API)提供
    - 種々の電力制御アルゴリズムを実装可能に
- 主要なSlumへの追加モジュール
  - *Power Analyzer*
    - ジョブの必要電力や制約下での性能(実行時間)を予測
  - *Power Allocator*
    - ジョブの最適な電力割り当て決定、job schedulerが連携してジョブスケジューリング
    - 動的にシステムの電力を管理(エマージェンシー対応)
  - *Power Monitor*
    - 各ノードの電力収集と電力制約の設定
  - *Node Power Manager*
    - 各ノードで動作し実際に電力のモニタリングや電力ノブの制御を行う



HA8000システムへのインストール

- 開発したスケジューラを九州大学情報基盤研究開発センターHA8000システムへ導入(※)
- HA8000の概要
  - CPU: Xeon E5-2670 v2 (2.5GHz), メモリ128GB/ノード、965ノード
  - 電力制約設定：RAPL
  - 電力モニタリング：RAPL & IPMI
- 開発スケジューラ運用方針
  - ログインノードにてSlurmサーバを実行
  - 占有の32計算ノードではデーモン(slurmd)の起動が常に可能
  - 全系利用時に全計算ノードのデーモン起動許可を頂くことで実験

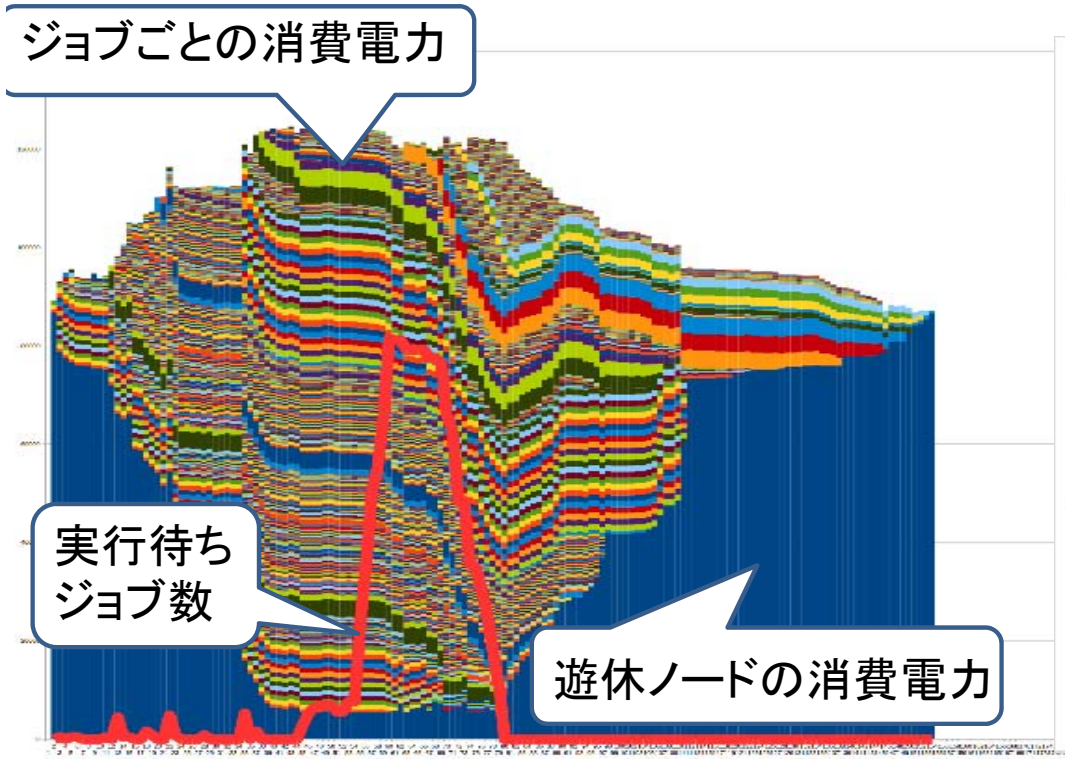
  → 全系で電力を考慮したジョブスケジューリング評価が可能に

（※センターの皆様にはご協力頂き感謝申し上げます）

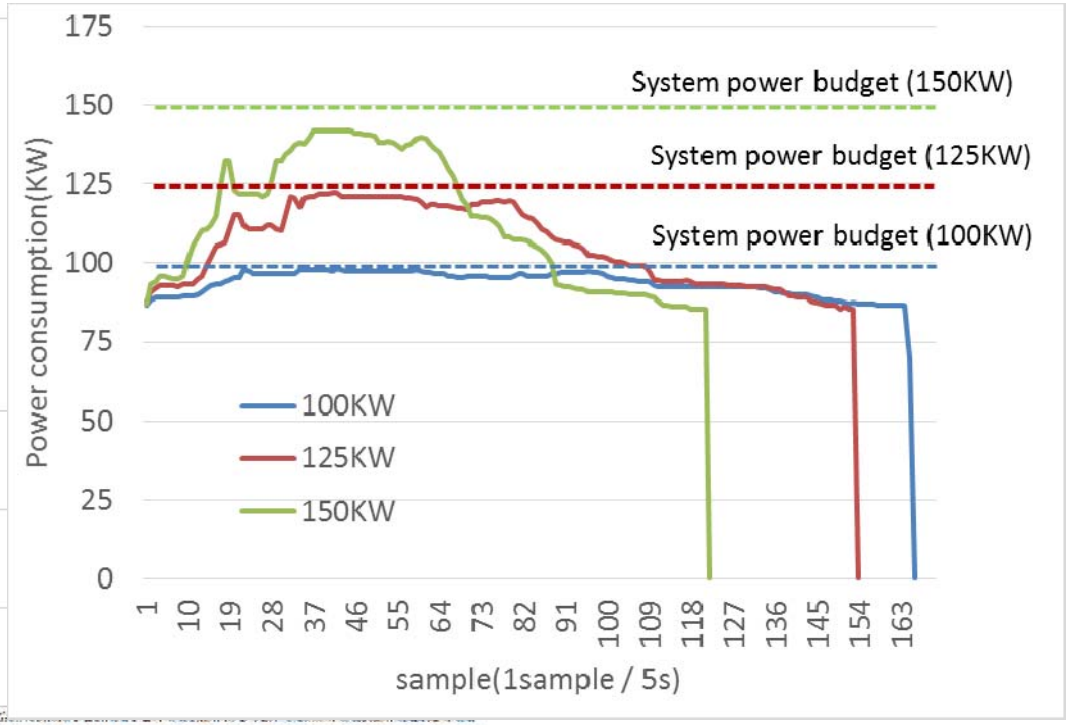
大規模システムでの機能検証と評価

- HA8000システムでの機能検証
  - 全体システム電力制約内でのジョブスケジューリング機能
  - ジョブスクリプトへの電力制約設定機能
  - 各計算ノードでの電力モニタリング・電力制約設定機能

