サイエンティフィック・システム研究会
Scientific Systems

情報システムの効率的エネルギー活用検討 WG
成果報告書

（WG 活動期間：2018 年 4 月～2020 年 3 月）

2020 年 3 月 17 日

サイエンティフィック・システム研究会

情報システムの効率的エネルギー活用 WG

商標について

記載されている製品名などの固有名詞は，各社/各機関の商標または登録商標です.

目次

1. はじめに

1.1. 活動方針

2011 年の東日本大震災以降，電力などのエネルギー消費の問題が顕在化し，特に電力の大需要家である大規模な実験装置を持つ研究機関や IT 機器の集約地であるデータセンターやサーバルームなどでは，電力料金の増加や国の施策であるエネルギー原単位の削減目標などは，達成すべく改善を進めている状況にある.

ただし，手段が目的化した対応がなされていることや，目的(省エネルギー)と目標(業務効率と生産性維持)と手段のバランスを考えず，闇雲に電力削減だけを目的にしてしまっているようなケースが散見され，その活動が本当に省エネルギーの効果があるのか，業務効率や生産性を低下していないかデータに基づいて検討するようなフェーズがないなど，本当の意味での省エネルギーになっていないことが見受けられる.

また，大学・研究機関の特徴として，昔の特殊なシステム導入の名残を残した設備であったり，竣工から年月が経過した建物や部屋をサーバルームとして利用しているケースも多く，スペースと電力量のバランスが合っていない部分や設置される機器が定型的なサーバ・ネットワーク機器ではないために問題が発生するなど，一般的なデータセンターなどとは異なる問題点が発生していることがあるが，特異すぎて課題として共有されていない問題も存在する.

上記の状況を考慮して，情報システムの現状を整理して，エネルギー利活用の指針を示すことを目指す.

1.2. 活動内容

情報システムの中でも消費電力の大きいスパコンを対象とし，検討を開始する.
多くの機関では歴史的な設備を有していたり，年月が経過した既設施設内に設置されていたり，スタンダードな電力利用量や様々な条件（施設，外部環境，利用状況など）の計測が簡単にできない構造になっている場合が多い.

本 WG 前半では，

- 各機関における現状の事例の調査
- 業界標準の指標・規格・将来動向の調査

を実施し，スパコン運用における課題を整理する.

WG 中盤以降は，

- 様々な条件や電力利用状況を調査・評価する方法の検討
- 業務効率や生産性を考慮したエネルギー効率の指標の検討
- 情報システムにおける効率的なエネルギーの利活用の検討

を実施し，報告書にまとめる.

1.3. 活動期間

2018 年 4 月～2020 年 3 月（会合　計 7 回）

1.4. WG メンバー

Table 1 WG メンバー

区分	名前	機関	役割
会員	姫野 龍太郎	理化学研究所(和光)	担当幹事 [~2019 年 4 月] アドバイザー [2019 年 4 月~]
	黒川 原佳	理化学研究所(和光)	担当幹事 [2019 年 4 月~] 推進委員(まとめ役)
	藤田 直行	宇宙航空研究開発機構(調布)	推進委員
	井上 弘士	九州大学	推進委員
	遠藤 敏夫	東京工業大学	推進委員
	塚本 俊之	理化学研究所 計算科学研究センター(神戸)	推進委員
賛助会員 (富士通)	松井 秀司	富士通(株)	推進委員(まとめ役)
	吉岡 祐二	富士通(株)	推進委員
	末安 史親	富士通(株)	推進委員
	クンワー,ラビン	富士通(株)	推進委員 [~2018 年 4 月]
	池田 美季子	富士通(株)	推進委員 [2018 年 4 月~]
	朽網 道徳	富士通(株)	推進委員 [~2019 年 4 月]
	石川 鉄二	富士通(株)	推進委員 [2019 年 4 月~]
	加瀬 将	富士通(株)	推進委員
	遠藤 浩史	(株)富士通研究所	推進委員
事務局	松本 孝之		
	青木 伸子		
	甲斐 友一朗	[~2018 年 4 月]	

1.5. 活動実績

第 1 回会合：2018 年 4 月 17 日（火）　14:00-17:15　富士通（株）本社
出席者：会員 6 名／富士通 7 名／事務局 3 名
・各機関における「システムおよび組織の効率的なエネルギー利用に向けた取り組み」の現状報告
・今後の取り組み(次回以降)についての意見交換，次回会合で取り上げる課題の検討

第 2 回会合：2018 年 8 月 3 日（金）　14:00-17:15　富士通（株）本社
出席者：会員 5 名／富士通 7 名／事務局 2 名
・将来のエネルギー状況，情報システム，施設に関する議論
・今後の取り組み(次回以降)についての意見交換，次回会合で取り上げる課題の検討

第 3 回会合：2018 年 11 月 26 日（月）　13:55-16:50　富士通（株）本社
出席者：会員 6 名／富士通 7 名（情報提供者 2 名）／事務局 2 名
・現状・動向調査
(1)ストレージの動作温度に関する議論
(2)再生可能エネルギーの現状及び今後のデータセンターでの利用方法
(3)冷凍機をはじめとした冷却システムの性能調査
・今後の取り組み(次回以降)についての意見交換，次回会合で取り上げる課題の検討

第 4 回会合：2019 年 1 月 21 日（月）　14:00-17:10　富士通（株）九州支社
出席者：会員 6 名（情報提供者 1 名）／富士通 6 名／事務局 2 名
・効率的なエネルギーの利活用に関する取り組み報告・調査と議論
(1)JAXA におけるノード単位の電源停止
(2)九大におけるエネルギーの効率的な活用に対する施策
(3)電力あたりの演算性能向上（HPC システムの電力に関するトピック）
(4)冷却設備電源の削減見込みに向けた施策
・報告書執筆に向けた目次案・内容に関する検討
・次回以降議論すべき内容に関する意見交換

第 5 回会合：2019 年 4 月 1 日（月）　14:00-16:40　富士通（株）本社
出席者：会員 6 名／富士通 7 名／事務局 2 名
・業務効率や生産性を考慮したエネルギー効率の指標の検討
・情報システムにおける効率的なエネルギー利活用の検討
・報告書の概要についての検討

第 6 回会合：2019 年 7 月 29 日（月）　14:30-17:00　富士通（株）本社

出席者：会員 3 名／富士通 5 名／事務局 2 名

・報告書ドラフト版レビュー（全体，各機関の事例，5/6 章）

・報告書の公開に関する議論

第 7 回会合：2019 年 12 月 10 日（火）　14:00-16:45　富士通（株）九州支社

出席者：会員 5 名／富士通 6 名／事務局 1 名

・SBT 達成に向けた中長期的な取り組みに関する報告

・公的機関の再生エネルギー導入に関する議論

・報告書レビューおよび報告書の公開に関する最終確認

2. 情報システムにおける効率的エネルギー活用の現状と課題

本章では，情報システムを取り巻く電力／エネルギーの活用に関するトピックを取り上げ，特に HPC システムやデータセンターの現状と解決すべき課題について示す．

2.1. 現在の電力事情とエネルギー原単位に関する課題について

2011 年の東日本大震災以降，日本国内では電力などのエネルギー消費の問題が顕在化した．こうした状況は電気料金の増加を招き，電力を多く消費する装置を所有する研究機関や IT 機器の集積地であるデータセンターやサーバルームの運用者を悩ませている．加えて，「エネルギーの使用の合理化等に関する法律」（以下「省エネ法」という．）により，空調，ボイラー，発電設備等を持つ事業所，特に多くの電力を消費する公的機関・研究所においては，エネルギー原単位を中長期的に年平均 1%以上低減させることを求められている．

エネルギー原単位とは，「エネルギー使用量」を「経済活動量」で割った値であり，単位量の製品やお金を生み出すために必要なエネルギー量として定義されている．しかし，実際に「エネルギー量」や「経済活動量」にどの値を用いるかについては統一された値はなく，対象により慣例的にいくつかの値が用いられている．

例えば，事務所などの場合は，「エネルギー使用量」として原油換算量（kl），「経済活動量」として床面積（m^2）が用いられる．他には，「エネルギー使用量」として電力量（kWh），ガス使用量（m^3N），重油換算量（l），熱量（J）が，「経済活動量」として製品個数，売上金額，付加価値金額，重量（t），体積（m^3）が用いられる場合がある．また，電力を多く消費する情報システムなどの機器を所有する機関では，経済活動の補正を行っている場合がある．その場合，機器の運転時間などが用いられる．

情報システムは最低でも数年間は同じハードウェアが稼働するため，「経済活動量」として床面積を用いる場合，エネルギー原単位を削減するためにはエネルギー使用量を削減することが求められる．しかし，システムを運用する立場としては，エネルギー使用量を削減するにはシステムを一部停止するしか方法はなく，その結果利用者に対して十分なサービスが提供できない事態に陥りかねない．これを防ぐためには，「経済活動量」としてシステムを運用することによる価値，あるいはシステムが提供する価値を利用することが課題である．これまで，情報システム，特に HPC システムが提供する価値に関する議論はほとんど行われておらず，特定の 1 組織だけでなく，業界全体で広く議論する必要がある．

2.2. HPC システムにおける電力消費の現状と課題

大規模な情報システムの集積地である HPC システムやデータセンターでは，必要とする電力も他に比べ非常に大きいことから，電気料金およびその削減に大きな関心が集まる．したがって，これらの施設の運用において効率的にエネルギーを活用することは，運用者が取り組むべき重要な課題の一つである．

ここでは，HPC システムに注目し，各機関および各機関が運用している HPC システムにおける現状と課題について以下に示す．

2.2.1. 理化学研究所（和光）

2.2.1.1. システム概要

理化学研究所（和光）（以下，「理研和光」）では，所内ユーザへの情報サービスの提供および所内外への情報提供を目的とした多数の IT 関連システムを運用している．今回は多数の IT 関連システムのうち所内共同利用の計算・データ処理システム（名称：HOKUSAI システム）のシステム概要図を示す．

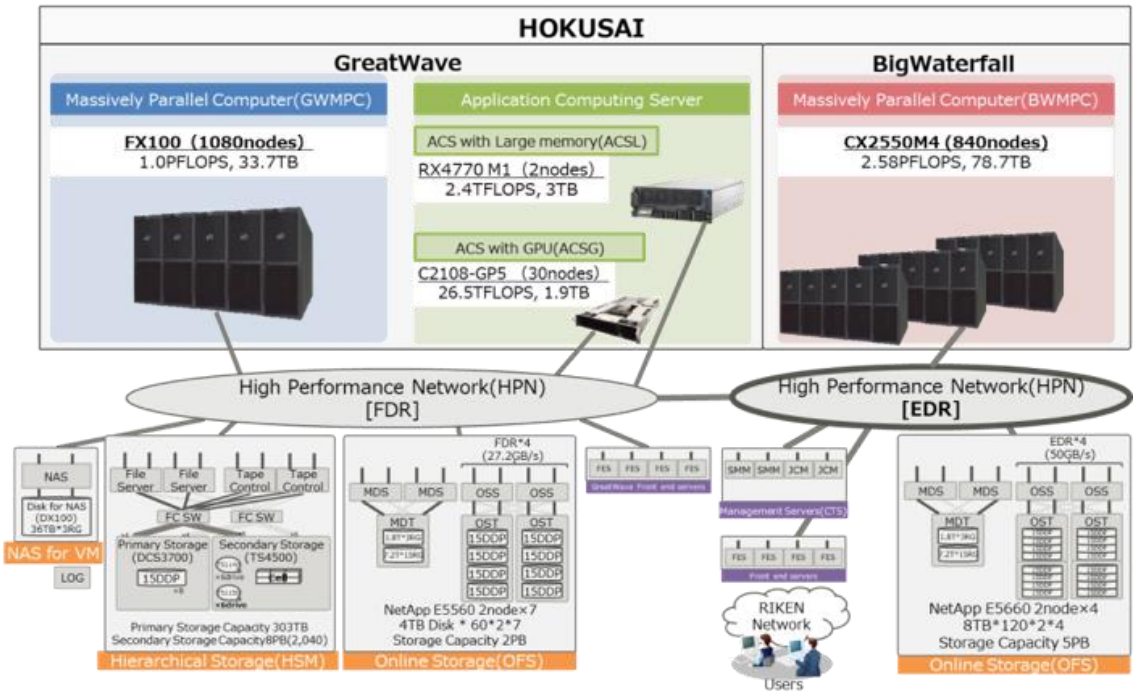


Figure 1 計算・データ処理システム概要

2.2.1.2. 施設設備概要

理研和光の IT 関連システムを設置しているデータセンターの特に冷却関連施設概要を以下に示す．上記，HOKUSAI システムは本データセンター内に設置されている．

本施設設備は 2013 年度から計画的に更新されてきたもので 2016 年度末に更新が完了した．

冷却設備更新前機器は電算室用パッケージエアコンと通称されるもので，2015 年当時で約 15 年間利用してきたもので，更新にあたって最新のパッケージエアコンにリプレースしてもエネルギー効率は低いものとなる．また，高密度実装タイプの IT 関連機器の冷却は空冷から水冷(液冷)へと転換が図られている時期ということもあり，更新後の冷却システムにはエネルギー効率の高い水冷をメインとなるような冷却システムに転換を図った．

IT 機器冷却用としての冷水の供給能力と AHU（Air Hundling Unit）経由で室内を冷却する空冷冷却能力は４：１程度の割合となっており，IT 機器の冷却の大部分は水冷で行う構成となっている．上記の HOKUSAI システムはその発熱の大部分は冷水によって対応しており，空冷を利用するのはネットワーク機器とストレージ機器となっている．

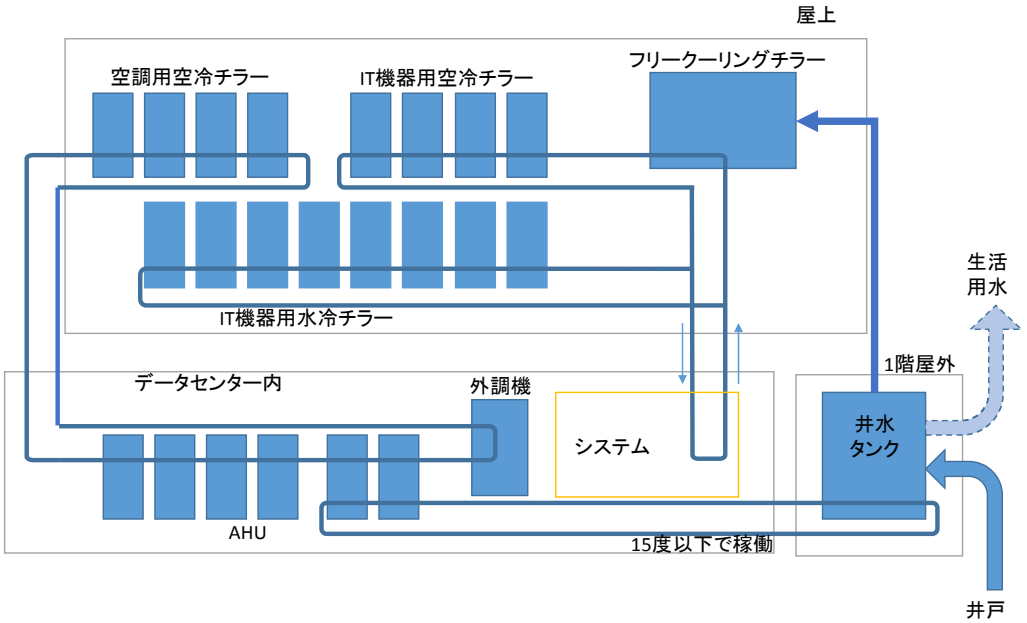


Figure 2 冷却システム概要

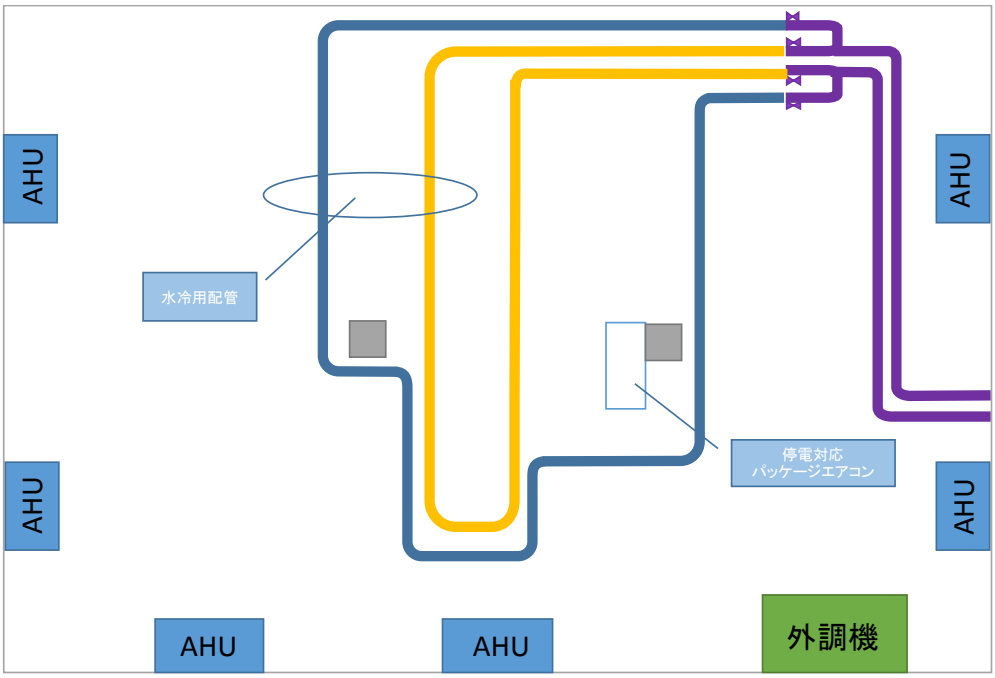


Figure 3 室内概要

2.2.1.3. 消費電力と PUE

本データセンターの改修にあたっての計画策定時に更新前システムの PUE を 2012 年 9 月頃に測定した結果は，概算の PUE が 2 を超える状況であった．そのため，2013 年からはじめた冷却設備更新においては，ピークで 1.5 以下，年平均が 1.3 以下という目標を立てて作業を行った．

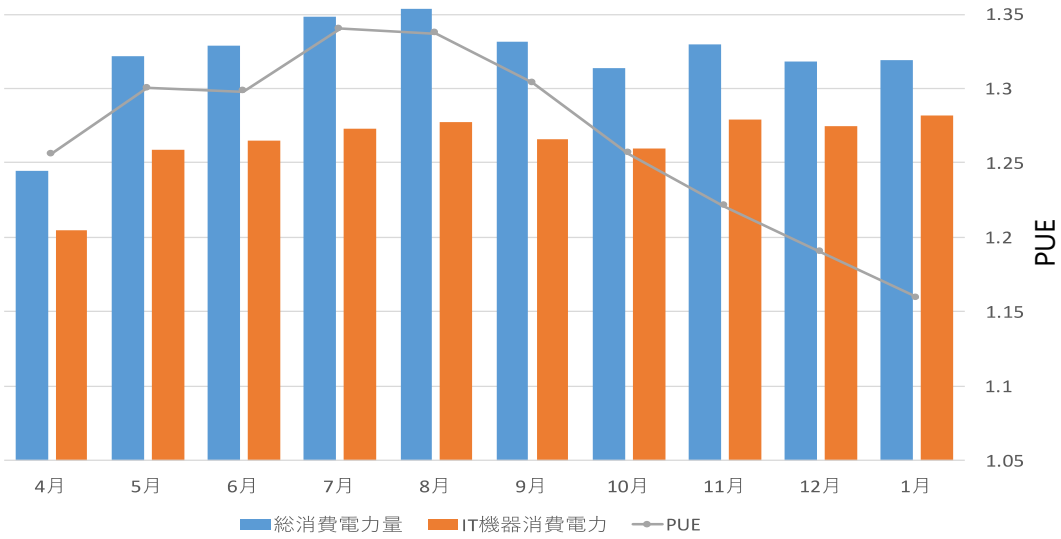


Figure 4 2015 年度の PUE 変遷

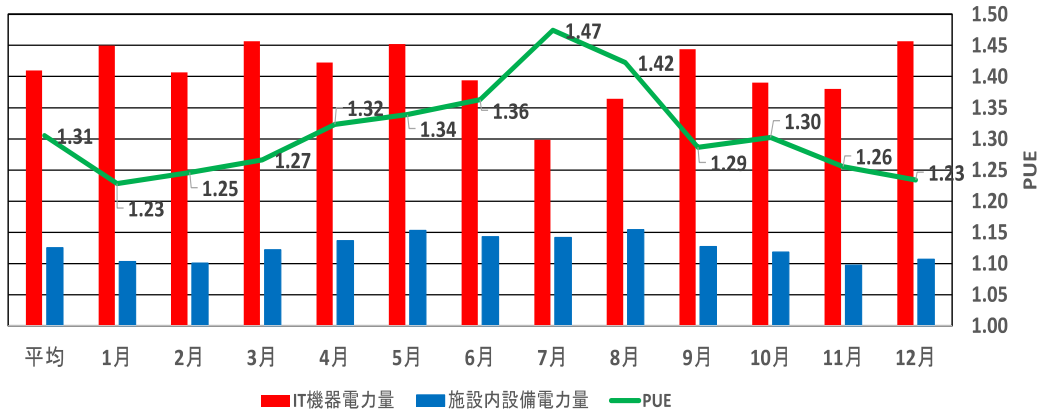


Figure 5 2017 年の PUE 変遷

2.2.1.4.　効率改善の取り組み

2012 年から開始した冷却設備の更新によって，IT システムにかかる冷却を含めた総消費電力は想定を上回る削減が行え，冷却設備だけでみた場合 50%以上の削減できたことになる．今後は冷却設備電力の上昇を IT システムの増加や性能向上に対していかにおさえるかが問題となってくる．

2020 年度頃の IT システムの消費電力の大部分は CPU/GPU の消費電力となっている．また，それらの冷却は高密度に実装を行うほど直接/間接あれ，ほぼ水冷となっていることが予測される．さらにエネルギー効率の高い冷却を行えるかは，いかに高い水温で冷却できる IT システムを導入できるかどうかであり，冷却水温が高い場合は，冷却設備電力が削減できることになる．

また，空冷機器においても，空調温度を高めに設定することや，冷却範囲を局所化するなどの方策を実施することでエネルギー効率が高められると推測される．

2.2.1.5.　課題

最終的に事業所として省エネルギーを達成するために，エネルギー原単位削減に結びつける必要があるが，データセンターなどにおいては，生産性を示す原単位の分母となるは，フロア面積とされることが多い．これは IT システムの生産性とはかけ離れたパラメータであることが多い．すなわち，設置面積が同じであっても IT システムの生産性は LSI ベースのデバイスの数や種類によって大きく変化し，フロア面積だけで，生産性を定義するのは実体との乖離が大きく，もっと実体に即した生産性を提案していく必要がある．す特に事業所の消費エネルギーのうち，IT システムの占める割合が多い場合は死活問題である．

2.2.2. 理化学研究所 計算科学研究センター（神戸）

2.2.2.1.　システム概要

理化学研究所計算科学研究センター（R-CCS）が運用するスーパーコンピュータ「京」のシステム概要図を以下に示す．

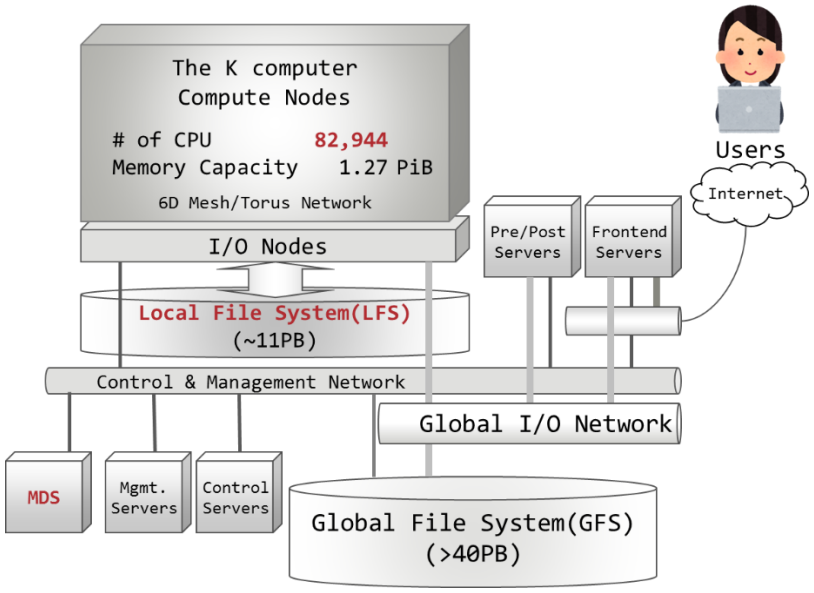


Figure 6「京」のシステム概要

「京」の消費電力は，数万ノード規模の大規模なジョブを実行する場合で最大 15MW 程度，それ以外の通常運用の場合で 11MW 程度であり，電力変動の幅が大きい特徴がある．

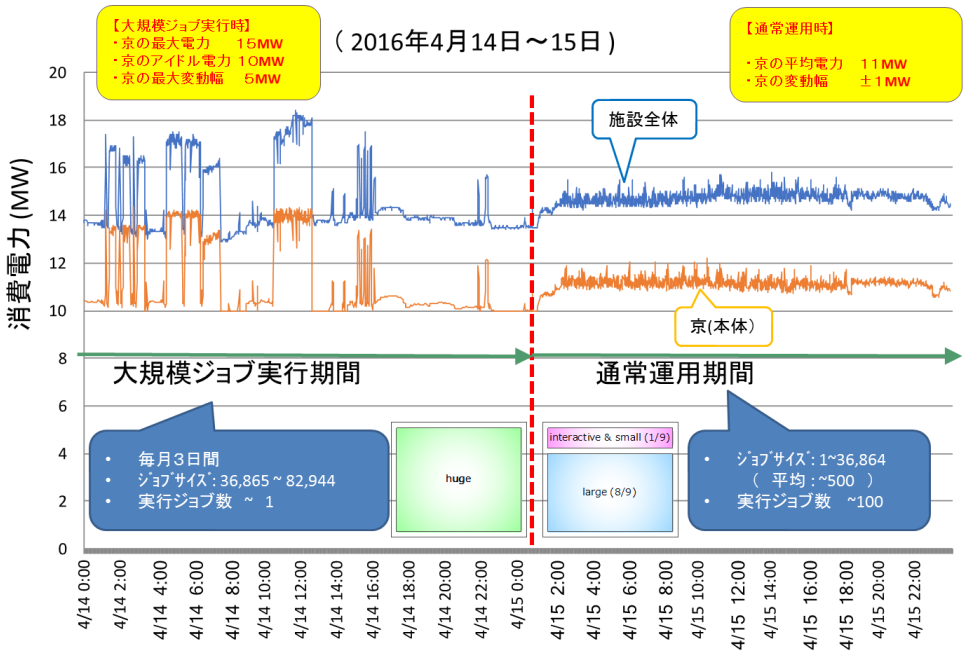


Figure 7「京」の電力変動

2.2.2.2.　施設設備概要

「京」を運用する R-CCS の施設概要を以下に示す.