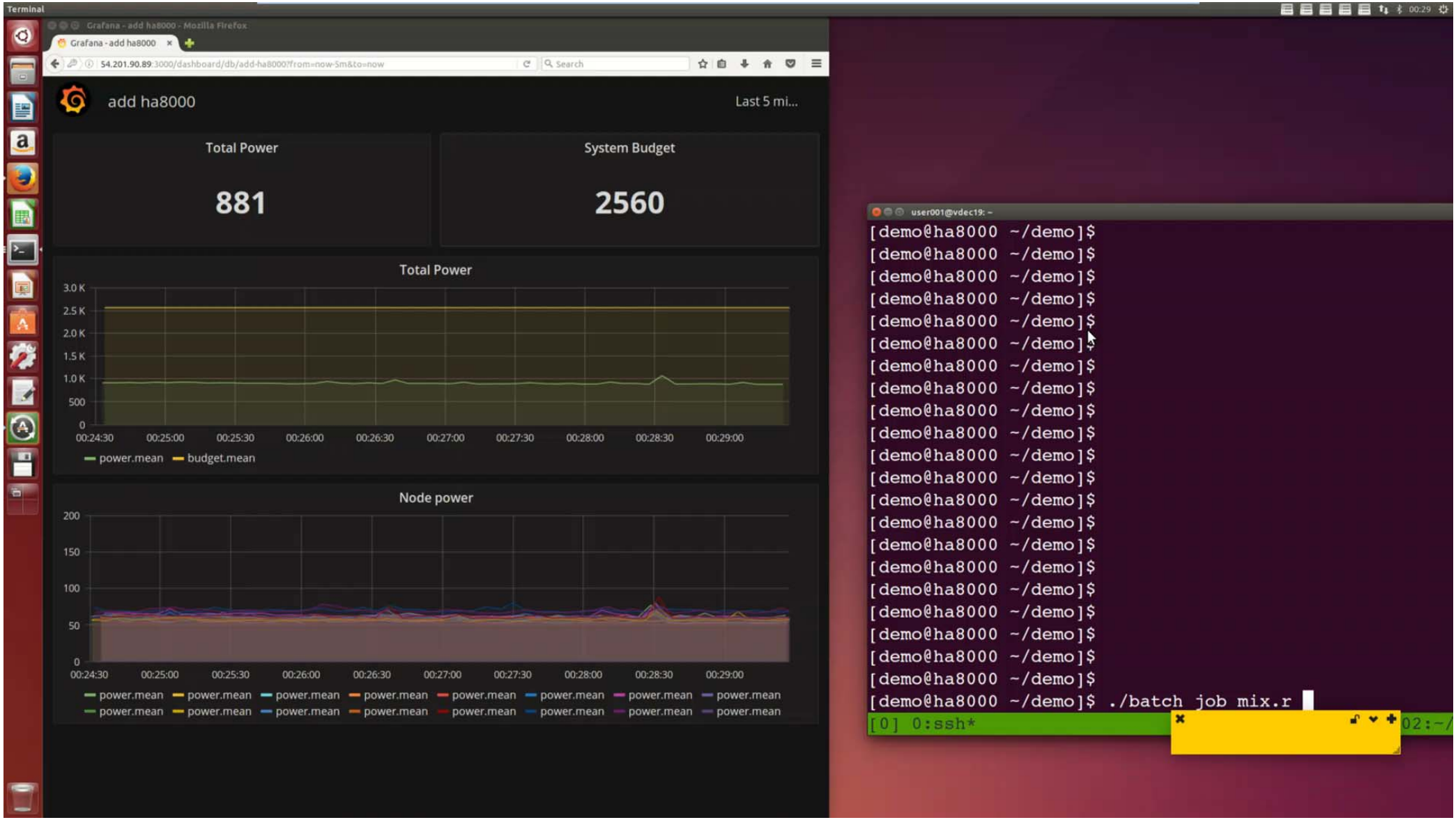
全ノードの電力モニタリング結果



電力制約を設定したスケジューリング時の電力



リアルタイムの電力モニタリングツール



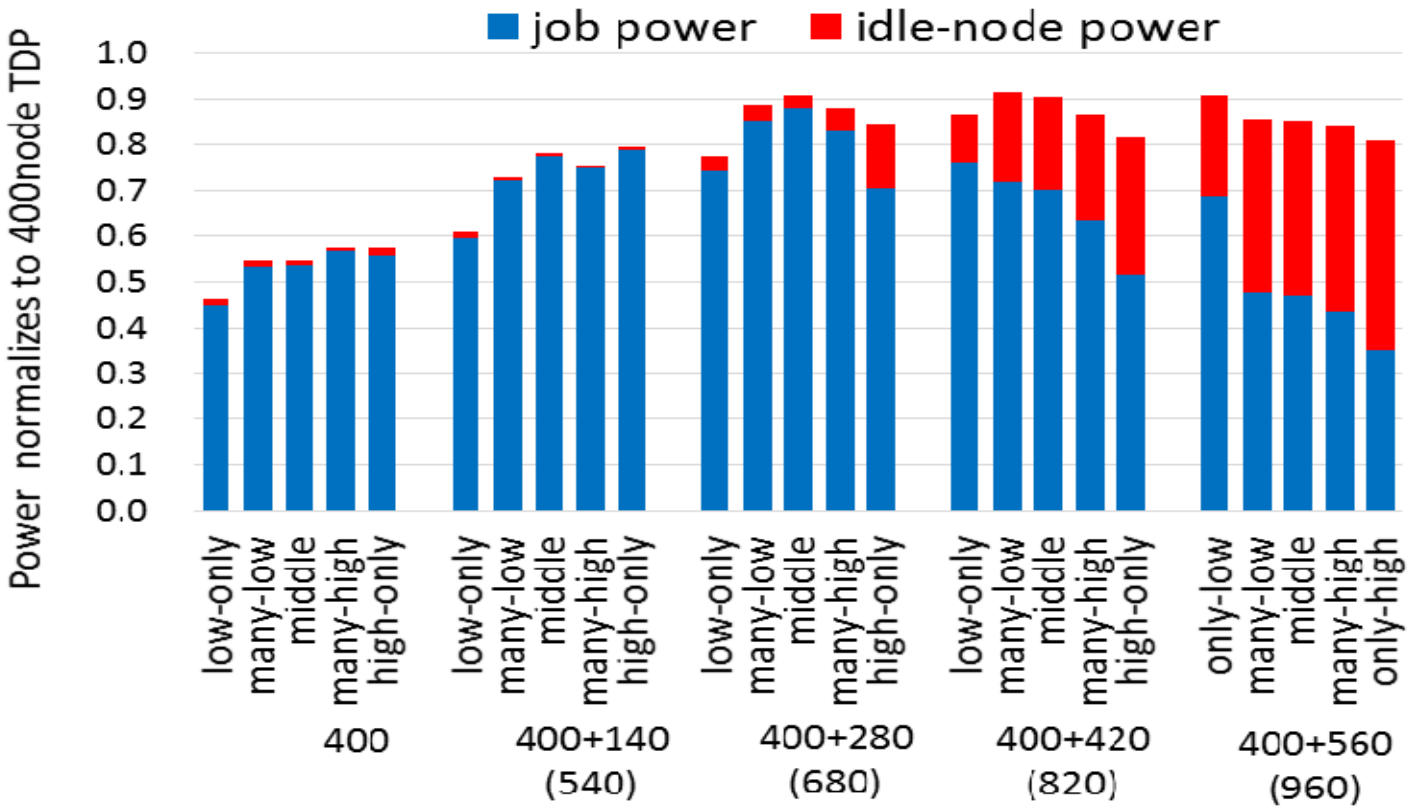
電力制約適応型システムを想定した評価 (1/3)

- 評価の仮定
  - 評価環境：九州大学情報基盤研究開発センターHA8000
    - 400ノードのCPU最大消費電力(TDP)を電力制約として仮定
  - 行列積を並列実行、電力キャップ値は85/70/55Wを各ジョブに割り当て