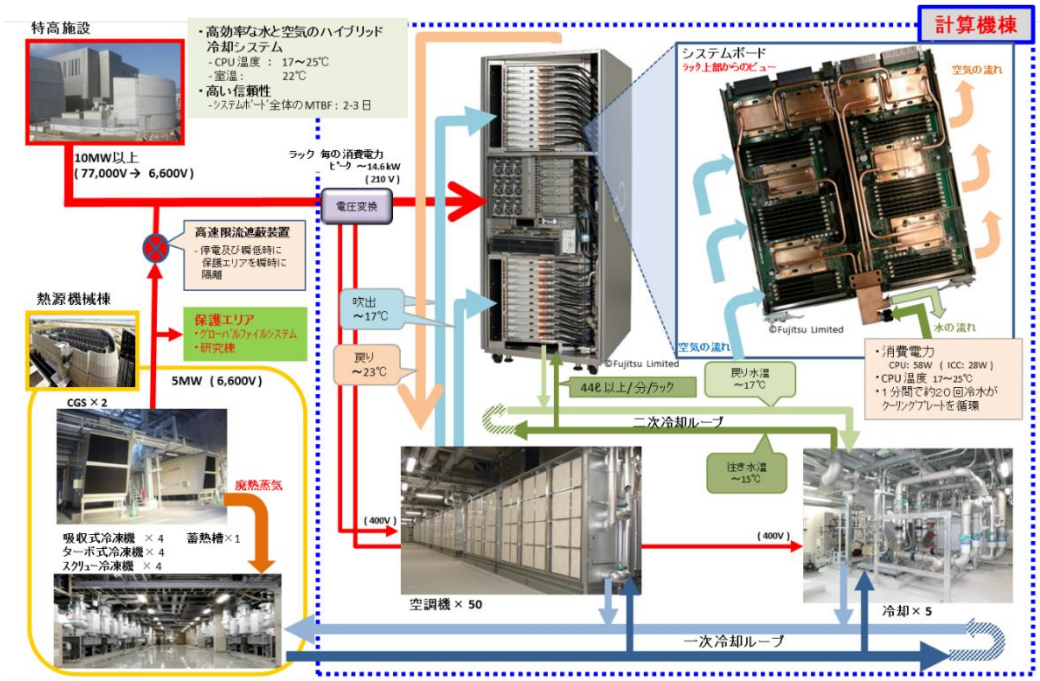


Figure 8 R-CCS 設備概要

2.2.2.3.　消費電力と PUE

R-CCS 施設全体の消費電力は通常 15MW 程度であり，そのうち 11MW±1MW 程度を商用電力で供給しており，不足分はコジェネレーションシステム（CGS）による自家発電によってまかなっている.

設備全体の電力は，運用開始当初は平均 3MW 超であったが，空調設備や冷却塔，冷凍機の効率改善，契約電力超過防止活動等の継続的な省エネ活動により，2017 年度には平均 2.5MW 以下まで低減することができた．これにより，運用当初 1.4 程度であった PUE を 1.3 程度まで低減できた．以下に PUE の変遷と効率化改善への取り組み（詳細は後述）の関係を示す.

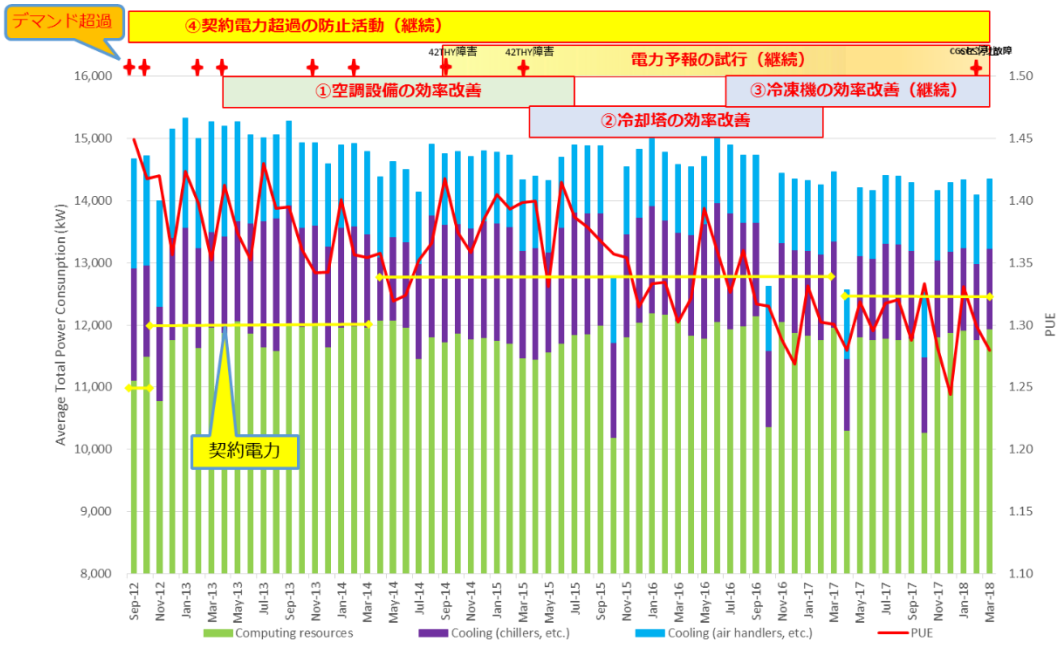


Figure 9 PUE の変遷

2.2.2.4.　システムの変遷とエネルギー原単位削減の取り組み

R-CCS では供用開始（2012 年）以降毎年 1%のエネルギー原単位の削減に取り組み，達成してきた．R-CCS におけるエネルギー原単位の算出式は以下の通りである．

エネルギー原単位　＝　原油換算(kl)　÷　換算延床面積(㎡)

換算延床面積＝延べ床面積＋当該年度の運転時間／基準年度の運転時間×対象延床面積

2.2.2.5.　効率改善の取り組み

以下に示す４つの効率化改善の取り組みの実施により，契約電力の超過（デマンド超過）を撲滅することが出来た（設備障害時を除く）．

①　空調設備の効率改善

空調機の風量を１／２に削減し，温度差を ２ 倍とすることで，必要な排熱量を確保しながら電力削減を実施した．これに伴い，計算機室の平均温度が下がり，結果をとしてメモリ故障率の 低減の効果を得る事が出来た．

空調機の風量半減：空調機の RAS 機能（2 個のファン）を利用し，ファン 1 個を停止することで，

風量 6 割，電力半分を可能とした．

温度差を 2 倍：空調機の吹き出し温度と戻り温度の差を，3℃差から 6℃差に変更

吹き出し設定温度を 21℃→17℃，戻り温度を 24℃→23℃

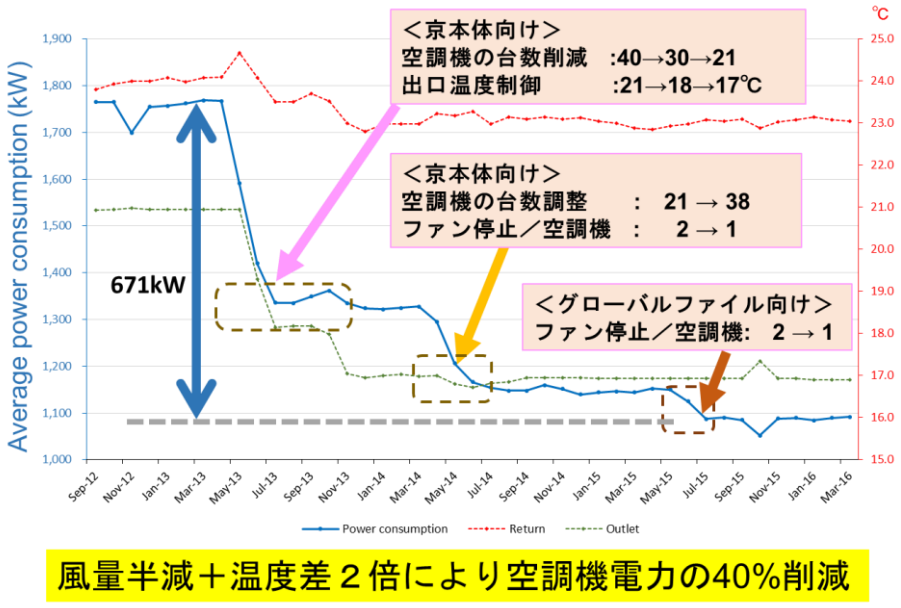


Figure 10 空調設備の効率改善

②　冷却塔の効率改善

R-CCS の熱源機械棟の屋上に設置されている複数の冷却塔について，以下の対応を実施した．これにより，ショートサーキットの消滅およびプリアラームの抑制ができ，結果として余分な冷凍機稼動回数を劇的に減らすことができ，2016 年度以降の効率化（電力削減）につながった．．

・プーリー交換による風量アップ

・白煙防止キャップ撤去による能力アップ

・熱源機械棟外壁パネル撤去による風通し改善

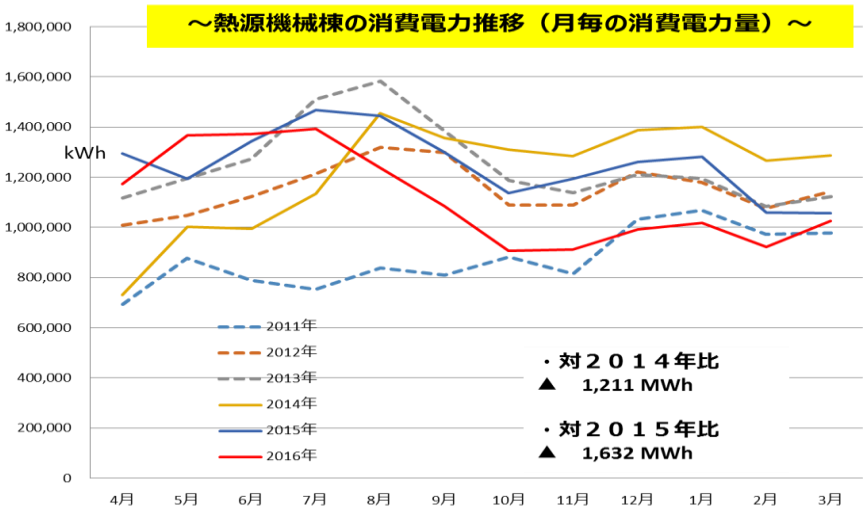


Figure 11 冷却塔の効率改善

③　冷凍機の効率改善

「京」では，フルノードジョブを実行する際，電力が 4MW 程度急激に増大する場合がある．通常大きな電力負荷の変動には，冷凍能力の高い蒸気吸収式冷凍機を追加起動や出力増加で対応するが，出力が安定するまでに数十分（出力増加）～数時間（追加起動）を要するため，そうした短時間での急激な電力増大には対応することが難しい．こうした状況については，起動や出力の安定が早いターボ冷凍機を追加起動して対応している．しかし，起動操作が手動のため，電力上昇を検知して即座に人手で起動することの作業負荷が高い．そこで，急激な電力上昇時に一時的に排熱を支援するため，温度成層方式の蓄熱槽を 2016 年度に導入した．

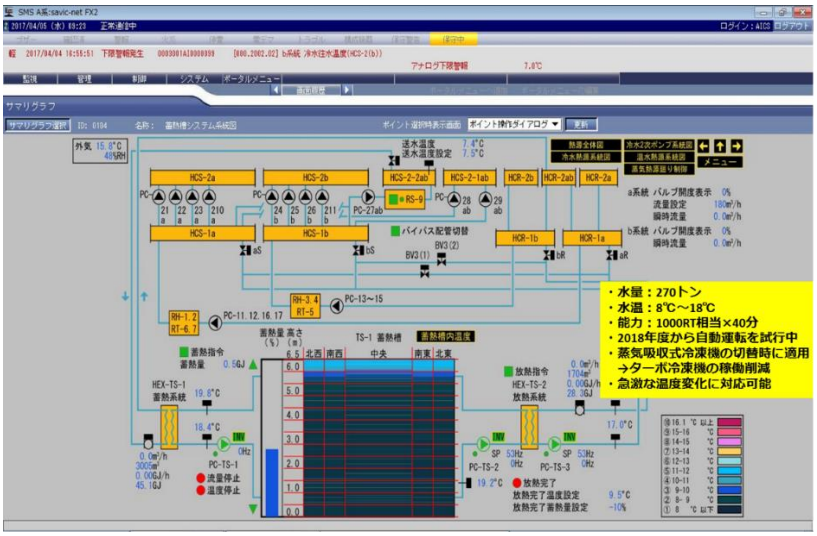


Figure 12 温度成層方式の蓄熱槽

- 契約電力超過の防止

「京」では過去何度か契約電力超過が発生している．その度にペナルティや契約電力上限の増大が発生し，運用コストに与える影響が非常に大きくなった．このことから，電力超過を防止する仕組みの確立に取り組んだ．

①大規模ジョブ実行時の投入ジョブの審査制度の導入

（小規模での実行時電力を申告し，大規模での実行前に最大消費電力を確認）

②施設監視データ（受電電力デマンド値ほか）と「京」システムの連携

③ジョブと消費電力の紐づけ（履歴を蓄積・評価）

④契約電力超過チェックの自動化

⑤契約電力超過リスク発生時のジョブ停止処理連動

⑥停止対象ジョブ選定アルゴリズムの開発・運用（計算資源を無駄にしないため）

⑦「京」本体の消費電力予測システムの開発・運用

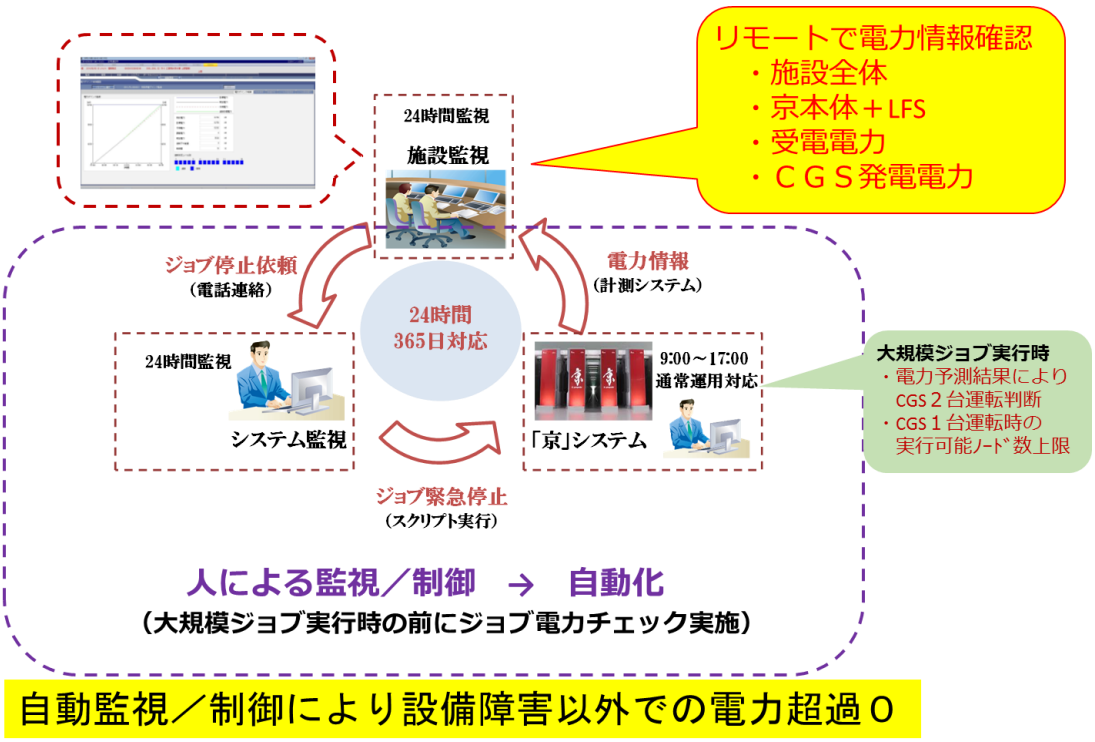


Figure 13 契約電力超過の防止

2.2.2.6. 課題

「京」本体の消費電力は最大 15ＭＷであるが，次のスパコン「富岳」の消費電力は「京」の 2 倍以上と想定されており，電力変動幅は「京」より更に大きくなると考えられる．現在，「富岳」向けの設備増強工事が実施中であるが，現状で以下の課題がある．

- 電力変動への対応
  必要となる排熱能力については，1 台の冷凍機追加で対応可能である．しかし，「富岳」規模の電力変動に連動する熱負荷変動に対応する「冷却システム」については，複数台の冷凍機での対応が必要となるため，全体としての追従能力や冷凍機保護の観点からリスク対策が十分とれているかを今後確認しながら調整していく必要がある．
- エネルギー原単位の見直し
  消費電力が 2 倍以上となるため，新たなエネルギー原単位の適用が必要となる．「京」で採用していた稼働時間を考慮した換算延床面積に対して，性能値を考慮した新たな換算延床面積を検討する必要がある．なお，検討にあたっては R-CCS だけでなく，理研全体への影響を考慮する必要があるため，関係部署との調整も必要となる．

2.2.3. 宇宙航空研究開発機構（調布）

2.2.3.1. システム概要

宇宙航空研究開発機構(以下，「JAXA」)のスーパーコンピュータシステム(JSS2: Jaxa Supercomputer System generation 2)のシステム構成を次に示す．

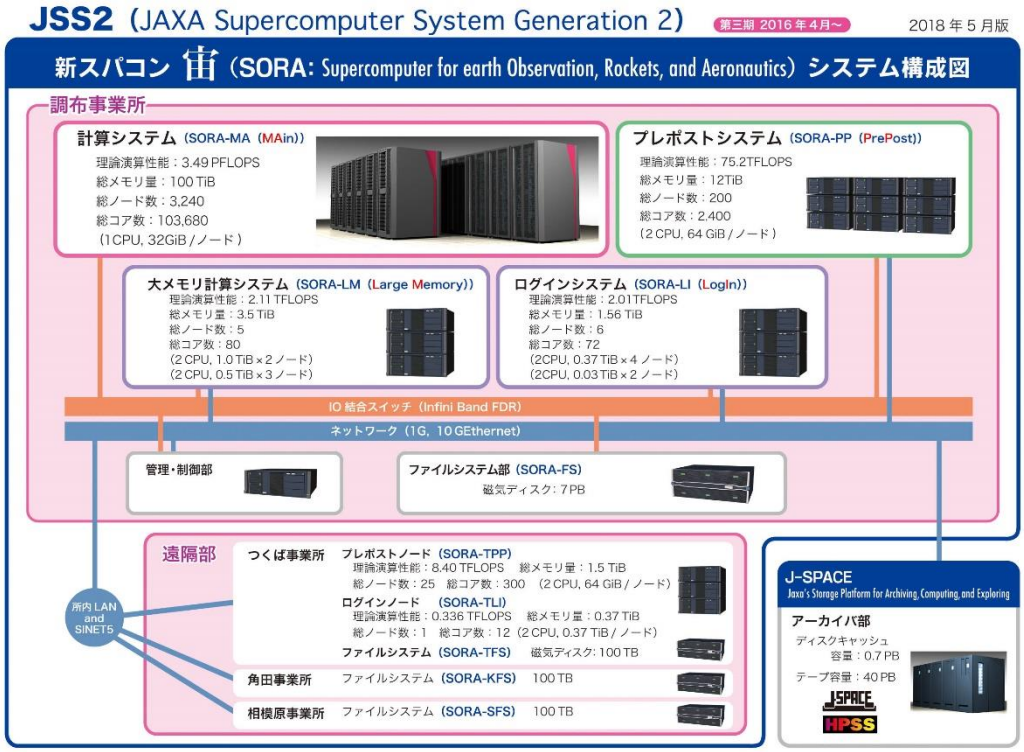


Figure 14 JAXA Supercomputer System generation 2 概要

2.2.3.2.　効率改善の取り組み

- 製造技術による効率改善

JSS2 の主計算システムである SORA-MA(Supercomputer for earth Observation, Rockets, and Aeronautics - MAin system)と前世代の主計算機である JSS-M システムを比較すると，設置面積(ラック)当たりの演算性能が約 200 倍に向上している．半導体技術の進歩により，単位計算性能当たりの消費電力は劇的に減少しているが，その減少を上回る計算性能向上の要求があるため，スーパーコンピュータシステム全体としての消費電力削減努力が必要となっている．SORA-MA と JSS-M を比較すると，理論性能当たり約 25 倍の電力性能の向上が見られる(Figure 16)．これには，プロセッサの性能向上(Figure 17 (a))や高効率電源(Figure 17(b))等が寄与している．

Figure 15 設置面積（ラック）当たりの演算性能向上

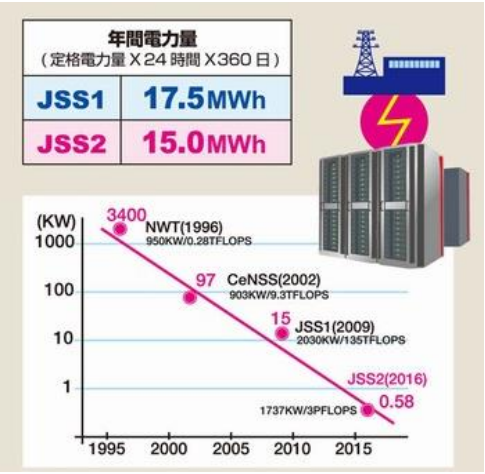


Figure 16 1TFLOPS あたりの歴代計算機の消費電力

(a)　(b)

Figure 17 省電力化に寄与する要素技術例

- 運用環境による効率改善

スパコンの回路は，集積度が上がると大量の熱を出すため，従来の空冷方式では冷却しきれなくなる．JSS2 ではコールドプレートを用いた水冷方式を採用し，効率の良い発熱除去を行っている(Figure 18)．

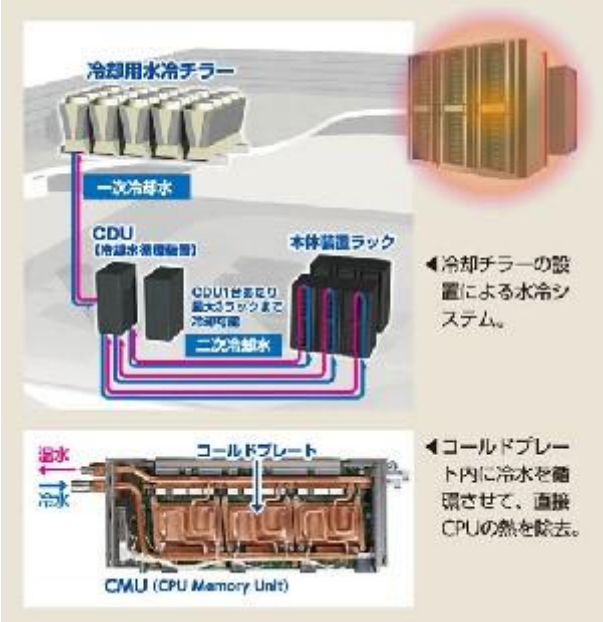


Figure 18 SORA-MA の冷却システム

また，計算機室の環境を設置システムに合わせて変更している．JSS2 システムの設置面積縮小に合わせて壁を設置し空調エリアを約 1/2 にすること，及び，前述の水冷化により空調機の稼働台数を 40 台から 14 台へ削減した．更に，JSS-M で採用した設置上のホットアイル・コールドアイルの考え方は引き続き採用している．

更に，継続的に空調温度設定を見直し，現在は JSS-M 当時よりも計算機室内の温度を 3℃上げた 21℃に設定し空調に必要なエネルギーの更なる減少に努めている．水冷技術の採用と相まって，JSS2 全体の空調に必要なエネルギー量は JSS(前世代のシステム)のそれに比べて約 1/6 となった(Figure 19)．

Figure 19 空調エリアと室温による省エネルギー化

- ジョブスケジューラと連携したノードの電源制御による効率改善

ジョブが実行されずに待機している計算ノード(待機ノード)の消費電力(待機電力)を抑える省電力方式は，ジョブの実行性能に影響を与えないことから有効な方式である(「PC クラスタにおける省電力化のための自動電源制御方式」，研究報告ハイパフォーマンスコンピューティング（HPC）,2017-HPC-160(2),1-7 (2017-07-19) , 2188-8841)．待機ノードの状態が長時間継続する場合，電源を停止することで省電力化を実現することができる．バッチジョブスケジューラがスケジュールした計算ノードへのジョブ配置情報を元に待機時間を予測し，待機ノードを電源停止する自動電源制御方式 JSCAPS(Job Scheduling Aware Power Save)を実装，運用している(Figure 20)．

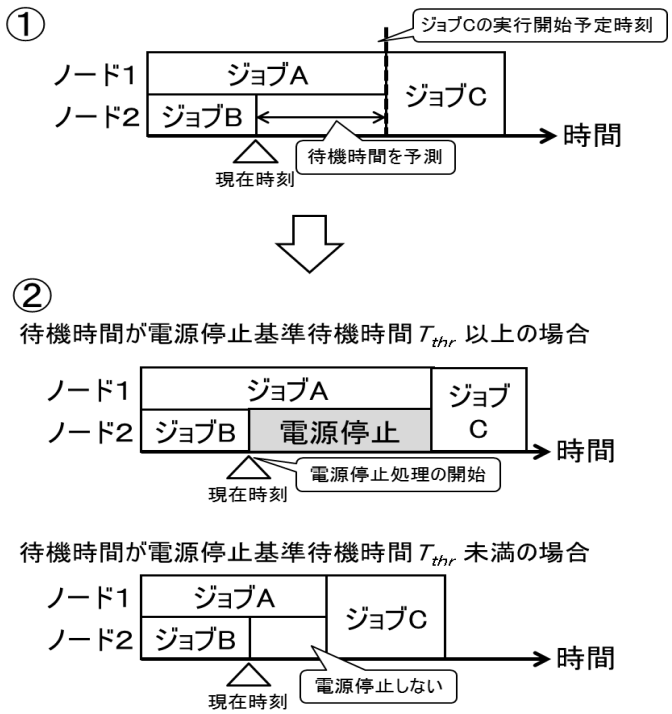


Figure 20 JSCAPS による電源停止タイミングの決定

- テープ媒体の利用による効率改善

ハードディスクは，書き込みや読み込みをしていない時でも電気的に動作し電力を消費する．システム全体ではあるが，Figure 21 に示す通り，ファイルシステム(媒体としてハードディスクを使用)とアーカイバ(媒体として主にテープを使用)の消費電力はシステムレベルで 10 倍の差がある．アーカイバの特性を理解した上でファイルシステム上のデータをアーカイバで保管・運用することでシステム全体としての電力効率の改善が期待できる．

	ファイルシステム			アーカイバ			
	全体	Disk	ｻｰﾊﾞ 等	全体	Tape	Disk Cache	ｻｰﾊﾞ 等
記憶容量 a	5.25 [PB]	5.25 [PB]		21.72 [PB]	21 [PB]	0.72 [PB]	
総I/O速度	50 [GB/s]	(144 [GB/s])	(50 [GB/s])	10 [GB/s]	(10 [GB/s])	(20 [GB/s])	(10 [GB/s])
電力 b	48.43 [KVA]	38.4 [KVA]	10.03 [KVA]	19.22 [KVA]	6.43 [KVA]	6.21 [KVA]	6.58 [KVA]
電力/容量 b/a	9.22 [KVA/PB]	7.31 [KVA/PB]		0.88 [KVA/PB]	0.31 [KVA/PB]		

システムで10分の1

ドライブで20分の1以下

Figure 21 とあるシステムにおけるファイルシステムのアーカイバの電力比較

2.2.3.3.　課題

技術の進歩により電力効率の良いシステムが登場し，それを積極的に採用したシステム構築を心がけているが，システム増強の需要が大きく，今後電力量の増加を検討しなければいけない状況にある．電気代の受益者負担による事業所内での電力量削減効果を狙う考え方もあるが，使用できる電力量が事業のボトルネックとなってきている．エネルギー原単位の削減以外に，東京都では CO2 排出量規制があり，その対応が大きな事業課題となっている．

2.2.4. 九州大学 情報基盤研究開発センター

2.2.4.1.　システム概要

九州大学情報基盤研究開発センターが運用するスーパーコンピュータ「ITO」のシステム概要図を以下に示す．本システムは，国の第 5 期科学技術基本計画に示された超スマート社会の実現，ならびに AI（人工知能・機械学習）・ビッグデータ，データサイエンスなどに対応した研究基盤の提供を目指

し，柔軟な利用形態を提供できるスーパーコンピュータシステムとして仕様が策定されたものである．異なるシステム間の協調作業やデータサイエンスにおける対話的な作業を支援する大規模プライベートクラウド環境（フロントエンド）と大規模シミュレーションや機械学習のための高性能計算ノード群（バックエンド）を高速ファイルシステムを介して連携運用する構成となっている．また，パブリッククラウドとの本格的な連携インタフェースを導入し，オープンデータと連携したスーパーコンピューティングの方向性を示し，新たな利用者層と研究課題に向けた研究基盤を提供する．さらに，本システムから新たに導入された詳細な電力モニタリング機構と制限電力内のジョブスケジューリング機能を活用し，インテリジェントな省電力運用の確立を目指す．

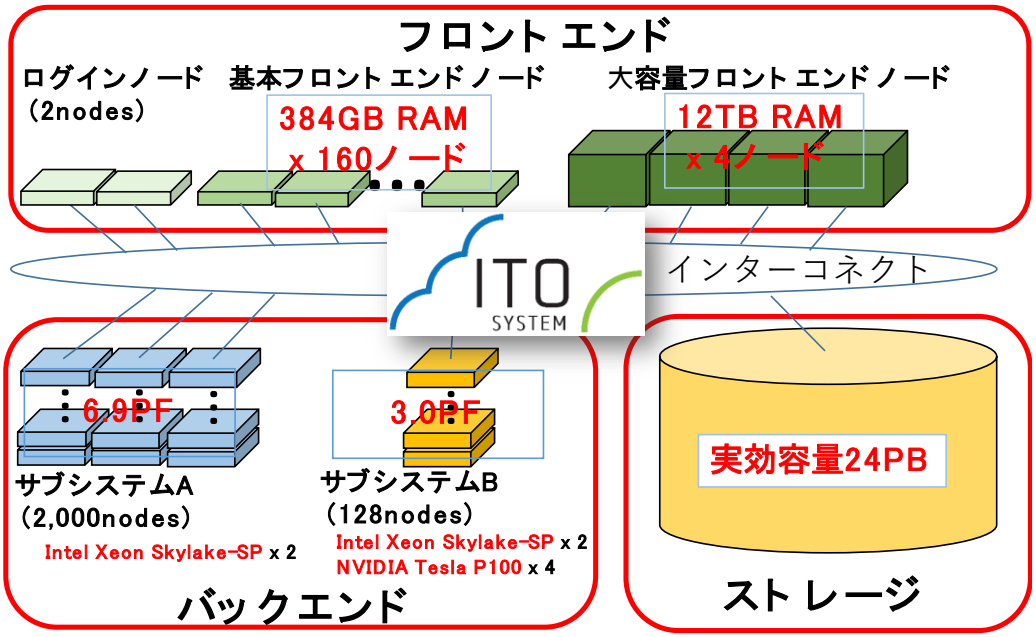


Figure 22「ITO」システム概要

2.2.4.2.　施設設備概要

「ITO」を運用する九州大学情報基盤研究開発センターの施設概要を以下に示す．冷却に関しては，計算ノードの CPU とメモリ，GPU が 2 次の水冷であり，計算ノード中の他の部品およびフロントエンド，ストレージ，ネットワーク装置は空冷である．

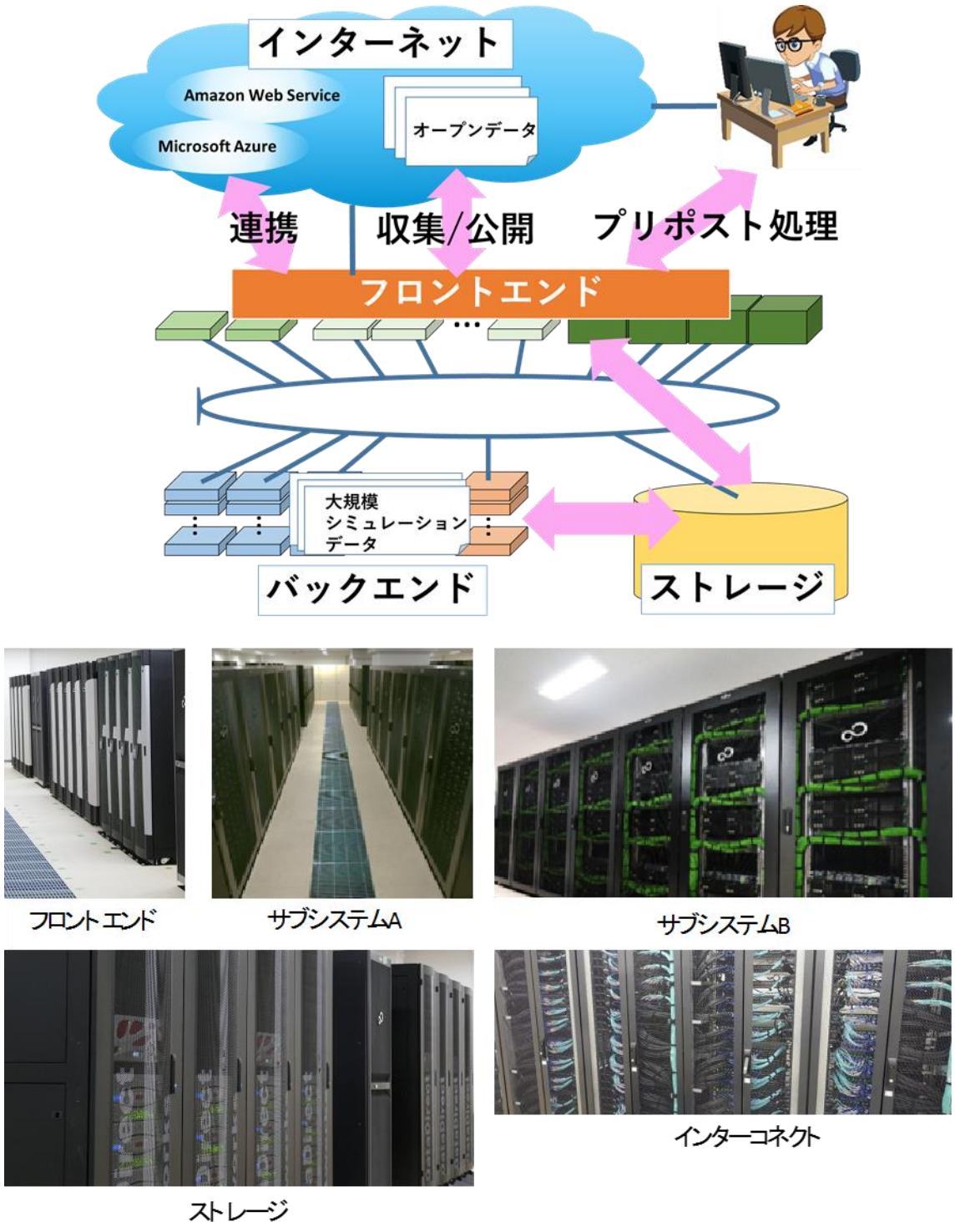
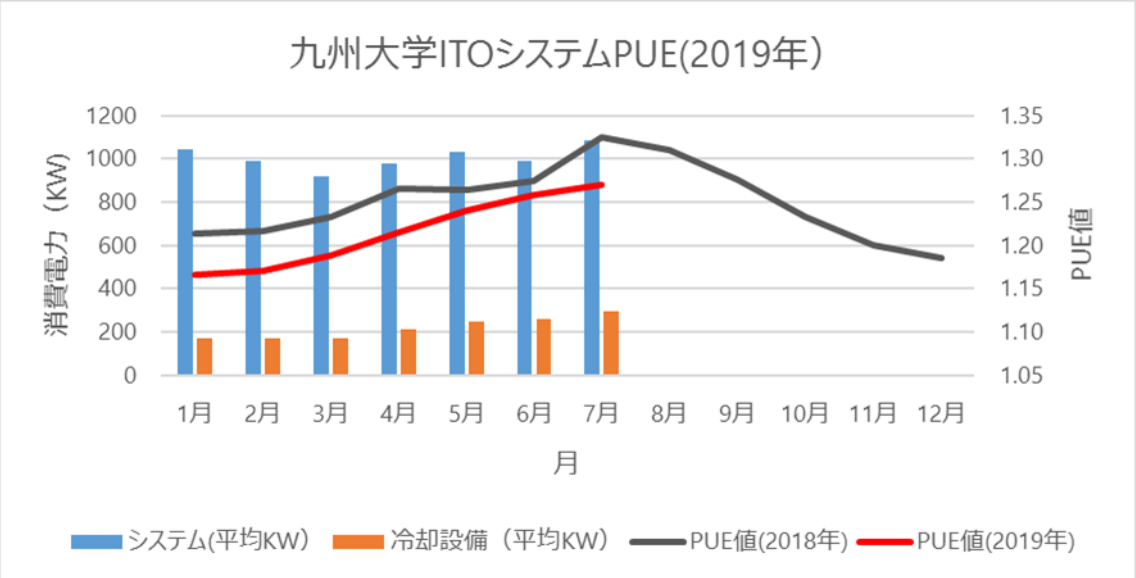


Figure 23 九州大学情報基盤研究開発センター施設概要

2.2.4.3.　消費電量と PUE

「ITO」稼働からのシステム，冷却設備の消費電力および PUE の状況を以下に示す．2019 年 7

月時点での 2019 年平均 PUE は，1.22 と水冷，冷却リアドアなどの効果により比較的良い値となっており，今後気温が下がることによりさらに良くなる見込みである．2018 年と比較し 2019 年の PUE が良くなっているのは，稼働率向上によるシステム消費電力の増加に比べ，冷却設備消費電力の増加比率が低いためである．



	2019年												平均
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	
システム(平均KW)	1043	988.8	920.5	979.7	1031.5	993.8	1086.2						1006.21
冷却設備（平均KW）	172.9	169.3	172.9	210.6	248.3	257.3	294.3						217.94
PUE値(2019年)	1.17	1.17	1.19	1.21	1.24	1.26	1.27						1.22

	2018年												平均
	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月	
システム(平均KW)	775.2	796.1	806.4	745.7	832.4	861.5	770.5	912.4	890	946.6	1008.9	1010.4	863.01
冷却設備（平均KW）	166.2	173.1	187.6	198.2	220	236.8	251.1	283.5	245.4	221.1	202.4	186.7	214.34
PUE値(2018年)	1.21	1.22	1.23	1.27	1.26	1.27	1.33	1.31	1.28	1.23	1.20	1.18	1.25

Figure 24 PUE の変遷

2.2.4.4.　システムの変遷とエネルギー原単位削減の取り組み

九州大学情報基盤研究開発センターにて運用したスーパーコンピュータシステムの変遷は以下の表に示す通りである．表で示す通り，ITO の導入により電力効率が大幅に改善されており，結果としてエネルギー単位元の削減に貢献していることが分かる．

なお，ITO は九州大学の新キャンパスに設置されたものであり，PRIMERGY CX400，PRIMEHPC FX10，ならびに，HA8000+SR16000 とは設置場所ならびに設置環境が異なる．参考までに，移転前の PRIMEHPC FX10 が設置された環境では供給電力：1,100 KVA，冷却能力：1,247 KW（水冷：400KW ＋ 空冷：847KW）であったのに対し，移転後の ITO 設置環境は供給電力：3,000 KVA，冷却能力：2,472 KW（水冷：1,800KW ＋ 空冷：672KW）である．

Table 2 システム変遷

システム名	PRIMERGY CX400	PRIMEHPC FX10	HA8000 + SR16000	ITO
システム稼動時期	2012-2017	2012-2017	2013-2017	2018-
理論演算性能	0.966 PF	0.182 PF	0.720 PF	10.43 PF
消費電力	692 KW	279 KW	824 KW	1,742 KW
電力効率	1.40 GFlops/W	0.65 GFlops/W	0.87 GFlops/W	5.99 GFlops/W

2.2.4.5. 効率改善の取り組み

- 電力モニタリング機構の拡充

ITO では，従来システムと比較して電力モニタリング機構を大幅に増強している．監視対象は，バックエンドサブシステム A，バックエンドサブシステム B，フロントエンドサブシステム，ストレージサブシステム，インターコネクトネットワークのスイッチ，ルータ(ネットワーク機器)，冷却装置，であり，時間粒度は 1 分毎，分解能はラック毎，である．また，投入・実行されたジョブに関し，実行中の最大消費電力を記録することが可能である．対象は CPU およびメモリの最大消費電力であり，Intel 社が提供する Running Average Power Limit（RAPL）を利用している．現状，ユーザへの電力情報開示方法は問合せベースである．

- 電力制御機構

消費電力を考慮したシステム運用のための機能を整備（ただし，運用上現在は未適用）している．また，ジョブごとに必要な最大消費電力情報に基づき電力を考慮したジョブスケジューリングに関する検討を進めている．例えば，サブシステム A 全体または B 全体の電力上限値設定や，ノードの電力当たり性能（電力効率）のばらつきを考慮したスケジューリングなどが挙げられる．さらに，消費電力削減機能としてアイドルノードの電源 OFF による電力削減も可能である．なお，電力キャッピング機能を用いた電力効率に関しては，深沢，南里，本田による「MHD シミュレーションコードを利用した CPU 電力キャッピング下でのスーパーコンピュータシステム ITO の消費電力特性評価」（情報処理学会研究報告 Vol.2018-HPC-167 No.18 ）にて報告されている．

2.2.4.6. 課題

現状，システムの運用における電力情報を蓄積している段階である．今後の課題としては，これら蓄積した電力情報に基づき，電力キャッピングの適用など様々な施策を適用することが挙げられる．

2.2.5. 東京工業大学 学術国際情報センター（GSIC）

2.2.5.1. システムの変遷とエネルギー原単位削減の取り組み

2017 年 8 月に運用開始した，東京工業大学　学術国際情報センター運用する TSUBAME3.0

スーパーコンピュータのシステム概要図を以下に示す.

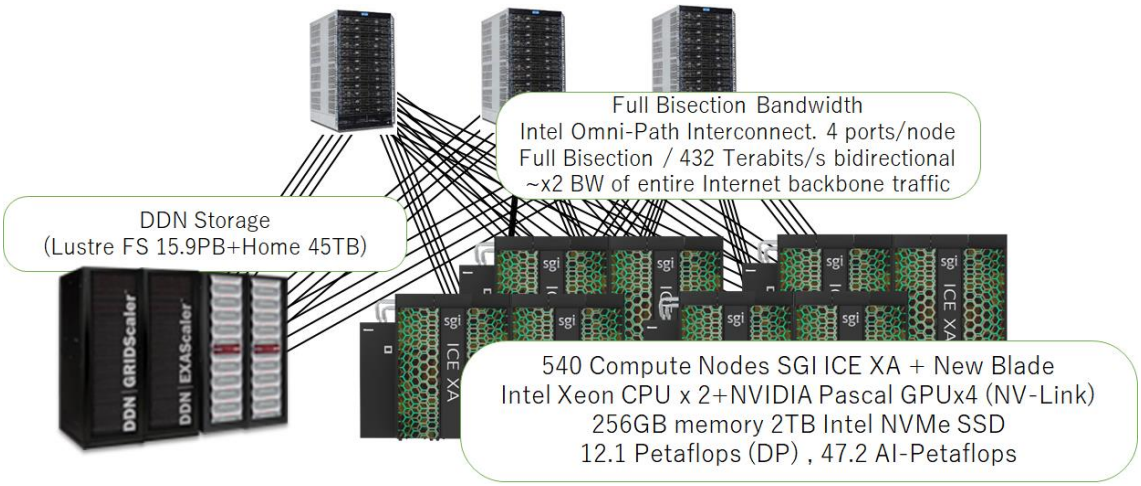


Figure 25 TSUBAME3.0 システム概要

2.2.5.2.　冷却設備概要

TSUBAME3.0 の冷却を中心とした設備概要を以下に示す.

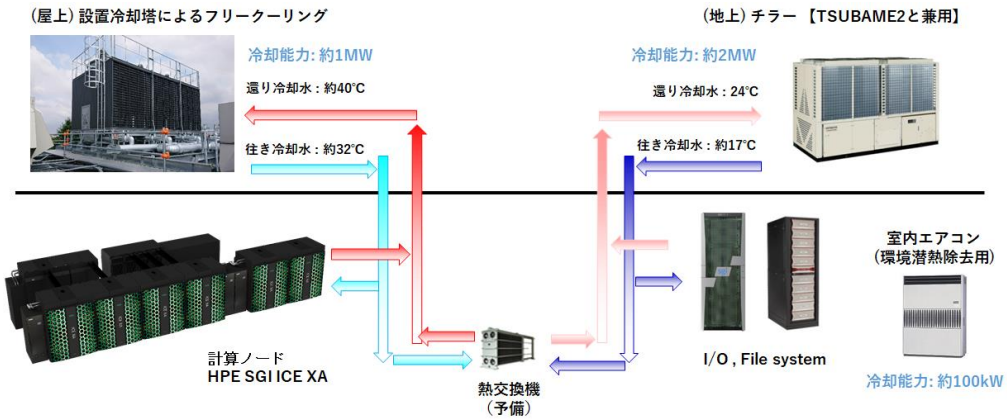


Figure 26 TSUBAME3.0 冷却設備

通常運用中の冷却水の流量は，おおよそ下記の通りである：

- 冷却塔⇔計算ノード (上図左側のループ)：約 2000 l/min
- チラー⇔I/O 等(上図右側のループ)：約 200 l/min

冷却能力の多くを担うのは冷却塔による 30℃台の高温の冷却水であり，フリークーリングを用いることにより省エネを目指す設計である．一方で，真夏の外気温が極めて高い場合に冷却能力が不十分となるおそれを考慮して，より水温の低いチラーによる冷却水との熱交換器を設けた(上図下部中央)．幸い，ほぼ 2 年間の運用期間において，この熱交換が必要となったことはない．

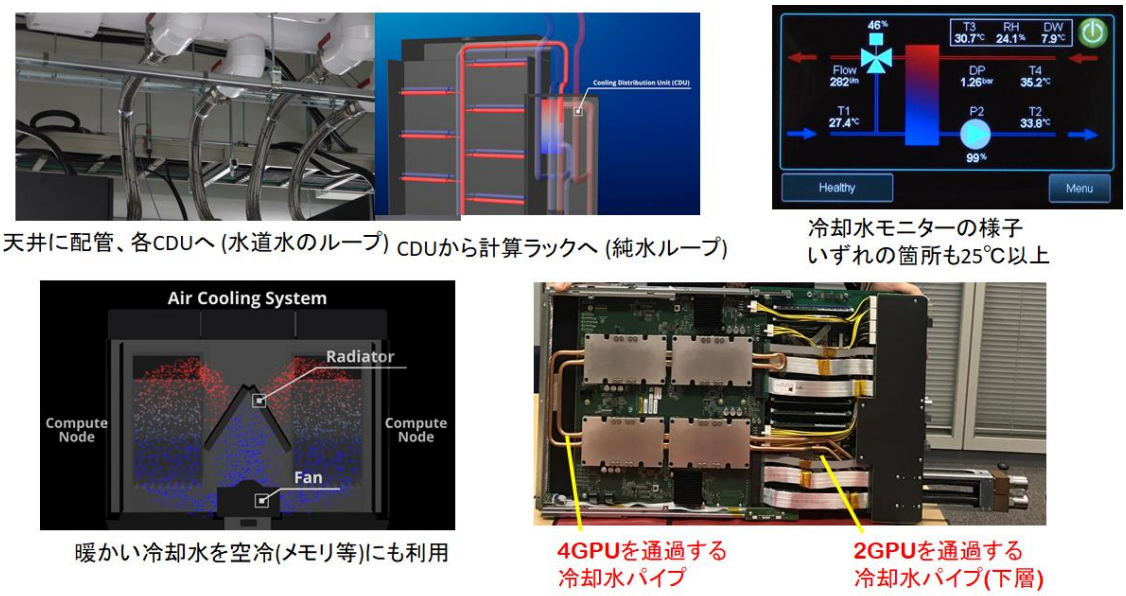


Figure 27 マシン室内の冷却設備

本図は，マシン室内でのより詳細な冷却整備を示す．ラック・CDU・計算ノードは高効率冷却のためにセットで設計されている．冷却対象となる計算ノードは，ラックあたり 36 台，システム全体では 15 ラック 540 台となる．

- 屋上冷却塔から来た冷却水は CDU に接続される．CDU は全体で 4 つ，1CDU あたり原則 4 ラックを担当する．
- CDU 内で冷却水は 2 つの役割を持つ：もう一段階の純水ループと熱交換，空冷のための空気と熱交換．
- 純水ループは各ラック内の各ノードへ接続し，重要な熱源である CPU と GPU を直接液冷する．
- 計算ノード内の他パーツ(メモリ，SSD など)は空冷される．

2.2.5.3. 消費電力と PUE

TSUBAME3.0 の消費電力は，通常運用時に 600kW 程度，Linpack 実行時等の事実上のピークは 1MW 程度である．計算ノードの電力は 3 相 415V により，それ以外のストレージ・管理ノード等は 200V 電源により供給される．

グラフは，TSUBAME3.0 の 23 か月間の電力推移を示す．チラー部においては電力計測器の都合で十分なデータがなく，現在の電力を基にした推定値となっている．またグラフにおいて 4 月/8 月に電力が下がっているのは定期メンテナンスの影響である．

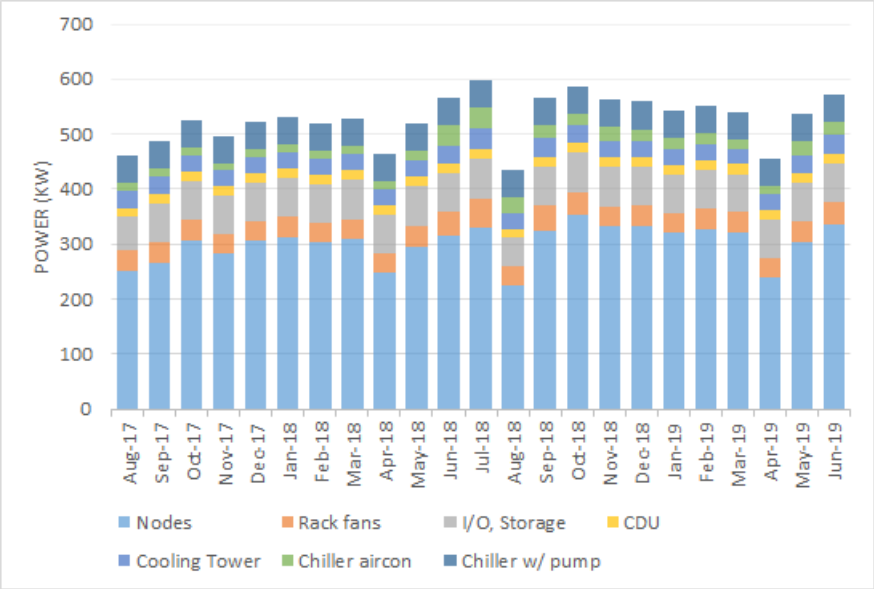


Figure 28 PUE の変遷

次に PUE の議論を行う．前節に示したように計算ノードと冷却設備が一体化しているために，PUE の定義が自明ではないが，今回は下記の通りとした：

- 計算ノード，ストレージに加え，ファンを持つラックの電力も IT 機器電力として計上した．例えば空冷計算機においては，筐体内のファンも IT 機器に含まれることが一般的であるため，この計上は許容されると考える．
- 冷却塔，チラーに加え，冷却水の熱交換器・ポンプを持つ CDU を IT 機器「以外」と計上した．

上記の前提においては，期間平均の PUE は 1.29 であった．また上記から，効率が相対的に悪いチラーによる冷却水ループを除いた場合の PUE は 1.14 となる．

2.2.5.4.　システムの変遷とエネルギー原単位削減の取り組み

GSIC では 2006 より TSUBAME スパコンシリーズを運用しており，順次エネルギー効率を改善してきた.

Table 3 システム変遷

システム名	TSUBAME1.0/1.1/1.2	TSUBAME2.0/2.5	TSUBAME3.0
システム稼動時期	2006/4~2010/10 (2007 年秋より 1.1, 2008 年秋より 1.2)	2010/11～2017/7 (2013/9 より 2.5)	2017/8～
ノード数	655	1442	540
理論演算性能	1.0: 85TFlops 1.1: 110TFlops 1.2: 160TFlops	2.0: 2.4PFlops 2.5: 5.7PFlops	12PFlops
システム全体の消費電力	平均約 1MW	平均約 1MW	平均約 0.5MW
運用時 PUE	1.44	1.28	1.29 (ノード部 1.14)
冷却方式	アイル分離方式空冷	ラック内空冷	直接水冷＋空冷
電力効率(Linpack 時)	0.04～0.08GFlops/W	0.9～3GFlops/W	14GFlops/W

スーパーコンピュータとしての電力効率として，Watt あたりの Flops (Linpack ベンチマーク時)について議論する．この指標は省エネスパコン世界ランキング Green500 で用いられる．なお Green500 に提出する値の算出時には，分母の電力に冷却やストレージ等の電力を含まない.