    - all low-power: 全ジョブ85W、all high-power 全ジョブ85W、Mix: 55Wと85Wをミックス
  - ジョブサイズとサブミットタイミングは理研RICCのログに従う

(R. Sakamoto, et al., “Production Hardware Overprovisioning: Real-world Performance Optimization using an Extensible Power-aware Resource Management Framework”, IPDPS2017.)

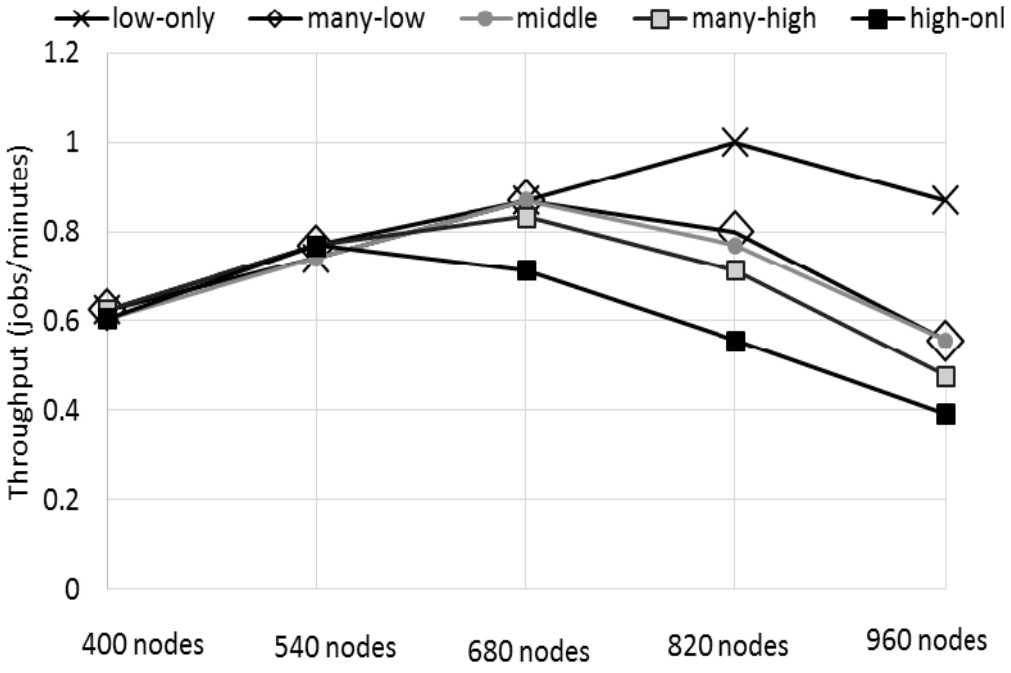
電力制約適応型システムを想定した評価 (2/3)