表に示されるように，TSUBAME の世代が変わるたびにこの値は大きく向上し，11 年間で 0.04GFlops/W から 14GFlops/W と 300 倍以上の向上を果たしている．さらには，それぞれ同世代の典型的なスーパーコンピュータに比較して高い水準を維持してきた．その表れとして，TSUBAME2.0 は Green500 において世界 3 位 (2010 年 11 月)，TSUBAME3.0 は世界 1 位 (2017 年 6 月)を達成した．それを実現した最大の要素は，GPU 等の電力効率のよいアクセラレータを世界に先駆けて大幅に活用してきたことである．さらに GPU のためのソフトウェア技術(独自 Linpack 含む)，ボトルネックを可能な限り排除するバンド幅重視のシステムデザイン技術，電力チューニング技術等の研究開発を継続し，上記の成果を得た.

2.2.5.5.　効率改善の取り組み

TSUBAME シリーズにおけるエネルギー効率改善の取り組みは大きく分けて，システム更新の際の設計面における取り組みと，システム運用中の取り組みの 2 つに分けられる.

設計面，主に冷却設備における取り組みは下記の通りである.

- TSUBAME1 における計算ノードは 4U 空冷サーバであった．それを空調によって冷却する必要

があったが，無駄を削減する取り組みの一環として，ホットアイルとコールドアイルを分離する手法とした．平均的な PUE は 1.44 であった．

- TSUBAME2 においては，TSUBAME1 より効率改善するために，部屋・アイル全体を冷却するのではなく，密閉されたラック(HP Modular Cooling System, MCS)の中のみを冷却する方式とした．計算ノードは空冷用のものであった．ラック内の吸気温度を 17℃程度とするために，冷却水は 10℃程度とする必要があり，冷却水を作るためのチラーによる消費電力が課題となった．平均的な PUE は 1.29 であった．
- TSUBAME2 の運用期間と並行して，先進的な冷却方式を実験するための小型スパコン TSUBAME-KFC を導入した．計算ノードは油槽に丸ごとつかり，油温を冷却塔で大気冷却された冷却水により下げる．詳細は論文"TSUBAME-KFC: a Modern Liquid Submersion Cooling Prototype towards Exascale Becoming the Greenest Supercomputer in the World" (IEEE ICPADS 2014)を参照のこと．
- TSUBAME3 では，TSUBAME-KFC の経験を経て，チラーの利用を抑制することおよび 30℃程度の高温液体により直接冷却することを主眼とした．その結果，上述のような CPU/GPU を直接液冷する方式とした．約 2 年間の計算ノード部の PUE は 1.14 である．ストレージ等の冷却のために，TSUBAME2 時代から用いているチラーを使っているため，システム全体 PUE は 1.29 に落ちる．

運用中の取り組みについて，TSUBAME2 時代のものを述べる．

- 運用開始時期には，チラーで冷却する水温を 10℃程度としていたが，これは保守的であり，順次水温を上げる取り組みを行った．最終的には水温は 17℃程度，ノード吸気温は 24℃程度とし，これでもシステムは問題なく動作した．
- 大学から夏季昼間のピーク消費電力を下げるよう(上限 800kW 程度)要請があったため，昼夜で目標電力を変えるピークシフト運用を行った．

2.2.5.6. 課題

現在の TSUBAME3.0 は，過去のシステムに比べてエネルギー効率は改善しているが，下記の課題がある．

- ストレージ等を冷却するチラー関係の消費電力が大きい．ここではチラー方式そのものの問題もあるが，TSUBAME2.0 時代から用いているチラーがオーバスペックである点も影響が大きいと考える．この冷却能力は全体で 2MW，ユニット当たりで 250kW である．通常は 1 ユニットのみが稼働していることを確認しているが，それでも平均電力 70kW 程度のストレージの冷却のためにはオーバスペックである．
- 計算ノードの冷却設備も，能力とのミスマッチの可能性がある．本冷却システムはラックあたり 60kW，計算ノード群全体で 900kW の冷却能力を持ち，これはピーク時の消費電力をもとに設計されている．しかし前述のグラフのように，計算ノード群の消費電力は平均運用時には

300kW 程度であり，ピーク時の 1/3 程度である．このような場合に，冷却設備自体の消費電力も，ピーク時の 1/3 に近づくことが望ましい．

将来のシステムに向けた，設計上の課題を述べる．

- TSUBAME3.0 の方式において，チラーを用いる部分がエネルギー効率を低下させている．産総研 ABCI システムのように，ストレージも温液冷却とすることにより改善すると期待されるが，ストレージの故障率への影響の可能性がある．

3. 効率的エネルギー活用に有効な事例

本章では，前章に挙げた課題に対して，解決や対策に有効な技術と事例について紹介する.

3.1. 技術

効率的なエネルギー利用および省電力には，各要素技術を改善するとともに，それらを適切に組み合わせることが重要である．本節では，そうした効率的なエネルギー利用や省電力に向けて，WG メンバーから提供された事例を列挙する．それぞれの詳細は Appendix1 に記載する.

- 半導体製造ばらつきを考慮した電力制御
- 自然エネルギーを活用した空調システム
- 予測技術を活用した空調制御システム
- ストレージと温度の関係
- 電力チューニング
- 縮退運転（自動ノード停止）

3.2. 制度

3.2.1. 再生可能エネルギーの購入について

省エネ法により，年間 1%のエネルギー原単位の削減が求められている．また，東京都条例により，温室効果ガス排出力削減の取り組みがなされている.

3.2.1.1. 東京都条例における排出量取引

（参考：排出量取引入門，東京都環境局，平成 30 年 12 月）

東京都では，あるべき姿として「省エネルギー・エネルギーマネジメントの推進により，エネルギー利用の高効率化・最適化が進展し，エネルギー消費量の削減と経済成長が両立した，持続可能な都市が実現している．産業・業務部門においては，事業者規模の大小にかかわらず，設備機器の効率的な運用・高効率化が進むとともに，低炭素なエネルギーの選択行動がとられている．」を掲げ，2016 年 3 月「東京都環境基本計画」において，次の目標を設定している.

- 温室効果ガス排出量　　：2030 年までに，30%削減(2000 年比)
- エネルギー消費量　　：2030 年までに，38%削減(2000 年比)
- 再生可能エネルギー　　：2030 年までに，電力利用割合 30%程度に

2016 年度速報値によれば，東京都 CO2 排出量(6,006 万トン)の内訳は次の通り.

- 業務・産業部門　51%
- 家庭部門　28%
- 運輸部門　18%
- 廃棄物　3%

業務・産業部門のうち大規模事業所(約 1,200)に対しては，総量削減義務が課され，排出量取引制度が設定されている.

● 排出量取引制度の概要

[制度概要]

- オフィスビル等を対象とする世界初の都市型のキャップ＆トレード制度
- 高効率機器への更新や運用対策の推進等，自らの事業所で削減対策を実験
- 自らの削減対策に加え，排出量取引での削減量の調達により，合理的に対策を推進すること
  ができる仕組み
- 大規模事業所間の取引に加え，各種クレジットの活用が可能

[削減計画期間]

削減計画期間は 5 年毎に設定されている.

- 第 1 期計画期間：2010～2014 年度
- 第 2 期計画期間：2015～2019 年度
- 第 3 期計画期間：2020～2024 年度

各計画期間終了後，1 年 6 か月間の整理期間の後，履行期限を迎える.

[削減義務率]

約 1,200 の対象事業所を 3 つの区分に分け，削減義務率を設定している.

Table 4 削減義務率

区分			第 1 期計画期間	第 2 期計画期間
Ⅰ-1	オフィスビル等※ 1 と地域冷暖房施設（「区分Ⅰ-2」に該当するものを除く.）	(※ 1) オフィスビル，官公庁庁舎，商業施設，宿泊施設，教育施設，医療施設等	8%	17%
Ⅰ-2	オフィスビル等※ 1 のうち，他人から供給された	事業所の全エネルギー使用量に	6%	15%

	熱に係るエネルギーを多く利用している事業所	占める他人から供給された熱に係るエネルギーの割合が 20%以上		
Ⅱ	区分Ⅰ-1，区分Ⅰ-2 以外の事業所	工場，上下水施設，廃棄物処理施設等	6%	15%

削減義務量は次の式で計算し，５年間の排出量を下記で定まる排出量上限以下にする.

削減義務量＝標準排出量×削減義務率×削減義務期間
排出上限量＝（標準排出量－単年度削減義務量）×削減義務期間

[排出量取引の位置付け]

削減対策の実施を排出量取引よりも優先するよう定めているが，一方で，排出量取引を実施する必要があると判断した場合には，計画的な取得に努めることを求めている．つまり，排出量取引は削減不足が確定してから検討するのではなく，早い段階から組織的な検討体制を構築して準備を進める必要がある.

[クレジット]

クレジットとは，削減対策の実施により得られる温室効果ガス削減量のことで，次の 5 種類がある.

1. 超過削減量
2. 都内中小クレジット
3. 再エネクレジット
4. 都外クレジット
5. 埼玉連携クレジット

[削減量口座簿]

排出量取引の結果は，東京都が管理する「総量削減義務と排出量取引システム」という電子システムに記録する．口座簿の記録は申請等に基づき東京都が行う．口座簿には(1)指定管理口座と(2)一般管理口座がある.

指定管理口座とは，知事が指定地球温暖化対策事業所の指定を行う際に，職権で開設される口座であり，削減義務の履行状況を管理するもの．指定管理口座に記録される数値は，対象事業所の排出状況を示す数値になる.

一般管理口座とは，事業者からの申請に基づき開設する口座であり，クレジットを売却，購入する際(排出量取引)に開設が必要となる.

[排出量取引の実際]

排出量取引を行うためには，次の 4 つのステップを実施する．

1. 削減量の確認
2. 口座の開設
3. 取引先の確保
4. 計画的な取引の実施

東京都の排出量取引は相対取引であり，取引価格は取引する当事者同士の交渉・合意により決定される．取引価格に対する上限価格，下限価格等の制約はない．

クレジットの販売先や購入先を見つけ方には，いくつかの方法がある．

1. 電子システムの「見積受付登録事業者照会」を利用する方法

電子システム内にある掲示板で，クレジットを売りたい，買いたい者が取引相手を探すために，自らの情報を登録できるシステムである．

2. 民間のクレジット仲介業者，グリーンエネルギー証書の発行事業者を利用する方法

排出量取引セミナーに出店したことのあるクレジットの販売・仲介を行っている業者の情報を東京都の Web サイトで公表している．

3. 公表データを利用する方法

「排出量取引実績等の情報」や「計画書のデータ」を参照し，購入先や販売先の候補を検討する．

前述の通り，取引価格は取引する当事者同士で決定されるが，東京都環境局の Web サイトには，取引価格の参考値が公表されている．

[バンキング]

削減計画期間中に削減対策を実施し超過削減量やオフセットクレジット等を発行したものの，当該削減計画期間の削減義務の履行に利用しなかったクレジット等を，翌削減計画期間に持ち越すこと．バンキングは期日の到来とともに自動的に行われるため，手続きは不要．

- 総量削減義務と排出量取引システムについて

クレジットの量や取引履歴などの情報を記録し，管理する電子システムであり，インターネットを通じて，Web ブラウザ上で操作できる．口座開設者は，自らの事業所の義務履行状況のほか，自分が開設した口座に記録されているクレジットの量や取引履歴などを参照できる．

- 利用時間　：開庁日（土日，祝日を除く）9:00 から 18:00 まで
- 利用料　　：無料

- 基準排出量の変更について

基準排出量が決定した特定地球温暖化対策事業所であって，次の 1，2 に掲げる要因による排出量の増減量として算定される量が，1 は変更部分における排出量の増減量の合計が基準排出量の 6%以上，2 は供給先の床面積の増減量が基準年度における供給先の床面積の 6%以上である場合に，基準排出量の変更を行うことになっている．注意すべきは，増加だけでなく，減少した場合も変更を行う必要があること．

1. 熱供給事業所以外の事業所で，次の①から③のいずれかに該当する場合．
   ① 床面積が増減した場合
   ② 用途が，排出活動指標に定める用途のうち異なる用途に変更した場合
   ③ 事業活動の量，種類又は性質を変更するための設備が増減した場合

2. 熱供給事業所において，熱の供給先の床面積(住宅用途を含む)が増減した場合．

JAXA においては，JSS1(Jaxa Supercomputer System Generation 1)設置時に，既設建屋を取り壊し，スーパーコンピュータ棟の建築を行った際に，床面積の増加に伴う基準排出量の変更を行っている．

3.2.1.2. 低炭素電力事業者からの電力購入

排出量を算定する際に使用する CO2 排出係数が小さい電気事業者から電力を購入することで，排出量削減を実現することもできる．例えば，2019 年 10 月に JAXA が調査したところでは，東京電力の排出係数(t-CO2/千 kWh)は，0.455 であるが，低炭素電力電気事業者 A，B，C の排出係数は，それぞれ，0.229，0.192，0 であった．C 社と電力購入の契約をすれば，計算上 CO2 排出量がゼロになる．その一方，電力の単価は上がるため，年間電力使用量が変わらない場合でも，年間数億円の電気代増加となるという試算であった．

4. 業務効率や生産性を考慮したエネルギー効率の指標の検討

2.1 で述べた通り，「経済活動量」としてシステムを運用することによる価値，あるいはシステムが提供する価値を利用することが課題である．そこで本 WG では，情報システム，特に HPC システムの「経済活動量」や価値に関して議論を行った．

4.1. HPC システムの「経済活動量」および価値について

HPC システムの「経済活動量」および価値の考え方については，システムのハードウェア・ソフトウェアによる価値はもちろん，それ以外にシステム上で動作するアプリケーションの性能や研究の成果なども考慮することが望ましい．しかし，それら多様な観点を総合的に考慮して一つの指標としてまとめ上げることは非常に困難である．また，システムの特性によってどういった観点を考慮する必要があるかが異なる場合があるということも，まとめ上げる困難さをより増大させている．

4.2. 「経済活動量」として用いる新たな指標について

こうした状況のなか，WG の議論の中では新たな指標として，いくつかの候補が挙がった．

■ 使用量算出に影響：分子

指標	エネルギー原単位の傾向	メリット	デメリット・課題
再生可能エネルギー利用率	再生可能エネルギー利用状況変化時に変化する	・使用量（分子）を低減できる	・利用が現実的でない場合や進歩が遅い場合は，毎年の改善は難しい

■ 経済活動量算出に影響：分母

指標	エネルギー原単位の傾向	メリット	デメリット・課題
FLOPS	システム更新時に大幅に変化する	・システム更新時に大幅に低減できる見込みがある	・システム更新時以外は変化しないため，毎年改善するには他の観点で改善する必要がある
スパコン利用で得られた効果（成果報告書，レポート）	成果公表時に変化する	・利用者が多い（多様な成果が出やすい）システムではより低減できる	・成果は運用開始から1-2年経過時点くらいがピークでありその後の改善がどんどん難しくなる可能性がある ・どう測るか，どう数えるか ・科学的なインパクトの大きい研究を優遇して成果を出すことは効率的とは言えない可能性がある ・研究の価値の定量化が困難
スパコン非利用時と利用時の業務効率／生産性の差	業務効率改善時に変化する	・経済活動量の観点に近い	・業務効率／生産性の定義と測定 ・継続的に改善できるか
ジョブのアベンド率（業務効率）	ジョブのアベンド率改善によって変化する	・システムの改善（ハード，ソフト）によって低減できる	・データ分析が必要で，算出が容易でない場合がある
アプリケーションの効率	効率向上時に変化する	・FLOPS/Watt偏重（GPGPU偏重）を是正できる	・ターゲットアプリの選定 ・効率の定義
専用システムの専用領域での性能／汎用システムの汎用的な性能	ベンチマーク結果改善時に変化する	・システムの長所を活かした指標にできる	・公平性の担保 ・継続的な改善のためにはベンチマークを取り続ける必要がある
単位演算処理量に対するターンアラウンドタイム	システム更新やスケジューリング性能向上，ジョブ運用改善の際に変化する	・システム更新時の他，運用改善にて低減できる見込みがある	
システム利用者への提供資源量とFLOPSの積	システム更新時や実行ジョブ数増加（システム停止時間短縮）時に変化する	・システム更新時の他，運用改善にて低減できる見込みがある	
ユーザ数	システム利用者増加時に変化する	・集計が容易 ・継続的に低減できる可能性がある	・登録さえすれば簡単に増やせる
FLOPSとジョブが使用するコア数の積の総和	システム更新時や実行ジョブ数増加（システム停止時間短縮）時に変化する	・システム更新時の他，運用改善にて低減できる見込みがある	
同時実行可能な演算器数（FLOPS+整数演算）	システム更新時に大幅に変化する	・システム更新時に大幅に低減できる見込みがある	・システム更新時以外は変化しないため，毎年改善するには他の観点で改善する必要がある

Figure 29 エネルギー原単位の指標見直しに関する検討結果

議論の結果，表中の黄緑で示したような指標を用いるのが良いのではないかという結論に至った．具体的には，「計算性能」という一般的な名称とし，実際にどの値を使用するかは各サイトの運用者がシステムの特性に合わせて複数から選択できるようにすることで，個々のシステムにより適したエネルギー原単位を算出することができるようになる．

5. 情報システムにおける効率的なエネルギーの利活用の検討

2.2 で示したように，HPC システムを運用するほとんどの機関ではエネルギーの効率的な利用のための活動を継続的に実施している．ここでは，そうした各機関のこれまでの取り組みについて整理し，これから改善を試みようとする HPC システム運用機関の参考となる，実践済みのノウハウとしての一覧表記について検討した．以下に一覧を示す．

Table 5 効率的なエネルギー利活用のための実践済みノウハウ一覧

観点	手法	説明
冷却方式の変更による改善	空冷から水冷への変更	空冷より冷却効率の良い水冷方式の採用
	液浸の採用	水冷より冷却効率の良い液浸方式の採用
	ホットクーリング（冷却温度：XX～XX℃）	従来より高い温度でも安定動作する（特に液浸の）ハードウェア／製品の採用
	ウォームクーリング（冷却温度：XX～XX℃）	従来より高い温度でも安定動作する（特に液浸の）ハードウェア／製品の採用
	熱源側冷却方式の変更	大気冷却の採用
設備更新・運用方法の変更による改善	空調機設備の更新	同じ冷却能力をより低い消費電力で実現できる空調機の採用
	空冷エリア縮小	必要な範囲のみ冷却
	冷却温度変更	より高温での冷却
	ホットアイル／コールドアイル分離	熱の流れの整理
	冷却塔の冷却能力改善	風通しの向上
	蓄熱槽の活用	急激な冷却需要の吸収
システム更新・運用方法の変更による改善	電力モニタリングと緊急ジョブ自動停止	システム全体あるいは特定部分の電力を常時監視し，超過が発生しそうな場合にジョブ（プログラム）実行を自動的に停止する仕組みの導入
	熱効率の高いプロセッサの採用	同じ計算をより低い消費電力で実行できるプロセッサの採用
	ジョブスケジューラ連携による待	しばらく使わない計算ノードは

	機ノード電源停止	電源を停止しておき，使用するタイミングで自動的に起動する仕組みの導入
	電力キャッピング	消費電力が上限を超過しないようハードウェア／ソフトウェア的に制御する仕組みの導入
	コールドストレージの活用	参照頻度が少ないデータを少ない消費電力で記憶可能なコールドストレージ（テープ装置など）の採用
	ピークシフト運用	電気料金の安価な時間帯を有効活用できる仕組みの導入

6. 提言

これまでに述べた WG での議論から，今後の情報システム（HPC システム，データセンター）の効率的なエネルギー活用に関して提言する．

6.1. エネルギー原単位の指標の見直しについて

2.2.1.5 および 4 で述べた通り，データセンターや HPC システムにおいては，従来のエネルギー原単位の算出式に用いられている指標は実態に即していない状況となっている．これを改善するため，本 WG としてはエネルギー原単位の指標の見直しを提案したい．

具体的には，現在の計算式の分母（経済活動量）を「計算性能」とし，4 に挙げたような値から各機関がシステムの特性に応じて選択できるようにするべきである．

6.2. エネルギー原単位削減の対象期間について

エネルギー原単位の削減目標は，ある年度を基準に毎年 1%ずつとなっている．しかし，4.2 で示したような指標の見直しが実際になされる場合，数年ごとに実施されるシステム更新によって原単位は大きく改善することが想定される．また，情報システムの設備の更新は十年程度の間隔である．これらの更新タイミングで，その時代の技術レベルで可能なエネルギー利用効率化策を積極的に講じることで，長期的に見ると単年で 1%削減するよりも大きな効果が得られることと期待される．これらのことから，エネルギー原単位削減の対象となっている全事業者に対して一律に単年度での削減を求めるのではなく，事業の特性に応じた期間と削減率を運用者・事業者が選択できるようにすることが望まれる．

6.3. 冷却設備電力の削減について

これまで述べてきたように，現状情報システムには冷却のための設備や仕組みが必要不可欠である．これは，現在の情報システムを構成する部品が以下のような特性（課題）を持つためである．

- 動作するために電力を必要とし，それらは最終的に熱として排出される（使うと温度が上がる）
- 動作を保証された温度帯が低い（高温での動作を保証されていない）

また，一般に部品の故障率は低い方がよいとされており，部品の低故障率の維持のためにも冷却設備の必要性は高いままとなっている．さらに今後，システム規模がますます大きくなることで，システム・設備全体として消費するエネルギーも増大を続ける傾向にあると考えられる．

しかし，考え方を変え，これまでにない発想でシステムを設計・実装することで，大幅な冷却設備電力の削減を実現することができる可能性があると考える．例えば，以下のような冷却設備電力案が考えられる．

- 熱耐性が高く，高温での実行効率の高い IC チップの採用
  - より少ない冷却設備（冷却電力）で効率よく計算するチップを開発する

- 部品温度の上昇による故障率の増大を許容するシステム運用保守
  - 冷却設備に費用をかけない
  - 部品が故障する前提でノード資源等の割り当てを計画する
  - 部品の交換作業が容易なハードウェア設計とする

- 冷却不要な半導体素子，ストレージの開発
  - 冷却設備をなくす
  - 50℃以上の高温でも問題なく動作する部品を開発する
  - 熱がほとんど発生しないデバイスを開発する

7. おわりに

本報告書では，まず情報システムにおける効率的エネルギー活用の現状と課題を示し，それらに有効な技術と事例，ならびに同様の課題を抱える他業界の動向について紹介した．そして，情報システム，特に HPC システムの運用者としての観点から，より受け入れやすいエネルギー原単位の指標について検討した．また，情報システムにおける効率的なエネルギー利活用のために必要な技術や取り組みについても検討した．最後に，これらの検討結果を提言としてまとめた．これらの提言については，情報システム，特に HPC システムを運用する他組織にも広め，業界全体として改善を図っていきたいと考えている．

なお，今回は情報システムの中でも HPC システムに特化して議論を行ったが，同様に IT 機器で電力を多く消費するデータセンターにおいても，効率的なエネルギー利活用が重要である．そうした他業界との意見交換や共同での取り組みを実現することも，社会全体がエネルギーを効率的に活用出来るようになるために重要である．

Appendix1　事例

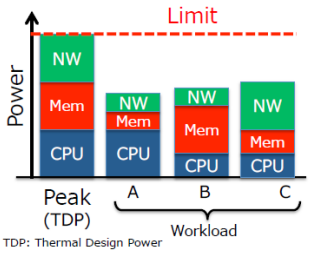
Appendix1.1.　半導体製造ばらつきを考慮した電力制御

Overprovisioned Systems

2

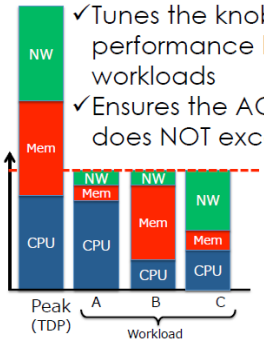
Under-provisioned (Conventional)

- HW Design
  - ✓Ensures the PEAK system power does NOT exceed the limit
- SW Design
  - ✓Tries to maximize the activity of HW components



Over-provisioned

- HW Design:
  - ✓Allows to install HWs w/o considering the power limit
  - ✓Provides power-performance knobs
- SW Design:
  - ✓Tunes the knobs to maximize the performance based on SW workloads
  - ✓Ensures the ACTUAL system power does NOT exceed the limit



Experimental Setup

3

- HPC Challenge: star DGEMM, star STREAM(Triad)
- NPB: BT, SP, EP
- Magneto Hydro-Dynamics(MHD) simulation
  - Typical stencil app. to simulate space plasma
  - Calculations and communications appear in turn
- Fiber benchmark suite: mVMC-mini (mVMC)
  - Variational Monte-Carlo simulation for strongly correlated electron system

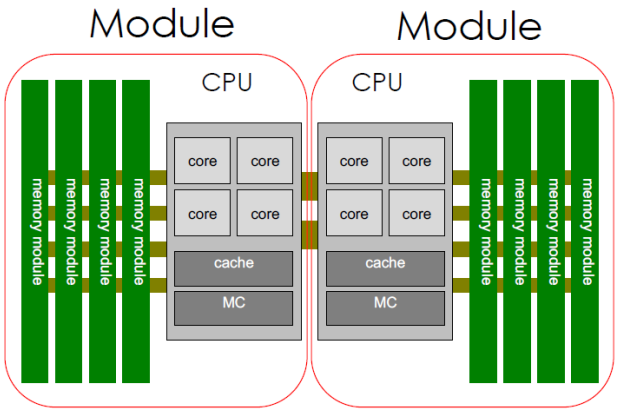
Blue=EP type
Red=With Comm. & Sync.