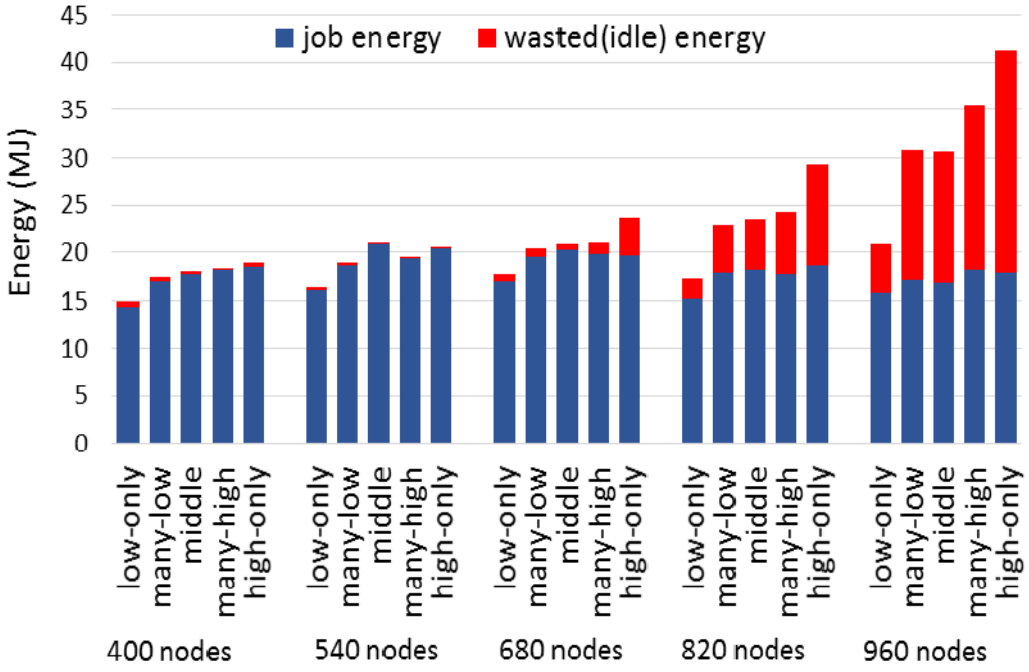
- オーバープロビジョンド環境における電力消費の分析
  - 400ノードの最大消費電力(TDP)を電力制約としノード数を増加
- 追加ノード数が増加するにつれ電力資源をより利用
- 一方でidleノードの電力が増加 → 無駄に電力が消費



電力制約適応型システムを想定した評価 (3/3)

- システムのスループット（性能）と消費エネルギー
  - ノード数増加により電力制約下での性能は向上する傾向、ただしジョブの特性により最良なノード数は異なる
  - ノード数が680ノードよりも多い場合はidleノードの電力が増大
  - オーバープロビジョンド環境ではアイドルノードの電力管理が重要