Site	Node Micro-Architecture	Total nodes	Procs. Per Node	Cores Per Procs.	Power Msrmt.
Cab(LLNL)	Intel E5-2670 Sandy Bridge	1,296	2	8	RAPL
BG/Q Vulcan (LLNL)	IBM PowerPC A2	24,576	1	16(compute)	EMON
Teller (SNL)	AMD A10-5800K Piledriver	104	1	4	PI
HA8K(Kyushu Univ.)	**Intel E5-2697v2 Ivy Bridge**	**965**	**2**	**12**	**RAPL**

KYUSHU UNIVERSITY

4

Terminology



CPU = Processor chip (including cores, cache, MC, etc.)
Module = A pair of a CPU and DRAMs directly connected to it

KYUSHU UNIVERSITY

5

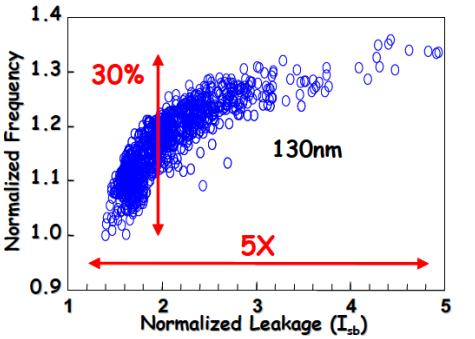
半導体製造ばらつきを考慮した電力マネージメント

KYUSHU UNIVERSITY

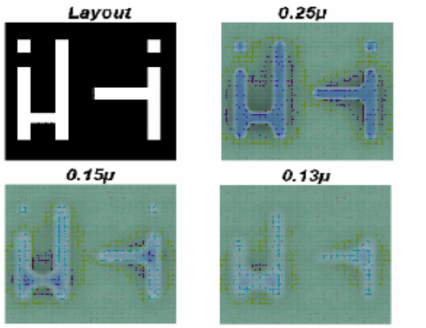
半導体の製造ばらつき

6

- イオン注入やエッチングにおいて非均一性が発生
- その結果，ゲート長やトランジスタ閾値がばらつく
- 微細化が進むにつれより深刻化



S. Borkar, Parameter variations and impact on circuits and microarchitecture, DAC, 2003.



[source: numerical technologies]

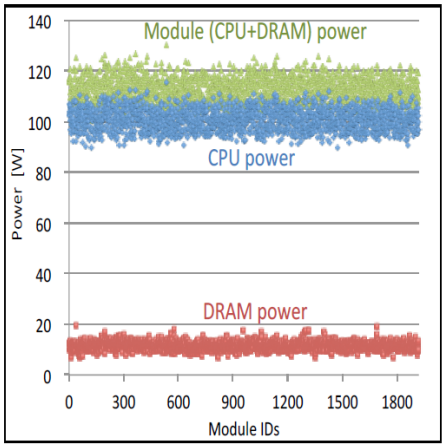
KYUSHU UNIVERSITY

Impact on CPU Frequency

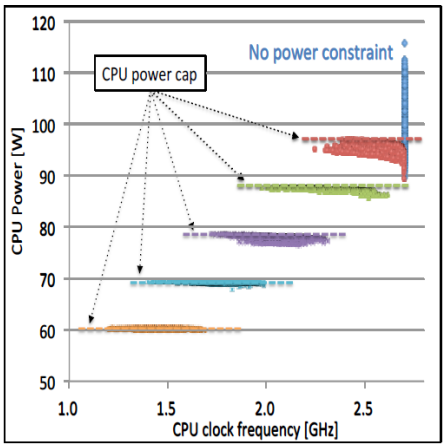
7

star DGEMM

w/ a uniform power constraint



30%

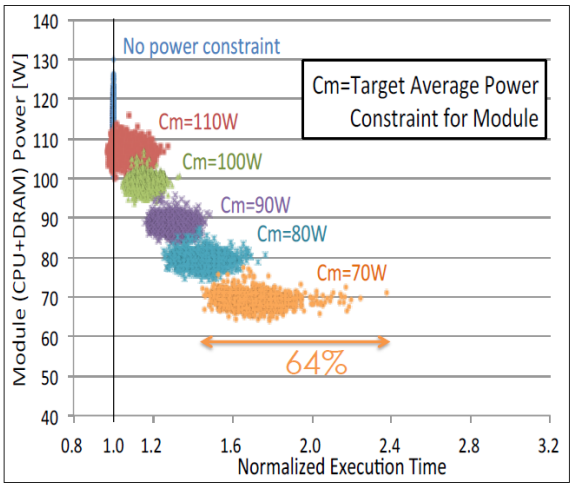


Power variation is translated into CPU frequency variation applying uniform power constraint!

8

Impact on Application Performance

star DGEMM　w/ a uniform power constraint



9

Problem and Goal

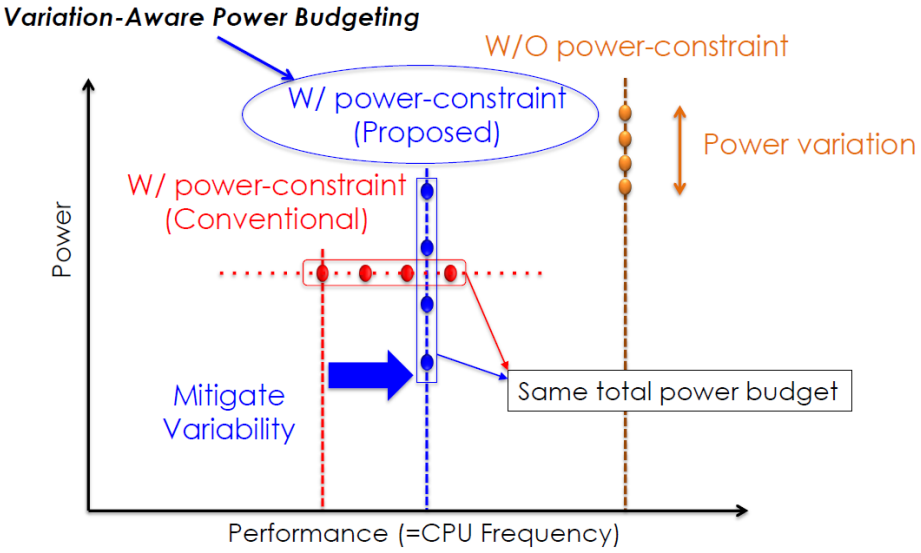
- Power-Constrained Supercomputing
  - will be applied to future HPC systems
- Manufacturing Variability
  - leads to performance variation under power constraint

Our Goal

Mitigate the impact of manufacturing variability on performance of HPC apps. under power constraint !

Concept

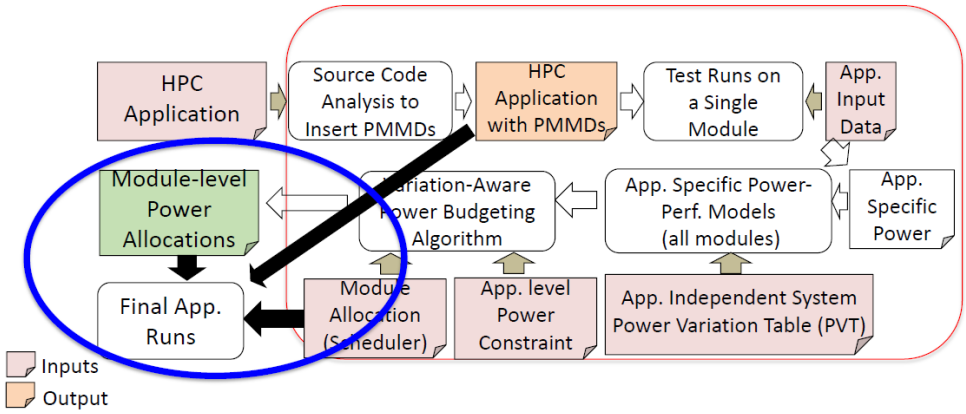
10

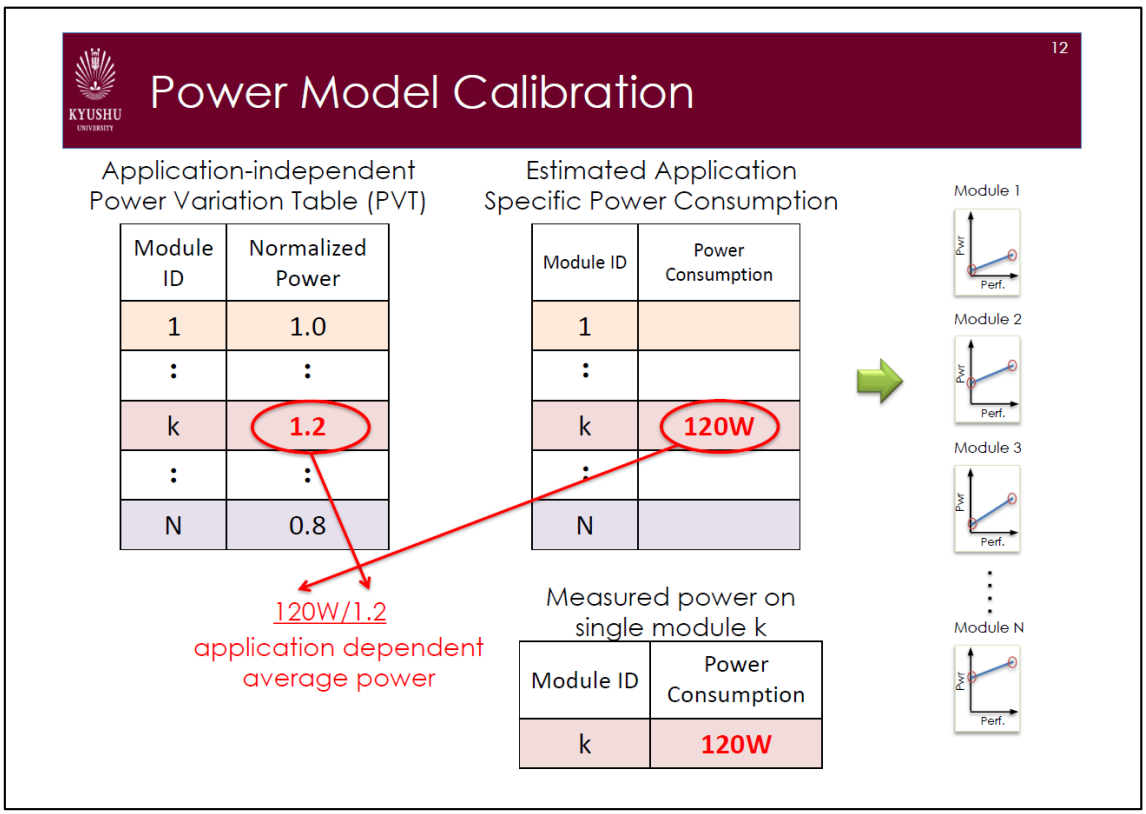


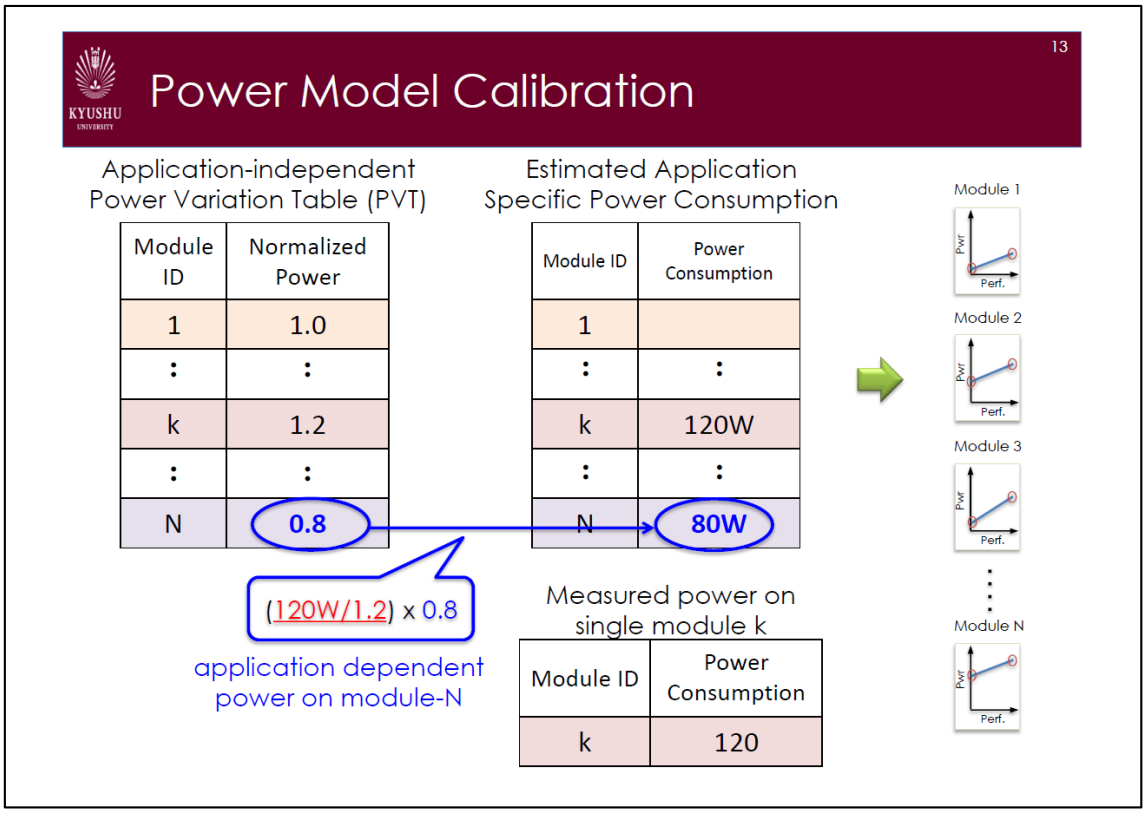
Variation-Aware Power Budgeting Strategy

11

How to manage power consumption

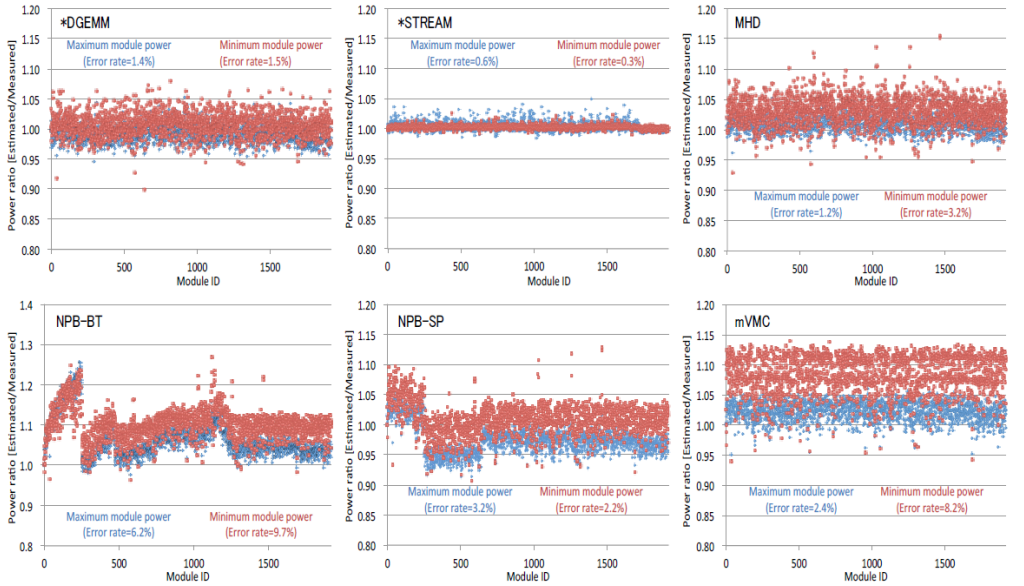



12
Power Model Calibration
Application-independent
Power Variation Table (PVT)
Module ID
Normalized Power
1
1.0
:
:
k
1.2
:
:
N
0.8
Estimated Application
Specific Power Consumption
Module ID
Power Consumption
1
:
k
120W
:
N
Module 1
Module 2
Module 3
Module N
Pwr
Perf.
120W/1.2
application dependent
average power
Measured power on
single module k
Module ID
Power Consumption
k
120W


13
Power Model Calibration
Application-independent
Power Variation Table (PVT)
Module ID
Normalized Power
1
1.0
:
:
k
1.2
:
:
N
0.8
Estimated Application
Specific Power Consumption
Module ID
Power Consumption
1
:
:
k
120W
:
:
N
80W
Module 1
Module 2
Module 3
Module N
Pwr
Perf.
(120W/1.2) x 0.8
application dependent
power on module-N
Measured power on
single module k
Module ID
Power Consumption
k
120

14

Power Estimation Accuracy



15

Options for Power Setting

Two options for power settings

- ***Power Capping*** (Pc) using RAPL
- ***Frequency Selection*** (Fs) using CPUFreqlibs

	Power Capping (Pc)	**Frequency Selection (Fs)**
Power Constraint	◎ Guaranteed	△ Not guaranteed
Performance Equivalence	△ Not guaranteed	◎ Guaranteed

16

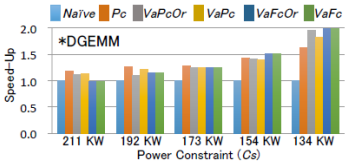
Tested Power Budgeting Methods

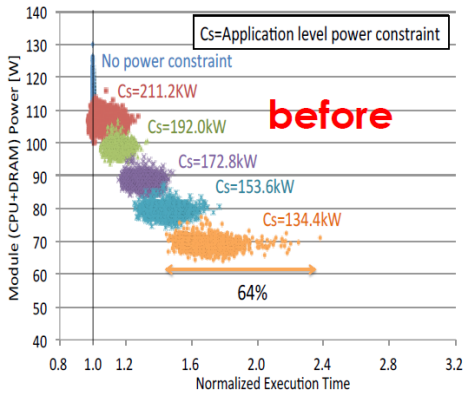
Method Name	Application Specific	Variation Aware	Power-Performance Model	Pwr. Set.
Naive	No	No	-----	Power Cap
Pc	Yes	No	Calibration	Power Cap
VaPc	Yes	Yes	Calibration	Power Cap
VaFs	Yes	Yes	Calibration	Freq. Sel.
VaPcOr	Yes	Yes	Oracle	Power Cap
VaFsOr	Yes	Yes	Oracle	Freq. Sel.

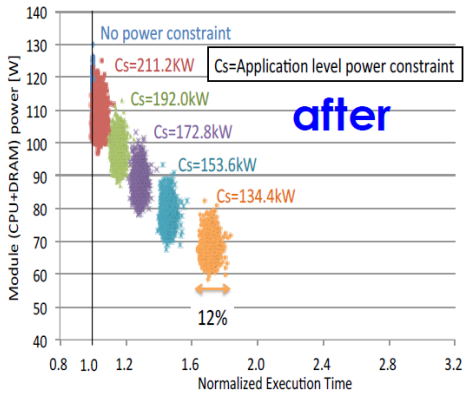
Va=Variation-Aware, Pc=Power Capping, Fs=Frequency Selection
Or=Observed power data are used

17

Speedup Ratios Normalized to Naïve (star DGEMM on 1,920 modules)



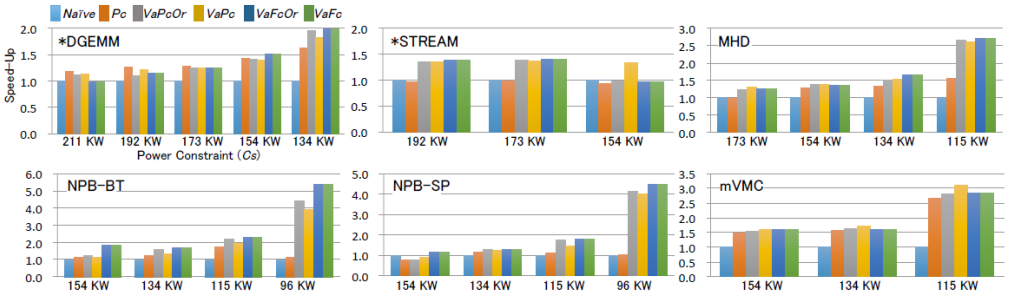




KYUSHU UNIVERSITY

18

Speedup Ratios Normalized to Naïve (All results on 1,920 modules)

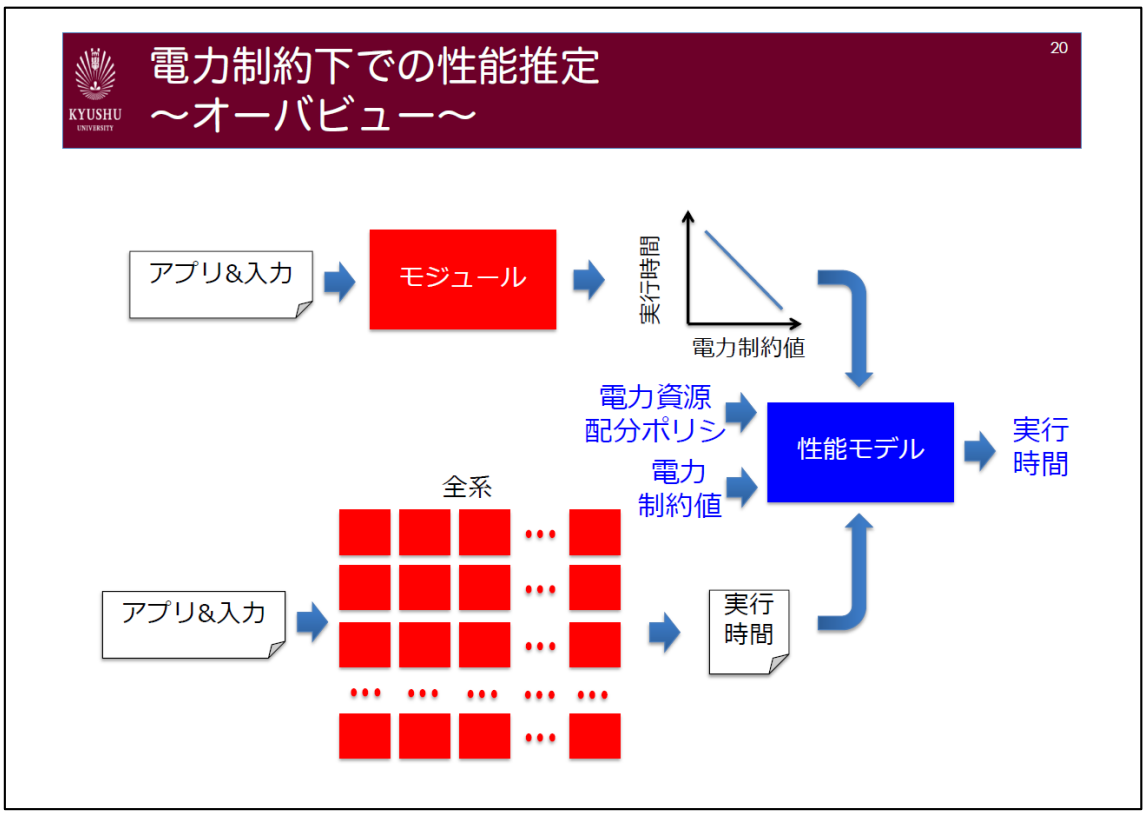


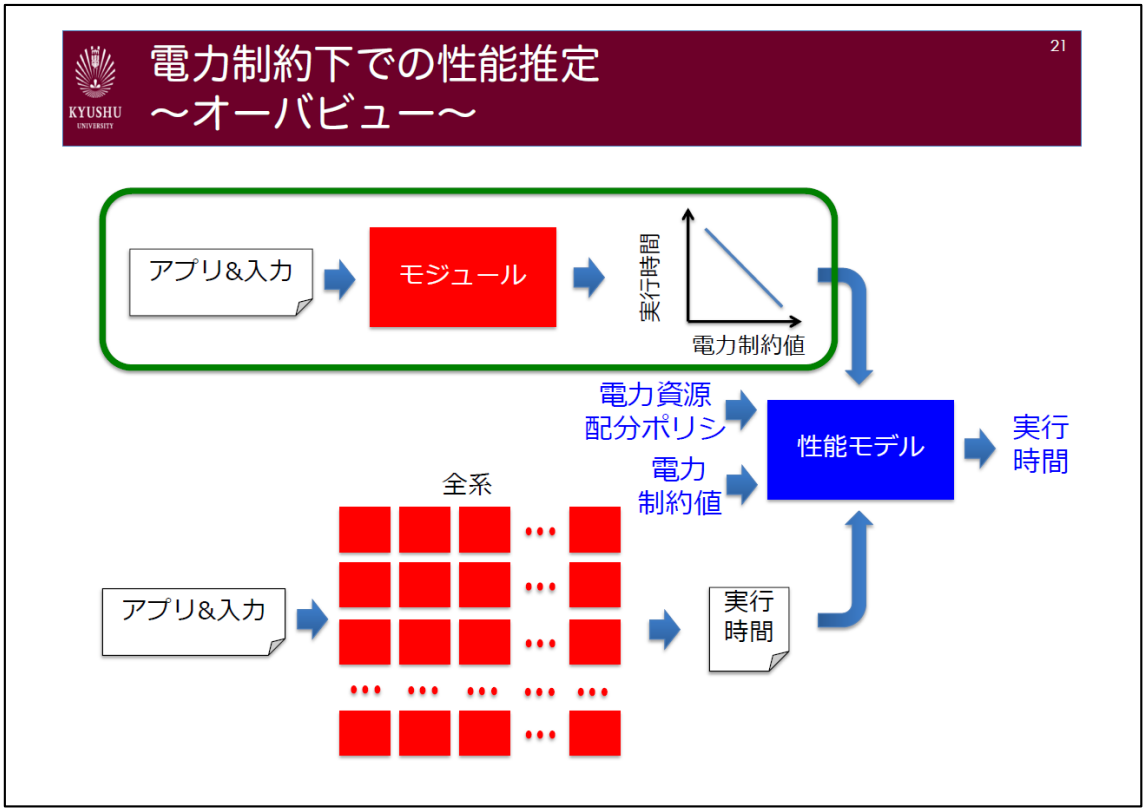
5.4X speedup at maximum (NPB-BT)
1.8X speedup in average

KYUSHU UNIVERSITY

19

システム性能電力推定


電力制約下での性能推定
～オーバビュー～
20
アプリ&入力
モジュール
実行時間
電力制約値
電力資源
配分ポリシ
電力
制約値
性能モデル
実行
時間
全系
アプリ&入力
実行
時間

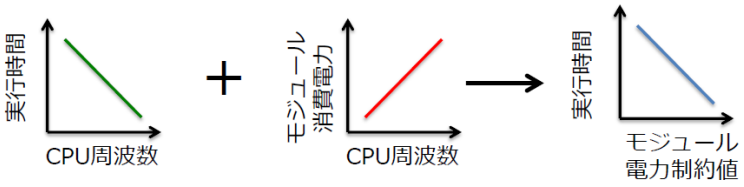

電力制約下での性能推定
～オーバビュー～
21
アプリ&入力
モジュール
実行時間
電力制約値
電力資源
配分ポリシ
電力
制約値
性能モデル
実行
時間
全系
アプリ&入力
実行
時間

モジュール電力性能特性のモデリング

22

2段階での特性モデリング

1.モジュール性能とCPU動作周波数の関係
2.CPU動作周波数とモジュール消費電力の関係

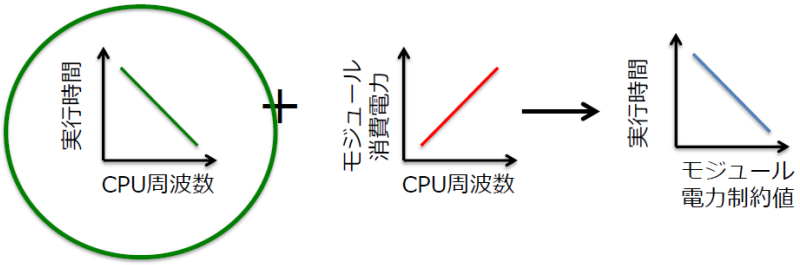


モジュール電力性能特性のモデリング

23

2段階での特性モデリング

1.モジュール性能とCPU動作周波数の関係
2.CPU動作周波数とモジュール消費電力の関係



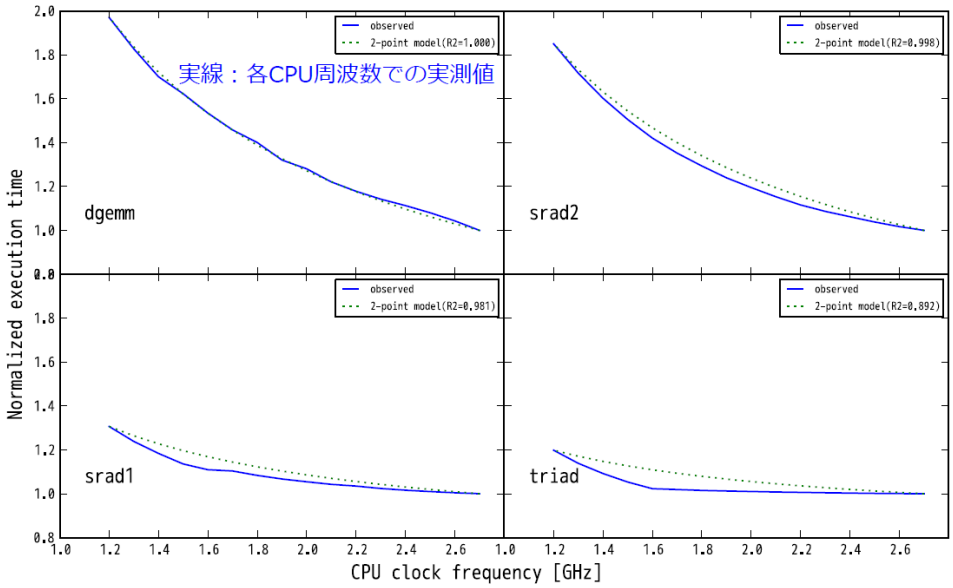
解決すべき課題

24

モジュール性能のCPU周波数依存性は
アプリ特性に依存する!

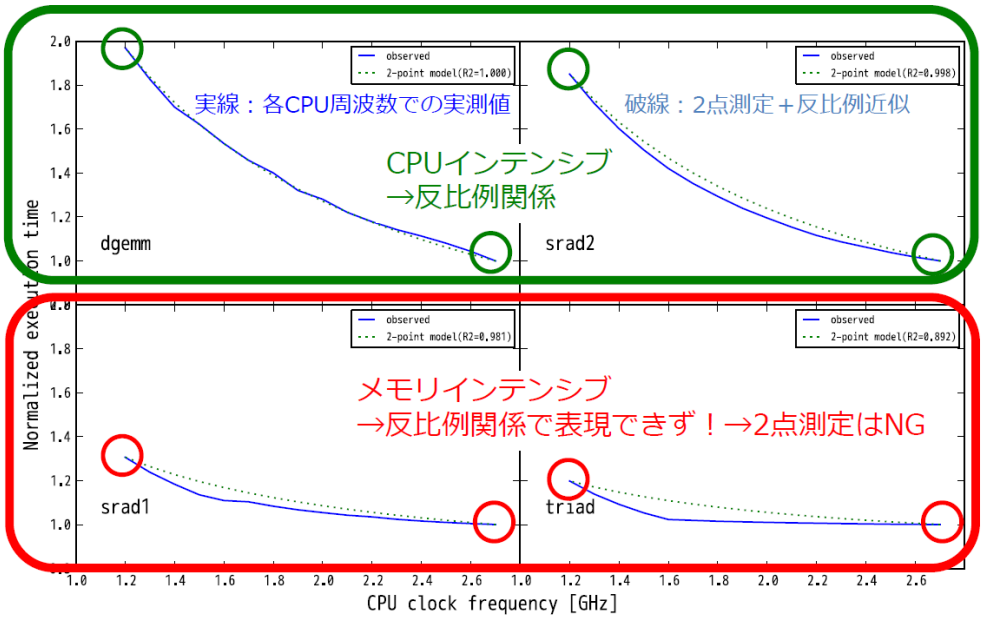
CPU周波数–実行時間の関係（1/3）

25



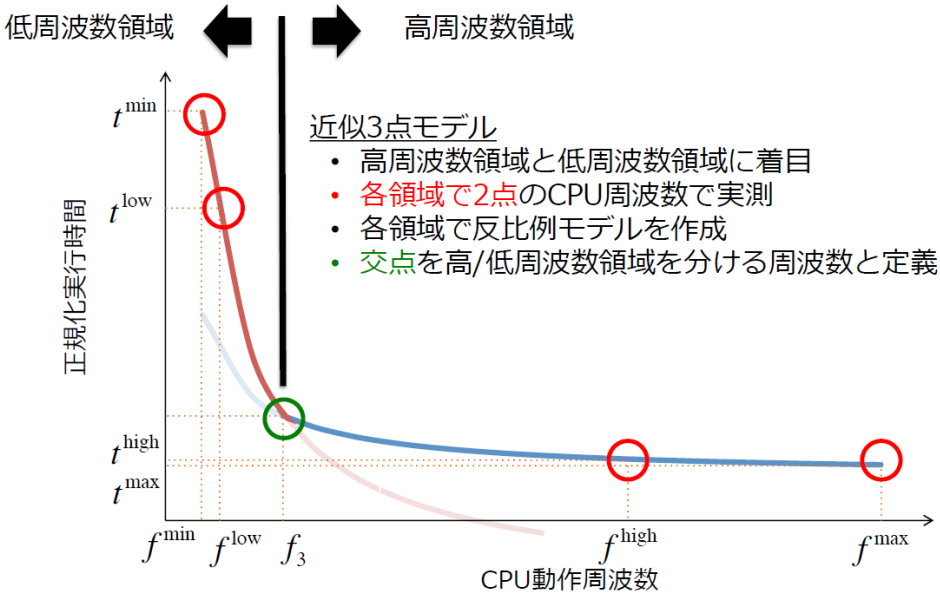
CPU周波数–実行時間の関係（1/3）

26



CPU周波数–実行時間の関係（2/3）

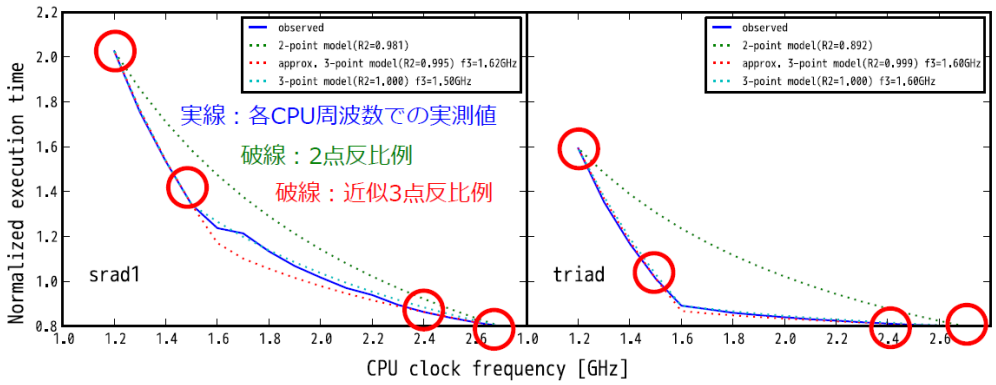
27



28

CPU周波数−実行時間の関係（3/3）

近似3点モデル（4回の事前実行）にて
精度良く実行時間を推定可能!

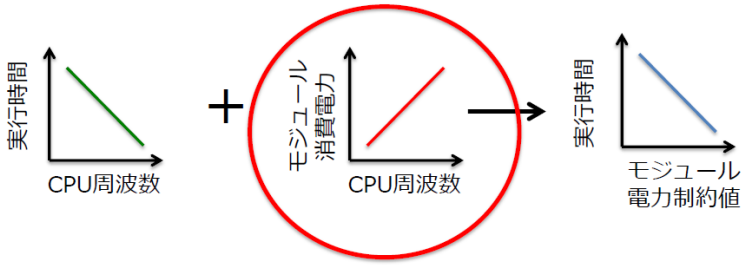


29

モジュール電力性能特性のモデリング

2段階での特性モデリング

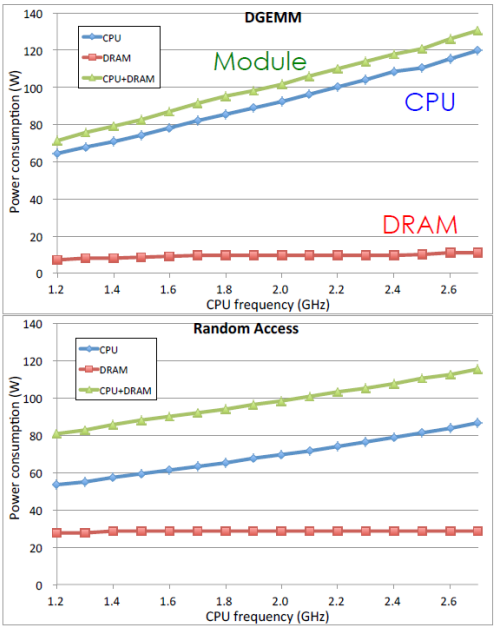
1.モジュール性能とCPU動作周波数の関係
2.CPU動作周波数とモジュール消費電力の関係



30

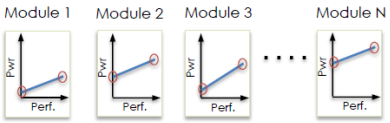
モジュール消費電力-CPU周波数の関係

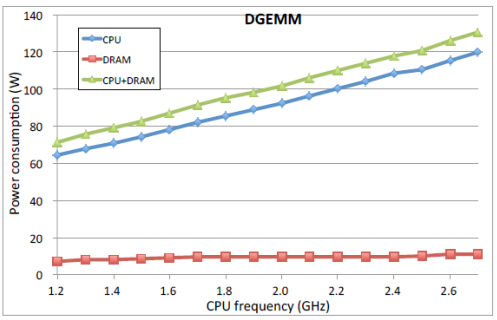
- CPU周波数とモジュール消費電力は線形関係
  - VFSを前提
  - DVFS対応は今後の課題
- アプリ特性によらず線形性を観測
- 最大/最小CPU周波数での電力測定でOK



31

解決すべき課題

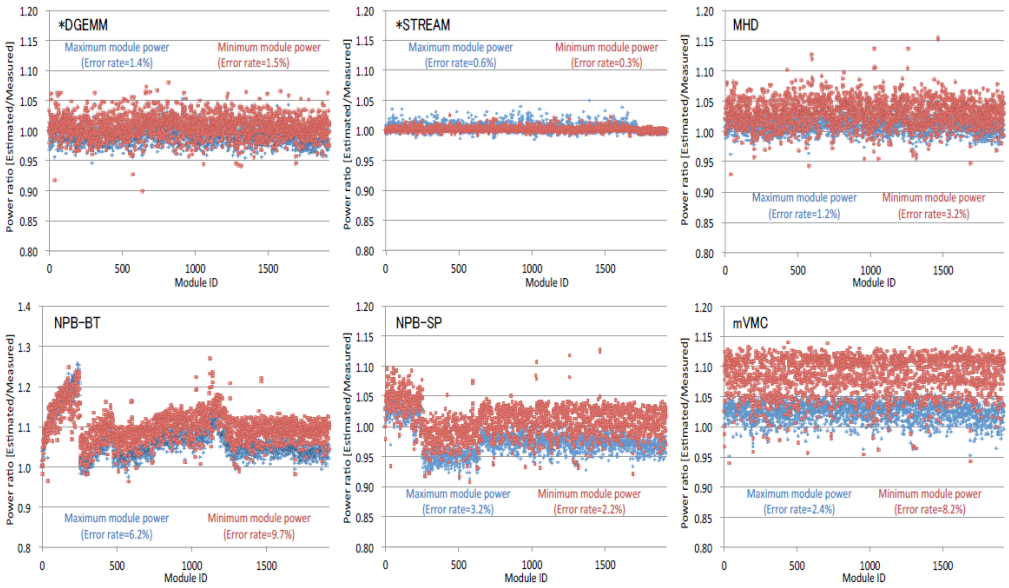




モジュール間消費電力-性能特性は
LSI製造ばらつきの影響を受ける!
→ Power Model Calibration を適用

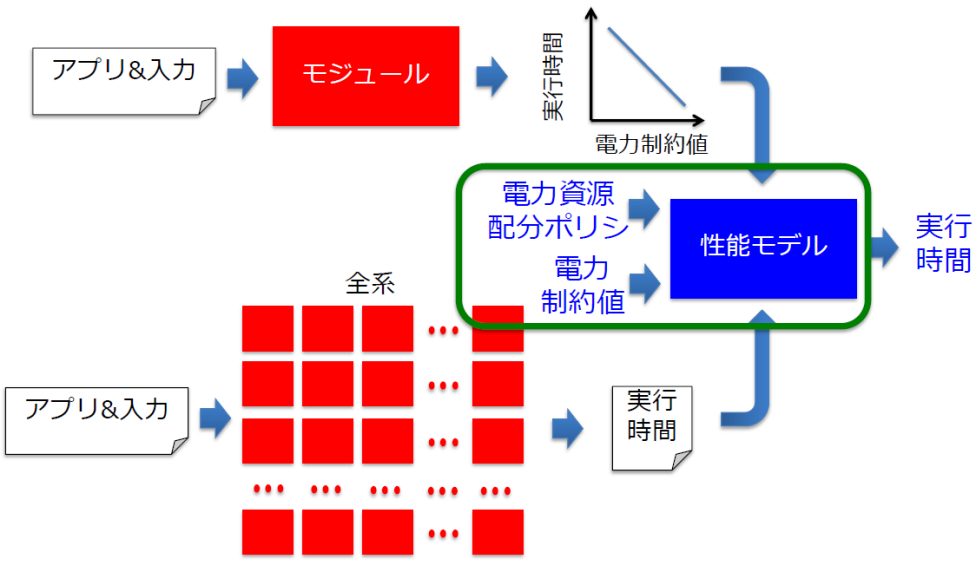
32

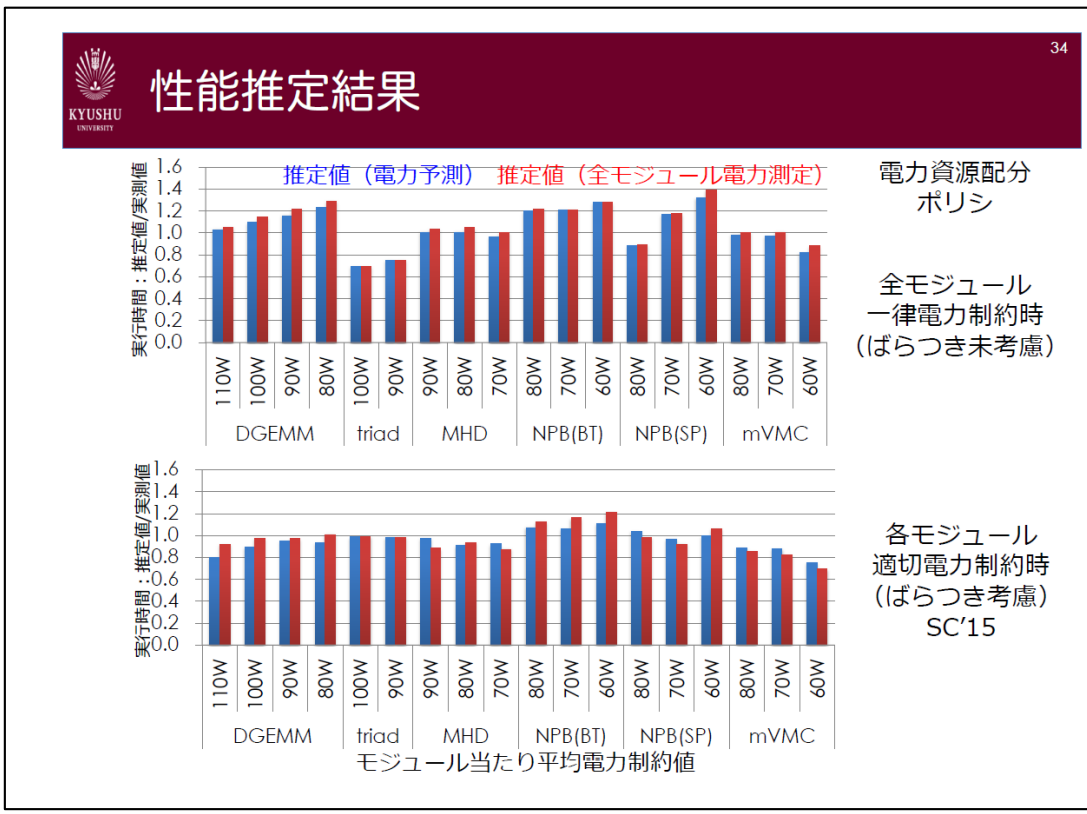
推定精度



33

電力制約下での性能推定 ～オーバビュー～




34
KYUSHU UNIVERSITY
性能推定結果
推定値（電力予測）
推定値（全モジュール電力測定）
実行時間：推定値/実測値
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
110W
100W
90W
80W
100W
90W
90W
80W
70W
80W
70W
60W
80W
70W
60W
80W
70W
60W
DGEMM
triad
MHD
NPB(BT)
NPB(SP)
mVMC
電力資源配分ポリシ
全モジュール一律電力制約時（ばらつき未考慮）
各モジュール適切電力制約時（ばらつき考慮）SC'15
モジュール当たり平均電力制約値

Appendix1.2.　自然エネルギーを活用した空調システム

本節では自然エネルギーを活用した空調システムについて，富士通研究所での開発例を紹介する．

Appendix1.2.1. センタの冷却方式とその特徴について

データセンタをはじめとした計算機センタでは，その利用用途や規模にあわせて様々な冷却方式が採用されている．下図は様々なセンタで採用されている冷却方式の種類と，関連する熱移動媒体や使用設備について記したものである．多くのセンタで採用されている空冷および水冷方式は計算機から出る排気と冷媒または水との熱交換が行われており，冷媒または水はそれぞれ空調機やチラーといった設備で放熱・冷却されるのが一般的である．

冷涼な外気を積極的に使用した冷却方式は外気冷却方式と呼ばれ，外気をセンタ内に取り込むか否かで直接または間接外気冷却方式に分けられる．これらの方式は自然エネルギーを利用した冷却方式とも呼ばれており，使用設備の使用電力を低く抑えることができ，空調機の効率を示す COP も高い．一方で計算機の熱を液体に伝達する方法として液浸・油浸冷却やコールドプレート冷却などがある．これらの冷却方式はスペース当たりの熱輸送量が高い事から，高発熱な計算環境での利用有利が適する．

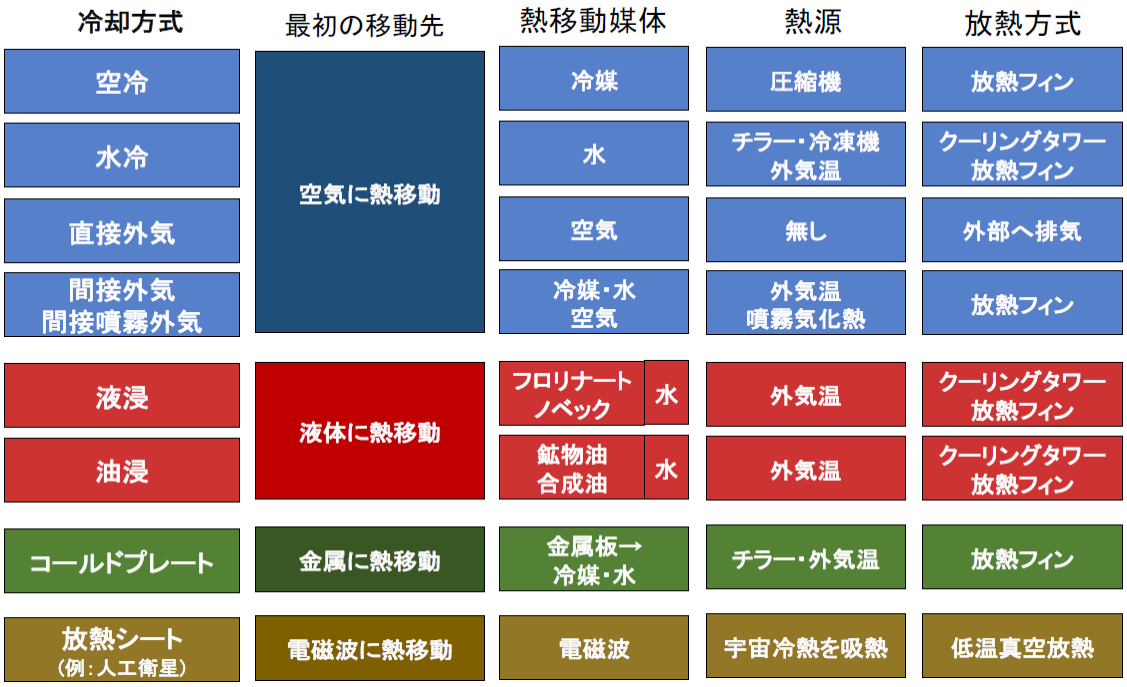


Figure 30 主な冷却方式一覧(引用：The Green Grid 資料より(2018))

現在，冷却方式が計算機のスペース当たり発熱量に応じて，現在の空冷から Figure 31 に示すように水冷・液浸冷却方式か，または自然エネルギーを利用した冷却方式にシフトし始めている．さらに

Figure 32 に示すように，これらの冷却方式を組み合わせた冷却方式も出始めており，今後 HPC の性能向上や需要増に伴い，より高効率な冷却方式にシフトしていくと考えられる．

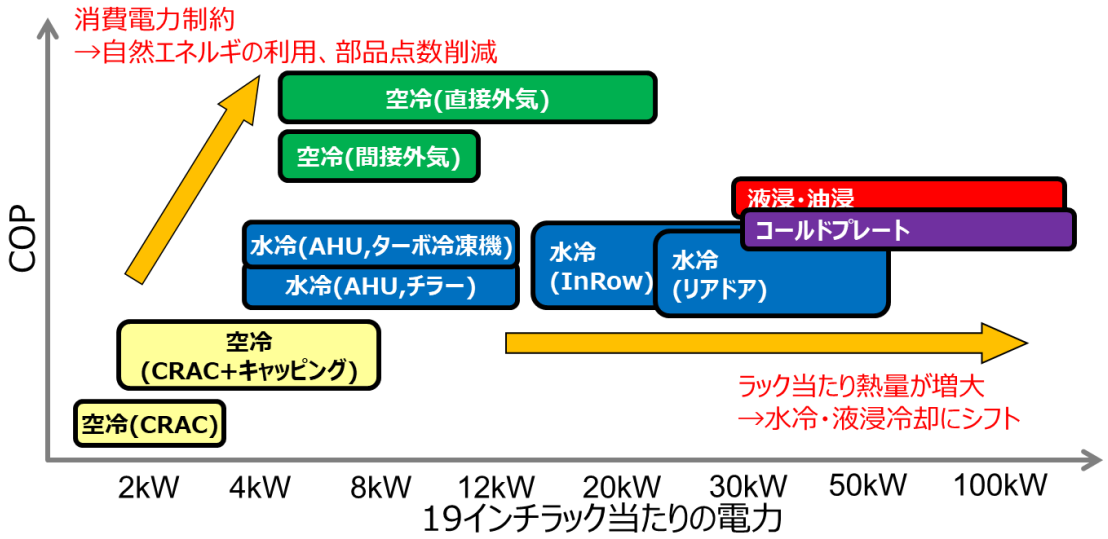


Figure 31 冷却方式の推移

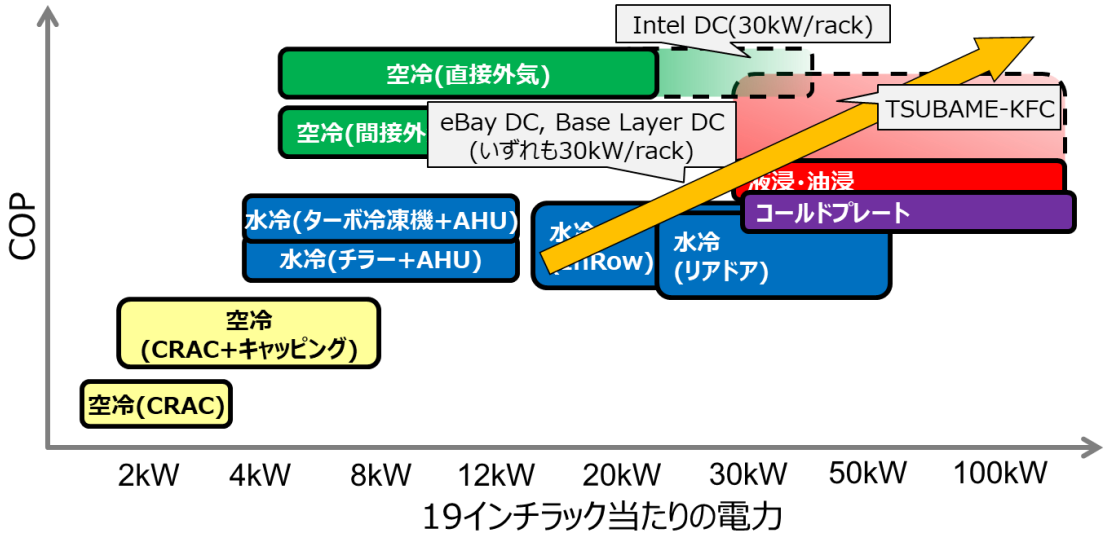


Figure 32 今後想定される冷却方式の推移

Appendix1.2.2. 自然エネルギー活用型モジュラ型データセンタの開発

富士通研究所では省エネを追及した自然エネルギー活用型モジュラ型データセンタの開発が行われた．開発したモジュラ型データセンタの外観と内観を下図に示す．このデータセンタは外気を最大限計算機の冷却に利用できるよう設計されており，空調機やチラーといった冷熱源を備えていない．外壁面に

は外気を給気する開口部と，暖気を排気する開口部を備える．2 つの小さな開口部は気化冷却機とつながっており，外気の温湿度調整が必要な時に，これらからも外気を給気する．

内部は 2 室に分かれており，気化冷却機，循環ダンパ，送風ファンを備えた空調室とサーバ等計算機が設置されたサーバ室からなる．空調室では気化冷却機と循環ダンパの起動によって外気が温湿度調整される．調整された外気は送風ファンを介してサーバルームに給気され，計算機を冷却する．サーバ室にはサーバラックが 3 台あり，ラック当たり 8kW 電力供給可能である．ホットアイルに排出された暖気は循環ダンパの開閉により一部を空調室に循環させることができる．