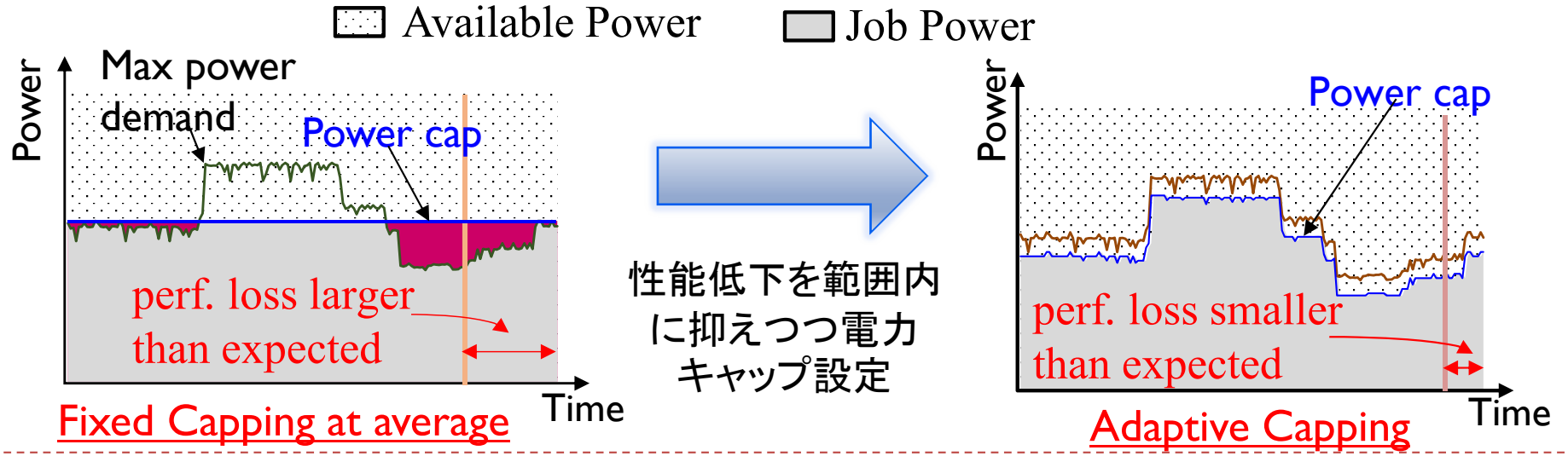
アドバンストなスケジューリング戦略

- エネルギー効率を考慮した電力制約指向スケジューリング
  - 電力余剰に応じて組み合わせ最適化問題により最適なジョブを選択
  - FIFOスケジューリングに比べ25%高速化、10%省エネを達成

  (黄ほか，“エネルギー効率を考慮した電力制約下でのスループット指向ジョブスケジューリング”，HPCS2015.)
- ジョブの要求性能を考慮した適応的電力制御手法
  - 各ジョブの電力モニタリング＋性能低下率予測
  - 要求性能範囲内で電力キャップ低下、余剰電力で他ジョブを実行

    → システムスループット向上、ターンアラウンドタイム削減

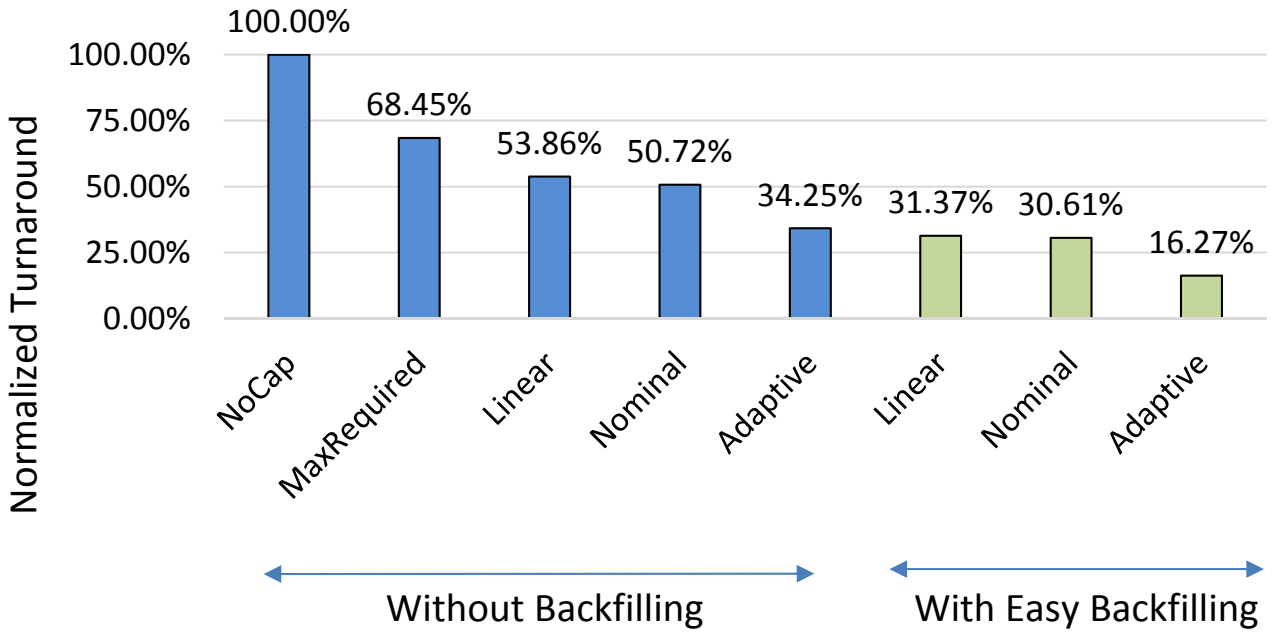
  (T. Cao, et al., “Demand-Aware Power Management for Power-Constrained HPC Systems”, CCGrid2016.)