Figure 33 モジュラ型データセンタの外観

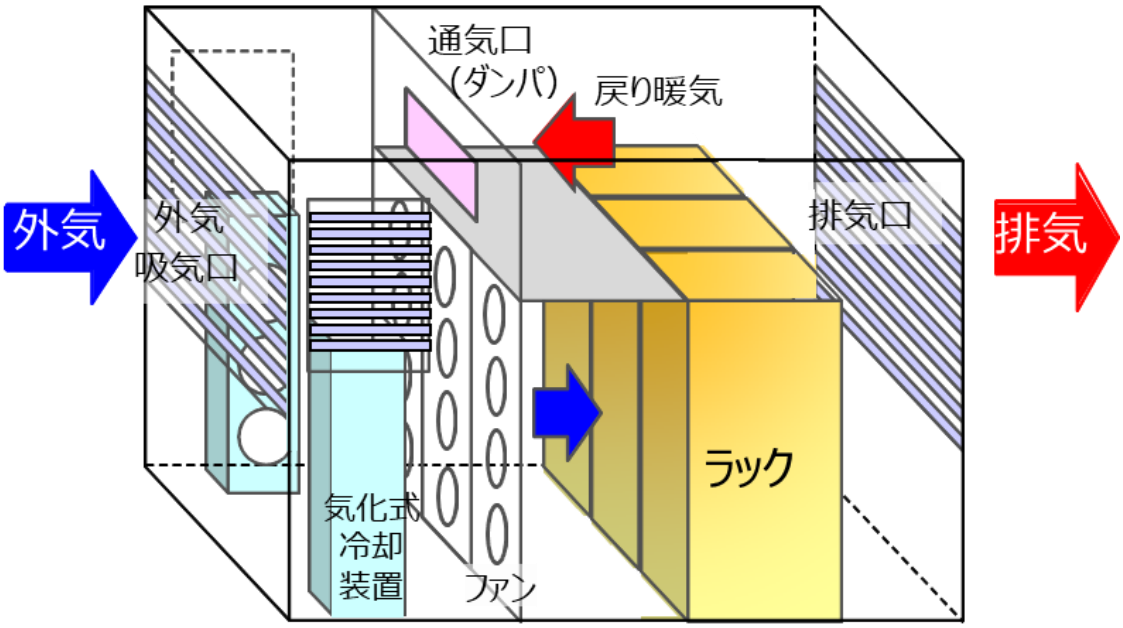


Figure 34 モジュラ型データセンタの内観

このモジュラ型データセンタの冷却システムで年間運用した時の結果を下図に示す．運用条件はコールドアイルの温湿度条件が一般的なサーバの設置条件である，10～35℃，10～85%である．循環ダンパは外気が 10℃以下または湿度 85%以上の時に，ダンパが開き暖気を循環するよう制御された．また気化冷却機は外気が 35℃以上または湿度 10%以下の時に，気化冷却機が稼働し外気を加湿冷却するよう制御された．循環ダンパおよび気化冷却機の制御によりコールドアイルの温湿度は運用条件内に制御された．

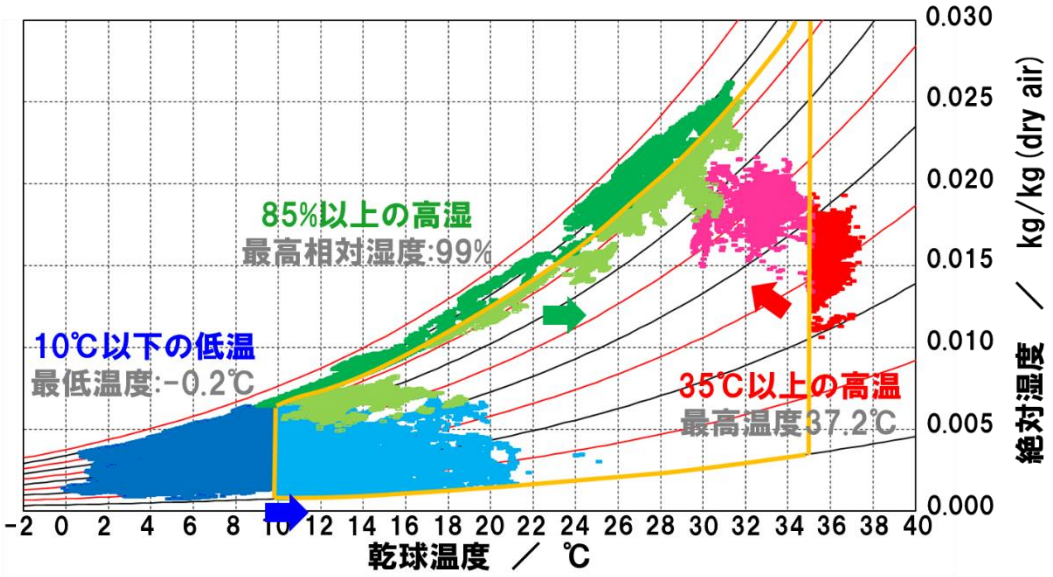


Figure 35 年間運用結果

Appendix1.2.3. 冷却機器と計算機連携した一体制御システムの開発

前項のモジュラ型データセンタにファンレスサーバを搭載することで，総電力の削減に取り組んだ．通常，サーバは内蔵部品を冷却するため冷却用のファンがある．このファンの回転数を制御して，部品の温度を適切に保っている．一方，冷却機器側にも，サーバに冷気を送風するためのファンがある．冷却コストを最小とするために，モジュラ型データセンタ全体を 1 台の大きな計算機として見立て，サーバの内蔵ファンを取り除き，下図のようなモジュラ型データセンタ側の冷却機器と連携する一体制御を実現した．

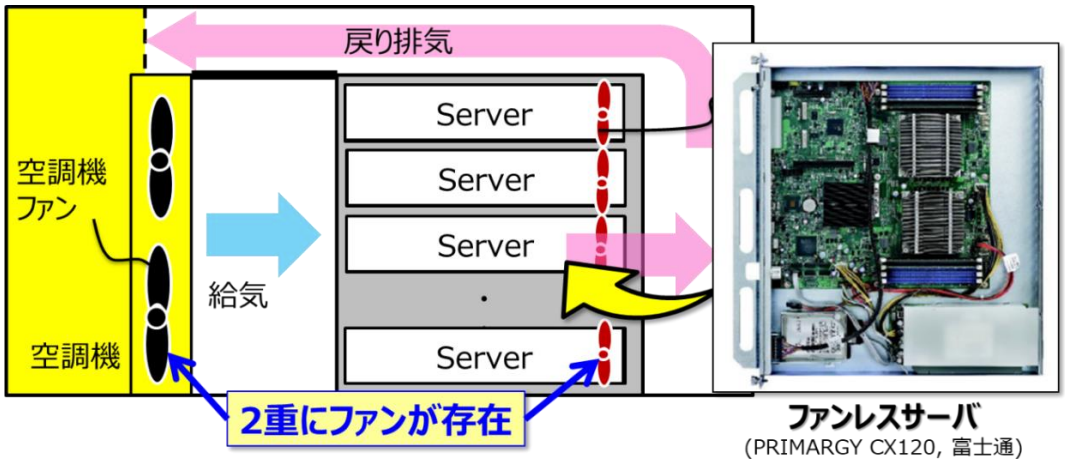


Figure 36 ファンレスサーバの採用

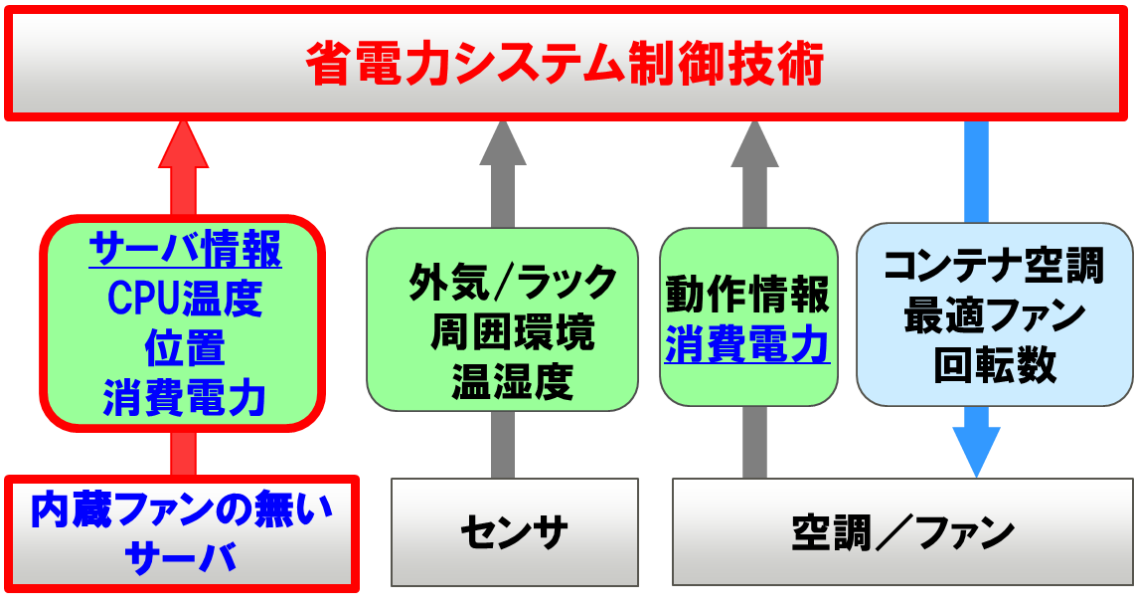


Figure 37 冷却機器と計算機連携した一体制御システム

下図に制御手順を示す．はじめに Figure 38 に示すように，CPU 温度は低温であるほどファン電力が増大し，一方で CPU 温度を高温にするほどサーバ電力がリーク電流起因により増大することから，これら 2 値を合わせた電力が最小となるような CPU 温度で運用させる．さらに Figure 39 に示すように，計算性能を維持する為，CPU 温度はサーマルスロットリングに入らない（クロック数を低下させない）上限温度以下で運用させる．

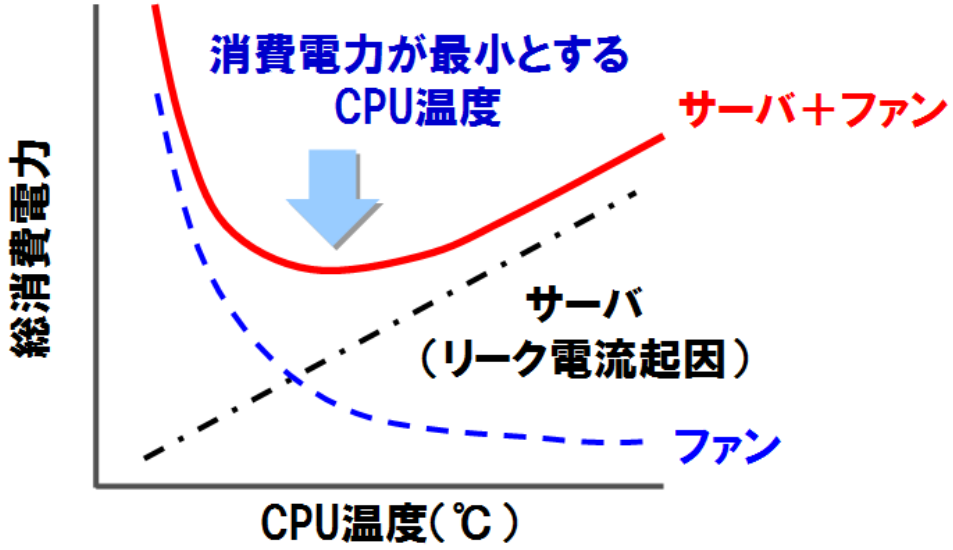


Figure 38 CPU 温度に対するファン・サーバ電力の相関

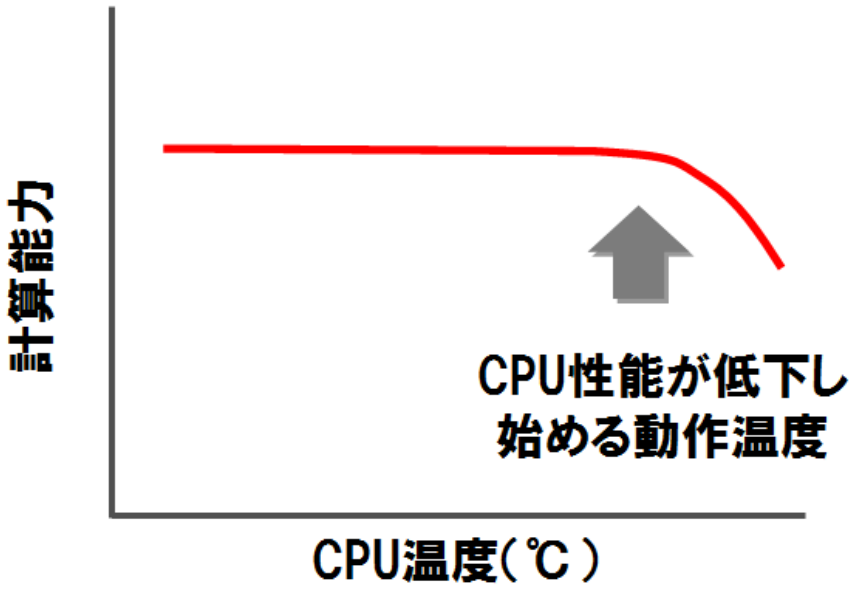


Figure 39 CPU 温度に対する計算能力の相関

今回実験で得られた総消費電力が従来空調を使用した場合，ならびにファン搭載サーバを使用した場合と比較してどれだけ削減効果があったのか比較した結果を以下図に示す．削減割合はエアコン・ファン付き計算機と比較し，夏場で最大約 41.3%と最も高く，冬場利用でも 36.8%減と極めて高い削減結果となった．本結果より外気利用，ファンレスサーバの採用，更には総電力最小値の効果が確認できる．PUE は，外気温 35℃の時は 1.14，外気温 13℃の時は 1.05 となり，年間を通して非常に高い冷却効率を実証した．

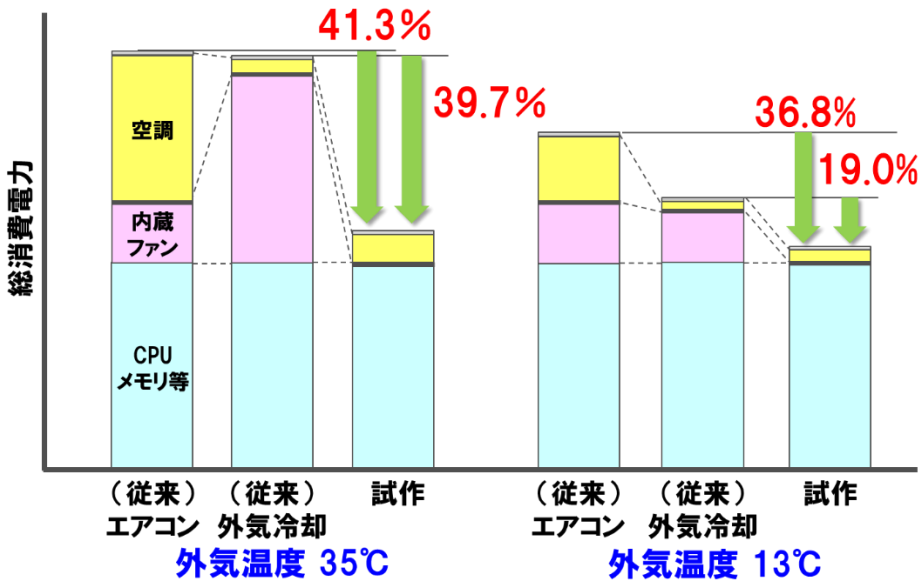


Figure 40 実証試験に基づく削減効果

Appendix1.3.　予測技術を活用した空調制御システム

本節では予測技術を活用した空調制御システムについて，富士通研究所での開発例を紹介する．

Appendix1.3.1. センタの空調システムの制御について

従来よりデータセンタの空調システムでは，外気温度や計算機温度の急激な変動でも室内を適切な温度管理条件内に維持できるよう，過剰な空調運用が行われている．特に HPC では電力消費の異なるお客様が多様なサービスを実行することから，ノード単位の発熱量は変動しやすく，より過剰な空調運用が行われる事になる．近年では空調管理システムやデータセンタインフラ管理(Data Center Infrastructure Management, DCIM)システムにより，冷却機器や ICT 機器に内蔵されたセンサ情報が一括で取得可能となった．このデータを活用した空調制御システムを開発することで，データセンタの空調電力のさらなる削減が期待できる．

過剰運用低減による空調電力削減の課題に対し，以下図に示すようなモデル予測制御と呼ばれる手法が提案されている．この手法は対象センサの将来状態を推定できる予測モデルを構築し，計測値とシステムの状態をこのモデルに入力し，予測値を算出する．得られた予測値から，空調電力が最小となるような動作を決定し，空調機器を制御する．

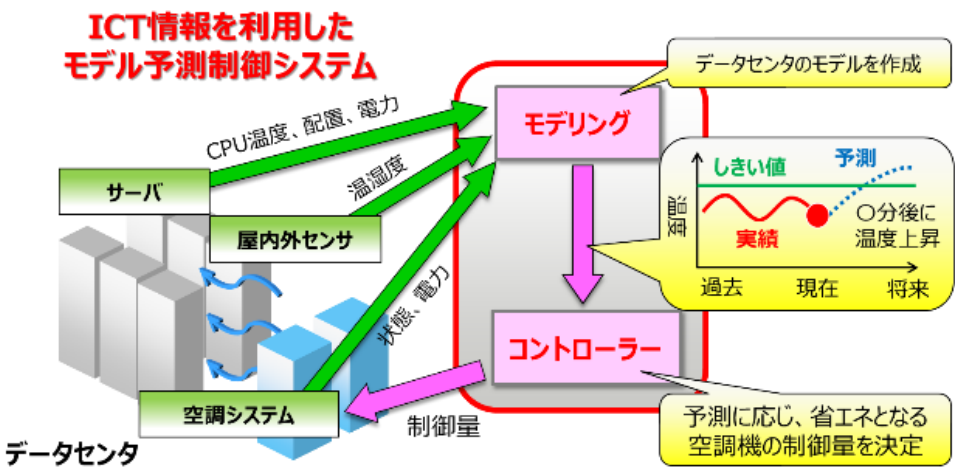


Figure 41 モデル予測制御システムの概要

Appendix1.3.2. Just-In-Time モデリングを用いた大規模センタ向け予測制御技術の開発

大規模なセンタでは以下図に示すように，計算機のレイアウトや空調設備の稼働状態が頻繁に変更され，構築したモデルが実際と乖離していくことから，固有のモデルを持たない手法が求められる．さらに空調機器や計算機の特性を十分に踏まえたうえで，予測値から空調機器を動的に制御する手法が

求められる.

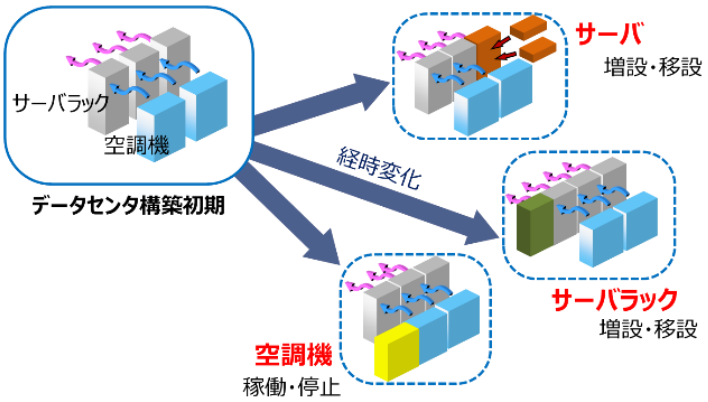


Figure 42 データセンタレイアウト変更の課題

上記を踏まえ，富士通研究所ではJust-In-Time（JIT）モデリングを使ったデータセンタ温湿度予測技術と，データセンタの空調電力を大幅に削減する空調制御技術を開発した．JIT モデリングとは固有のモデルを持たず，過去の温湿度計測値から回帰モデルを逐次作成する手法である．はじめに，取得した現在のセンサデータと類似した過去データを，データベースから検索する．次に検索した過去データから，ローカルモデルを構築する．構築したモデルと現在のセンサデータから予測値を算出する．モデルは逐次廃棄され，都度新しいモデルを構築する点が，本手法の大きな特徴である．モデルは自 然に更新されるため，本手法は，対象が例え非線形システムであっても高精度で将来の状態を予測できるメリットを有している.

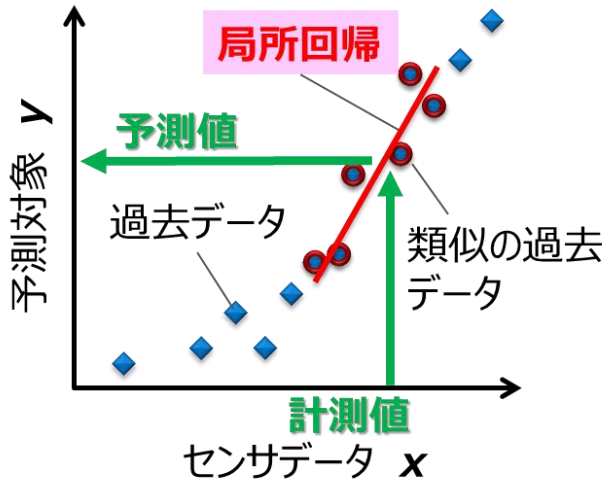


Figure 43 モデリングの概要

次に空調の消費電力を削減する制御技術として，本稿では外気導入を効率化する判 定制御，

および空調機電力を最小化する設定温度制御について紹介する．空調機電力効率に基づいた外気導入制御は Figure 44 に示す．室内の空調機近辺や屋外に温湿度を測定するセンサから，内気循環時および外気導入時の冷却・除湿に要する消費電力を計算する．この結果に基づき内気循環と外気導入の比率を制御することで温度および湿度を低消費電力で適切に管理できる．

空調機電力を最小化する設定温度制御は Figure 45 に示す．はじめに設定温度を変更した際に，過去のサーバ・室内温度分布の変化を分析し，空調機ごとの各エリアへの影響の大きさを算出しておく．あるエリアのサーバ温度が上がった時に，サーバが設置されているエリアへの影響が大きい空調機の設定温度を制御することで，最低限の消費電力での温度管理が可能となる．

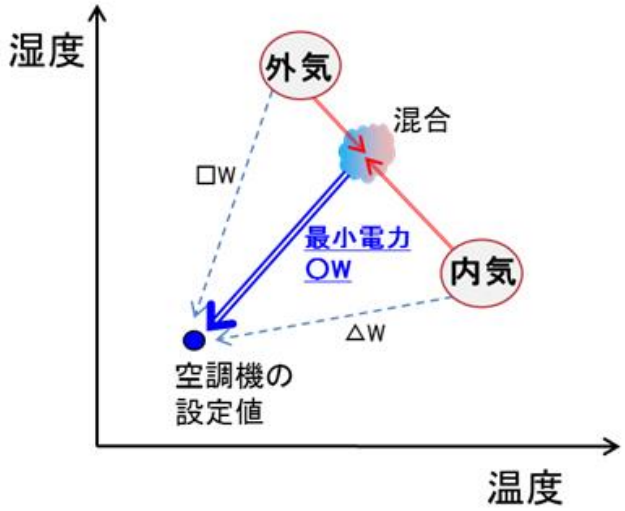


Figure 44 空調機電力効率に基づく外気導入制御

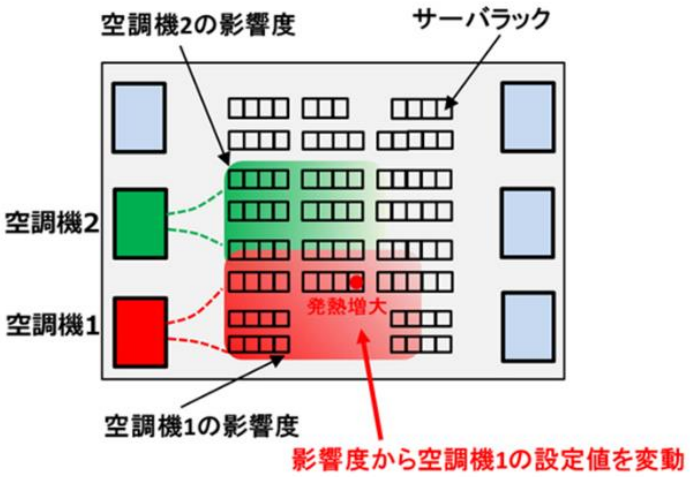


Figure 45 空調機電力を最小化する設定温度制御

本提案手法を運用中のデータセンタで実証し，従来手法と比較し提案手法の有効性を検証した．その結果提案運用の実施により，従来よりも 28.9%の削減効果が得られた．

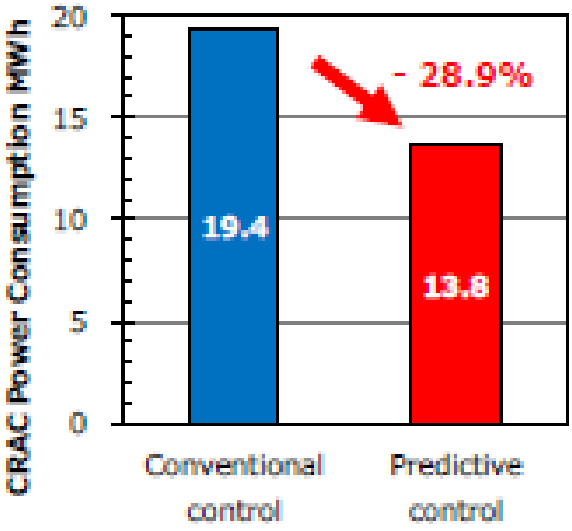
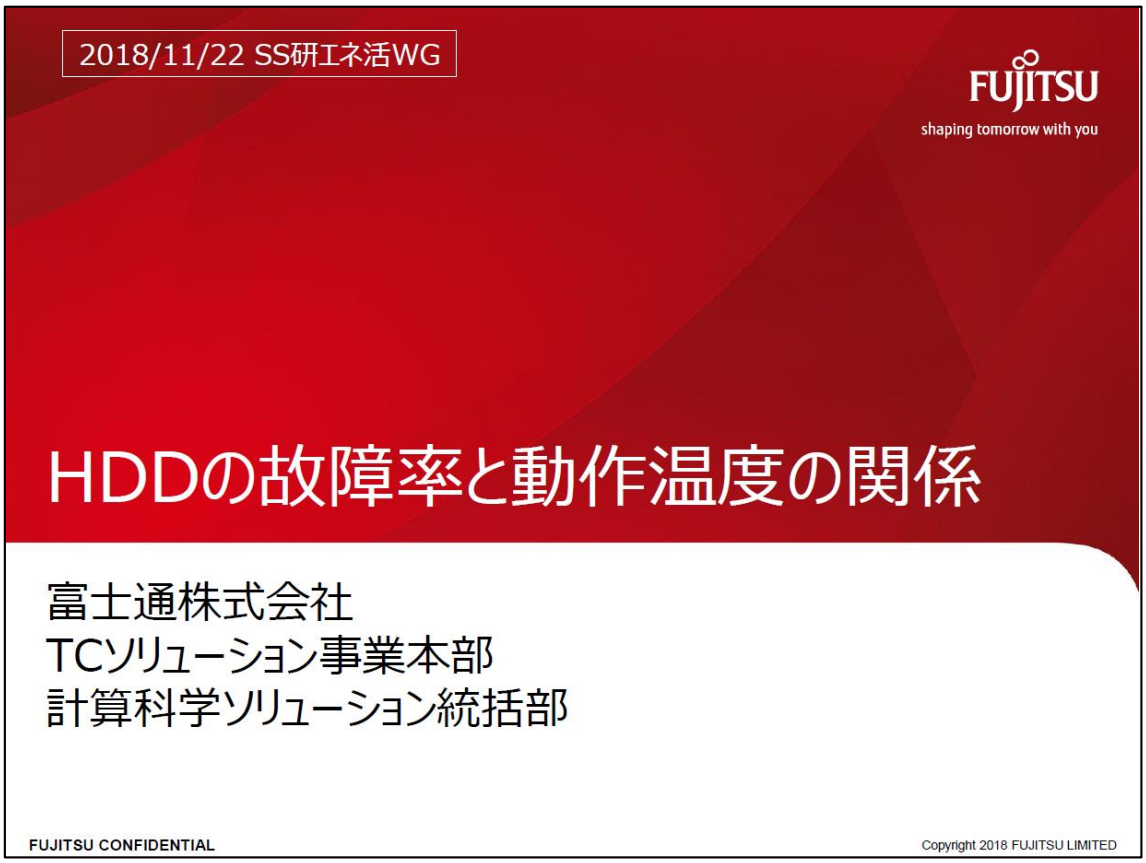