適応的電力制御手法の評価結果

- 評価環境
  - ジョブログをベースにしたスケジューリングシミュレーション
    - 性能向上に合わせてジョブサブミッションレートを2.5倍
  - ジョブログ：Intrepid – Blue Gene/P system (8core、10,240ノード)
  - 電力制約:1,812kW (TDP：3,624kW)
- ターンアラウンドタイムの評価結果

- NoCap：TDP固定
- MaxRequired　性能要求を満たす範囲の中で最大値
- Linear：許容性能低下率に比例する相対電力値を利用
- Nominal：性能要求を満たす範囲の中で平均値を利用
- *Adaptive*：性能範囲内に納まる電力キャップを適応的に設定(提案手法)



関連発表文献（H28年度）

- 国際会議（査読付き）
  - R. Sakamoto, T. Cao, M. Kondo, K. Inoue, M. Ueda, T. Patki, D. Ellsworth, B. Rountree, and M. Schulz, “Production Hardware Overprovisioning: Real-world Performance Optimization using an Extensible Power-aware Resource Management Framework”, 31st IEEE International Parallel & Distributed Processing Symposium (IPDPS 2017), May 2017 (to appear),.（センター利用に関する記載あり、センター職員の方著者に含む）
  - T. Cao, Wei Huang, Yuan He and M. Kondo, “Cooling-Aware Job Scheduling and Node Allocation for Overprovisioned HPC Systems”, 31st IEEE International Parallel & Distributed Processing Symposium (IPDPS 2017), May 2017 (to appear). （センター利用に関する記載あり）
  - T. Cao, Y. He, and M. Kondo, “Demand-Aware Power Management for Power-Constrained HPC Systems”, IEEE/ACM 16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid2016), May 2016.（センター利用に関する記載予定）
- 研究会発表（査読なし）
  - 坂本龍一, カオタン, 和遠, 近藤正章, 深沢圭一郎, 上田将嗣, 稲富雄一, 井上弘士, “電力制約を考慮した資源管理ツールによるHPCシステムの電力性能解”, 情報処理学会ハイパフォーマンスコンピューティング研究会, 2016年8月. （センター利用に関する記載あり、センター職員の方著者に含む）

まとめ

- 電力制約適応型の資源管理ツールの開発
  - 電力制約に応じたジョブスケジューリングと電力制御
  - HA8000上への導入とオーバープロビジョンドシステムの評価
  - 開発した電力制約適応型資源管理ツールはGitHubで公開
  - 電力を有効利用するジョブスケジューリングアルゴリズムの開発
- 今後の課題
  - ツールの完成度向上
  - 電力制御手法・アルゴリズムの改良
  - 大規模アプリでの評価
- 謝辞
  - スケジューラの導入や全系システムでの評価においてはセンターの皆様に大変なるご尽力を賜りありがとうございました