Figure 46 予測制御による電力削減効果

Appendix1.4.　ストレージと温度の関係



はじめに

FUJITSU

- 世界の電子データの90%は、磁気媒体に蓄積され、その大部分がHDDである
- HDDは現代のデータ社会において重要なコンポーネントだが、HDDの故障や寿命に影響を与えるkey factorに関する調査・研究は乏しい
- 故障に関するデータや情報は、製品の品質保証に関わるため、製造元が積極的に公開することもない
- そこで、本日はHDDの故障に関する先行研究のまとめを報告させて頂く

調査文献

FUJITSU

1. Failure Trends in a Large Disk Drive Population (Google, 2007)
   → 温度と故障率の相関 部分的にあり
2. Datacenter Scale Evaluation of the Impact of Temperature on Hard Disk Drive Failures (Microsoft and University of Virginia, 2010)
   → 温度と故障率の相関あり
3. Hard Drive Temperature – Does it matter? https://www.backblaze.com/blog/hard-drive-temperature-does-it-matter/ (BACK BLAZE, 2014)
   → 温度と故障率の相関なし
4. Environmental Conditions and Disk Reliability in Free-cooled Datacenters (University of Rutgers, GoDaddy and MS, 2016)
   → 温度と故障率の相関なし
   → 相対湿度と故障率の相関あり

1. Google (2007)

FUJITSU

■ 実施概要

- Googleのデータセンタで利用している9モデル10,000以上のHDDを対象に、9ヵ月にわたってデータ収集(数分間隔)
  - 環境データ（温度など）/ HDDのactivity level / SMART※値のデータをモニタリング
- 「HDD故障率と利用率、ライフタイム、動作温度、SMART等の相関」を統計的解析を用いて調査

■ 結果（詳細は次頁参照）

1. 利用率(平均R/Wバンド幅)は、ライフタイム前期(2年未満) のみ、故障率と相関あり
   - 初期不良フェーズをsurviveした後は、利用率の影響小
2. 温度は、ライフタイム後期(3年以降)のみ、故障率と相関あり
   - ライフタイム前期は、温度よりも他の要因の影響大
3. SMARTエラーは、故障率と高い相関あり
   - Scan Error：初回error後、60日以内に故障する確率は39倍高
   - Reallocation Counts: 初回error後、60日以内の故障する確率は14倍高　など
4. ただし、SMART エラーなしで故障するHDDも存在するため(約56%)、SMARTのみで高精度な故障予測モデルを構築するのは困難

   性能異常など他の要因との組み合わせが必要

1. Google (2007)

FUJITSU

0. AFR*1と寿命の関係

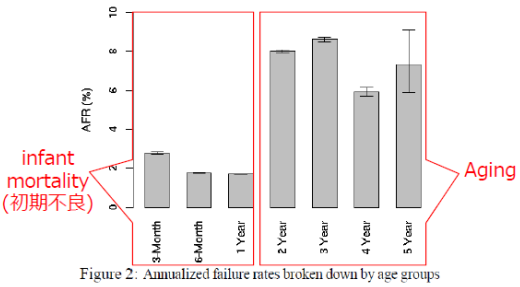


Figure 2: Annualized failure rates broken down by age groups

2. AFR*1と動作温度の関係

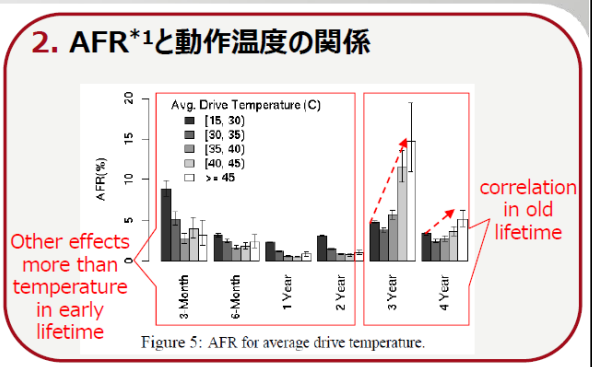


Figure 5: AFR for average drive temperature.

1. AFR*1と利用率(RWバンド幅)の関係

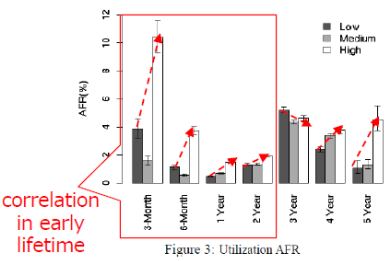


Figure 3: Utilization AFR

*1AFR: Annual Failure Rate(年間故障率)

3. AFR*1とSMART値の関係

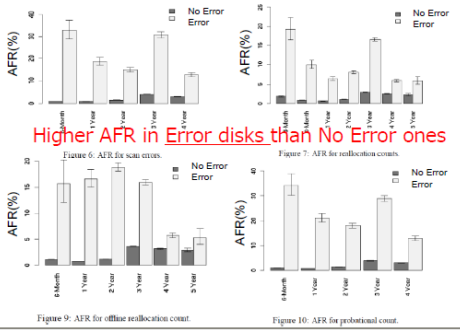


Figure 6: AFR for scan errors.
Figure 7: AFR for reallocation counts.
Figure 9: AFR for offline reallocation count.
Figure 10: AFR for probational count.

2. Microsoft and Virginia Univ. (2010)

FUJITSU

■ 実施概要

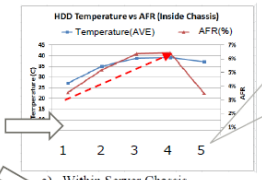
- MSデータセンタで利用している10,000以上のHDDを対象に「温度（a.サーバ内のHDD位置，b.ラック内のサーバ位置，c.DC内のラック間）とHDD故障率の相関」を1年間のデータで調査
- HDDは、ライフタイム前期の故障期を経ているものが対象

■ 結果

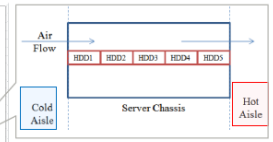
- 温度とAFRは、高い相関あり
  - a. サーバ内のHDD位置温度との相関 (R=0.79)
  - b. ラック内のサーバ位置温度との相関 (R=0.91)
  - c. DC内のラック位置温度との相関 (R=0.30)
- *Arrhenius*の式を用いたAFRの推定では、55℃で運用した場合、40℃の2倍故障率が高くなる

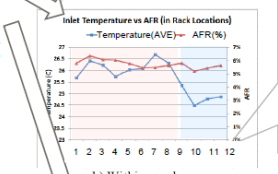
HDD Temp	Acc Factor (AF)	AFR	AFR relative to 40 C
40 C	1.257	2.75	100%
41 C	1.328	2.91	106%
42 C	1.402	3.07	112%
43 C	1.480	3.24	118%
44 C	1.562	3.42	124%
45 C	1.648	3.61	131%
46 C	1.737	3.80	138%
47 C	1.831	4.01	146%
48 C	1.930	4.23	153%
49 C	2.033	4.45	162%
50 C	2.141	4.69	170%
51 C	2.254	4.94	179%
52 C	2.372	5.19	189%
53 C	2.495	5.46	198%
54 C	2.625	5.75	209%
55 C	2.759	6.04	219%

×2

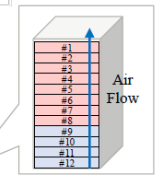


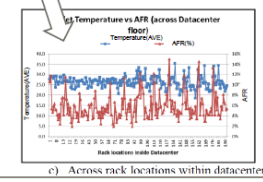
a) Within Server Chassis





b) Within a rack





c) Across rack locations within datacenter

3. BACK BLAZE (2014)

FUJITSU

■ 実施概要

- BACKBLAZEの運用する34,000HDDについて、平均動作温度と故障率の相関を統計解析で調査

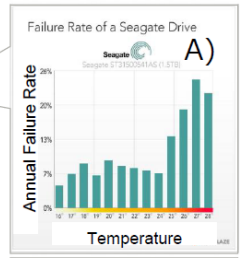
■ 結果

- 動作温度と故障率は相関無し
- 保証温度以下に冷やしても、寿命が長くなるわけではない

1. 相関係数・・・全体的に小さく、温度と故障に高い相関があるとは言えない
2. P値・・・帰無仮説「0.05未満のHDDは、故障率と温度に相関がある」
   - A) Seagate Barracudaは、温度が高くなるにつれて故障率も高くなる
   - B) 一方で、Hitachi DeskStarはその傾向がみられない

⇒ AとBで傾向が異なるため、帰無仮説が正しいとは言えない

Model	Correlation	Significant?	p-value
Western Digital Caviar Green	0.18	no	0.07
Seagate Barracuda 7200.11	0.17	YES	0
Seagate Barracuda LP	0.12	YES	0
Seagate Barracuda Green	0.05	no	0.61
Seagate Barracuda 7200.14	0.03	YES	0.02
Western Digital Red	0.02	no	0.67
Western Digital Caviar Green	0.01	no	0.88
Hitachi Deskstar 5K4000	0	no	0.82
Seagate Desktop HDD.15	0.01	no	0.25
Seagate Barracuda LP	−0.02	no	0.71
Hitachi Deskstar 5K3000	−0.02	no	0.13
Western Digital Caviar Green	−0.04	no	0.39
Hitachi Deskstar 7K2000	−0.04	YES	0.01
Seagate Barracuda XT	−0.04	no	0.56
Hitachi Deskstar 7K3000	−0.04	no	0.15
Toshiba DT01ACA Series	−0.05	no	0.73
Seagate Barracuda XT	−0.05	no	0.35



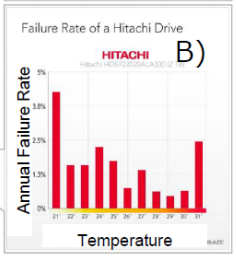


Fig.1 平均動作温度と故障率の相関係数

4. Rutgers Univ.&GoDaddy&MS (2016)

FUJITSU

■ 実施概要

- 9つの大規模データセンタを対象に、HDD故障と相対湿度及び温度の関係を調査
- データセンタの特徴およびHDD数は、右表の通り

DC Tag	Cooling	Months	Refresh Cycles	Disk Popul.
CD1	*Chiller*	48	2	117 K
CD2	*Water-Side*	48	2	146 K
CD3	*Free-Cooled*	27	1	24 K
HD1	*Chiller*	24	1	16 K
HD2	*Water-Side*	48	2	100 K
HH1	*Free-Cooled*	24	1	168 K
HH2	*Free-Cooled*	22	1	213 K
HH3	*Free-Cooled*	24	1	124 K
HH4	*Free-Cooled*	18	1	161 K
Total				**1.07 M**

Table 2: Main datacenter characteristics. The "C" and "D" tags mean cool and dry. An "H" as the first letter of the tag means hot, whereas an "H" as the second letter means humid.

■ 結果

- 温度よりも相対湿度の方が、AFRに高い相関がある

1. 冷却方式に関わらず、湿度の低いDCはAFRが低い
2. 相対湿度が高いDCほど、AFRも高くなる
3. フリークーリングのDCが、必ずしもAFRが高いわけではない
4. 室内温度の高いDCが、必ずしもAFRが高いわけではない

DC Tag	Cooling	AFR	Increase wrt $AFR = 1.5\%$
CD1	*Chiller*	1.5%	0%
CD2	*Water-Side*	2.1%	40%
CD3	*Free-Cooled*	1.8%	20%
HD1	*Chiller*	2.0%	33%
HD2	*Water-Side*	2.3%	53%
HH1	*Free-Cooled*	3.1%	107%
HH2	*Free-Cooled*	5.1%	240%
HH3	*Free-Cooled*	5.1%	240%
HH4	*Free-Cooled*	5.4%	260%

Table 3: Disk AFRs. HH1-HH4 incur the highest rates.

※9つのDCを対象とした大規模な調査だが、DC間の違い（location, workload, lifetimeなど)が不明

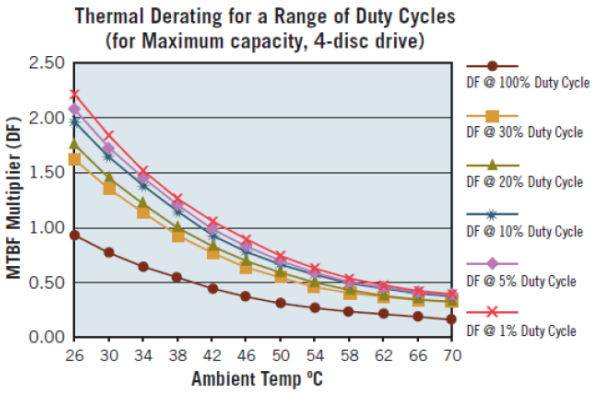
(参考) ハードウェアの寿命予測

FUJITSU

■ *Arrhenius Model* による温度とMTBF*2の関係

$$\ln k = -R_a/RT + lnA$$

k: 反応速度定数,　R_a: 活性化エネルギー,　R: 気体定数,　T: 絶対温度,　A: 定数



Seagate, Estimating Drive Reliability in Desktop Computers and Consumer Electronics Systems. *Seagate Technology Paper*, TP-338.1 (2000).

*2MTBF: Mean Time Between Failures（平均故障間隔）
本質的にはAFRと同じ　$AFR = 1 - \exp(-8760/MTBF)$　#8760は1年間の稼働時間

温度-故障率の相関に関するまとめ

FUJITSU

- 「相関あり」と主張する報告もあれば、「相関なし」と主張する報告もあり、決定的な結論は得られなかった。
- HDDの故障は、環境(温度、湿度)やワークロード、モデル、ライフタイム等、多くの要因が影響する。そのため、運用中のDCでの調査は、サンプル数を十分確保したとしても難しく、解析によって結論が異なる結果になったと考えられる。
- その中で、Googleの報告は、故障要因の大部分を考慮した解析であり、比較的信頼性が高いと言える。「ライフタイム後期」に限定すれば、MicrosoftやSeagateの信頼性試験の結果：「温度と故障率は相関あり」に一致する。
- 製造メーカーの試験では、実際のライフタイムにわたった評価が困難なため、一般的に*Arrhenius Model*を用いた加速試験 (高温下での試験)によって、故障率や寿命を予測する。

Appendix1.5.　電力チューニング関連

RAPL概要と関連研究

FUJITSU

- Intel製プロセッサ（Sandy Bridge以降）に搭載されている電力取得／制御機能
  - カウンタ情報や温度から電力を見積もる
  - チップ全体／コア／メモリ（DRAM）に対して以下の操作が可能
    - 消費電力に関する情報を取得（ミリ秒間隔で更新される）
    - 消費電力上限を設定
- 使い方
  - 測定対象プログラム内部にRAPLによって計算される電力を記録する仕組みを追加する など
- 関連研究
  - RAPL自体の検証とノード電力推定に関する検討（カオら，2013）
    - RAPLで取得した電力からノード電力を高精度に推定できる
  - コアごとの電力の推定に関する検討（小野ら，2015）
    - コアごとのカウンタ情報を基にRAPL電力を分配することでコア単位の電力を推定できる

11

FUJITSU

HPC分野での電力チューニング ～Sandia Power API～

12

Sandia Power APIとは

FUJITSU

- HPCシステムにおいて共通で必要とされる，電力管理機能／シナリオを実現するために，必要な計測・制御APIの標準化を目指した仕様
  - ソフトウェア／ハードウェア間の計測・制御APIの標準化
  - システムのデバイス構成の問い合わせやデバイスに対する電力計測・制御，計測値の時系列の変動に対する統計量の問い合わせなどに関するAPI仕様の策定
- 最新バージョンは2.0（2017年3月公開）
  - https://powerapi.sandia.gov/docs/PowerAPI_SAND_V2.0.pdf

13

Sandia Power APIでできること

FUJITSU

- 利用者側での利用
  - アプリケーションプログラム中で自ノードの電力量を取得し，それに応じて周波数を変更する など
- システム運用者側での利用
  - ジョブ実行前に，ジョブ割当範囲のノード群に対して電力上限を設定する など
- ハードウェアの様々な階層について属性値，メタ情報が取得可能
  - 属性値：電力や電力量，温度，周波数，電圧，電流 など
  - メタ情報：属性値の文字列，推定計測誤差，サンプリング回数，時間幅 など
- 一部属性については値の動的変更が可能
  - 周波数
  - 電力上限／下限
  - サンプリング回数

14

FUJITSU

HPC分野での電力チューニング ～低精度計算～

需要の高まりと適用のメリット

FUJITSU

- AI分野（DL）の発展とともに低精度演算への需要が高まっている
  - GPU，FPGAが利用されることが多い
- HPC分野では精度保証が重要なため，これまで積極的には利用されてこなかったが，電力削減の要請が強くなったことから，徐々に利用範囲が広まってきている
  - 研究も盛んにおこなわれている
- 一般的には以下のようなメリットが挙げられる
  - 小メモリ化
  - 高速化
  - 省エネルギー化

Appendix1.6.　縮退運転（自動ノード停止）

本発表の構成

FUJITSU

- 背景と目的
- 自動電源制御方式：JSCAPS（SWoPP2017）
- JAXAスーパーコンピュータJSS2におけるJSCAPSを用いた省電力化
- 実環境における省電力効果の評価
  - シミュレーション結果との省電力効果の比較
  - 長期間運用時の省電力効果と運用への影響の確認
- JSS2省電力化のパラメータチューニングによる再評価
- JSCAPSによるJSS2省電力化の課題と今後の取り組み
- まとめ

背景と目的

FUJITSU

- 背景

PCCにおける省電力化への要請の高まり

自動電源制御方式が注目

  - 待機状態となった計算ノードの電源を停止
  - シミュレータ上で既存の自動電源制御方式より高い省電力効果を確認

SWoPP2017における発表 ※情報処理学会研究報告, Vol.2017-HPC-160 No.2 (2017)

自動電源制御方式：JSCAPS (Job Scheduling Aware Power Save)を提案

  - スケジューリング情報をもとに、電源停止、再投入を実施
  - シミュレータ上で既存の自動電源制御方式より高い省電力効果を確認

- 目的

JAXAスーパーコンピュータJSS2上にJSCAPSを適用

  - 実環境における省電力効果を評価
  - JSS2の省電力化における運用への影響を確認

FUJITSU

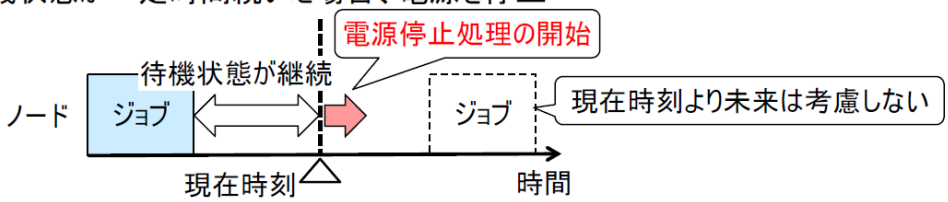
自動電源制御方式：JSCAPS（SWoPP2017）

JSCAPSにおける電源制御の特徴

FUJITSU

■ Slurm、LSFにおける電源停止処理

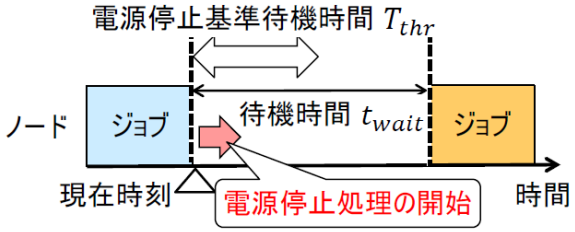
待機状態が一定時間続いた場合、電源を停止



電源停止までに余分な待機電力を消費

■ JSCAPSにおける電源停止処理

$T_{thr} \leq t_{wait}$ の場合、電源を停止



待機状態の開始直後から待機電力を削減

JSCAPSのパラメータ

FUJITSU

■ 電源停止基準待機時間

電源停止中の消費電力=0Wのとき、

（電源停止基準待機時間）

$$= \frac{\text{（電源停止処理中の消費電力量）＋（電源再投入処理中の消費電力量）}}{\text{（待機電力）}}$$

■ 最大停止ノード数

- JSCAPSが電源停止する計算ノードの上限
- 電源停止が一度に大量に行われる状況を抑止
- 設定例：最大停止ノード数=30
  … 停止ノード数が30であれば電源を停止しない

FUJITSU

JAXAスーパーコンピュータ”JSS2”における JSCAPSを用いた省電力化

JSCAPSの適用環境

FUJITSU

JAXAスーパーコンピュータ”JSS2” にある実行環境”PP”にJSCAPSを適用

■ 実行環境”PP”

消費電力が異なる2種類のサーバから構成

機種	PRIEMRGY RX350 S8	PRIEMRGY CX2550 M2
ノード数	138ノード	40ノード
CPU	Intel Xeon E5-2643 v2 3.5GHz×2	Intel Xeon E5-2643 v4 3.4GHz×2
メモリ	64GB	64GB
ジョブ実行中の平均消費電力	250W	170W
待機中の平均消費電力	180W	80W
電源停止時の平均消費電力	0W	0W

※実行環境”PP”のPRIMERGY RX350 S8は”JSS2”のノード群”SORA-PP”の一部
※平均消費電力はIPMI計測結果(2018年9月1日～30日)から計算

7

実行環境PPのジョブ投入状況

FUJITSU

■ 実行環境PPに投入されるジョブ

- バッチジョブと会話型ジョブの2種類
- 2018年9月1日～9月30日の投入状況

	ジョブ数	全ジョブに占める割合
バッチジョブ	59,489	96.5%
会話型ジョブ	2,180	3.5%
合計	61,669	100%

バッチジョブが大半を占め、会話型ジョブの割合は少ない

8

実行環境PPのスケジューリングポリシー

FUJITSU

- 実行環境PPのスケジューリング
  - バッチジョブ、会話型ジョブともに同じバッチジョブスケジューラがスケジューリング
  - 時間帯によりスケジューリングの優先度が異なる

時間帯	優先ジョブ
5:00～20:00	会話型ジョブ
20:00～5:00	バッチジョブ

  - 会話型ジョブのwait-timeは0秒
    …ジョブ投入から実行までの許容待ち時間

→ 投入された会話型ジョブは即時実行されない場合、キャンセル

会話型ジョブの実行を優先した運用

9

FUJITSU

実環境における省電力効果の評価

10

評価方針

FUJITSU

評価1．シミュレーション結果との省電力効果の比較

実行環境PPとシミュレーション環境の比較

	シミュレーション環境（SWoPP2017）	実行環境PP
ジョブの種類	バッチジョブのみ	バッチジョブと会話型ジョブが混在
スケジューリング優先度	投入順	5:00～20:00…会話型ジョブ優先 20:00～5:00…バッチジョブ優先

- 実行環境PPにおいて、バッチジョブ中心の期間の省電力効果を比較

評価2．長期間運用時の省電力効果と運用への影響の確認

- 実行環境PPの運用における長期的な省電力効果の評価
- 実行環境PPの運用への影響を確認

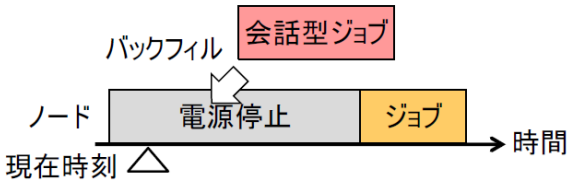
想定される運用への影響

FUJITSU

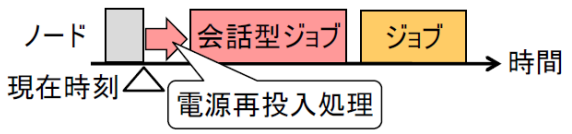
実行前にキャンセルされる会話型ジョブが増加する可能性

- 停止ノードに会話型ジョブが割り当てられる場合

1．バックフィルにより会話型ジョブが電源停止ノードに割り当て



2．電源再投入処理の完了後に会話型ジョブが実行



→ 実行までの待ち時間によりキャンセルされる可能性

JSCAPSの適用による会話型ジョブへの影響について確認が必要

評価指標

FUJITSU

■ 省電力効果の評価指標

$$\text{待機電力の削減割合}(\%) = \frac{\text{JSCAPS適用により削減された待機電力(W)}}{\text{JSCAPS未適用時の待機電力(W)}}$$

JSCAPS適用により削減された待機電力(W)
＝ JSCAPS未適用時の待機電力(W) － JSCAPS適用時の待機電力(W)

■ 会話型ジョブへの影響を示す指標

$$\text{会話型ジョブ実行前キャンセル率}(\%) = \frac{\text{実行前にキャンセルされた会話型ジョブ数}}{\text{投入された会話型ジョブ数}}$$

13

評価環境と設定

FUJITSU

■ 実行環境PP

	シミュレーション（SWoPP2017）	実行環境PP
電源停止処理に要する時間	33秒	50秒
電源再投入処理に要する時間	300秒	450秒
計算ノード数	160ノード	RX300 138ノード CX2550 40ノード
待機中の消費電力	180W	RX350 180W CX2550 80W

■ JSCAPSパラメータ設定

	シミュレーション（SWoPP2017）	実行環境PP
電源停止基準待機時間	335秒	500秒
最大停止ノード数	160ノード	評価1．178ノード 評価2． 30ノード

■ 実行環境PPの評価期間

評価1．2018年8月25日、26日　←バッチジョブの実行が中心となった期間
評価2．2018年4月3日～30日

14

評価1．シミュレーション結果との省電力効果の比較

FUJITSU

バッチジョブの実行が中心となった期間の省電力効果と比較

	シミュレーション SWoPP2017	実行環境PP 2018年8月25日,26日
平均ノード稼働率	50.3%	54.3%
全待機ノードの総消費電力の平均値	1.6kW	0.7kW
JSCAPS未適用時の待機電力の平均値	14.3kW	13.1kW
待機電力の平均削減割合	待機電力を89%削減	待機電力を95%削減

実行環境PPにおいても高い省電力効果が得られることを確認

15

評価2．会話型ジョブを考慮した実環境における省電力効果

FUJITSU

■ 省電力効果の確認

平均ノード稼働率	45.0%
全ノードの総消費電力の平均値	28.2kW
全実行中ノードの総消費電力の平均値	16.4kW
全待機ノードの総消費電力の平均値	11.8kW
JSCAPS未適用時に想定される待機電力の平均値	17.0kW
待機電力の平均削減割合	待機電力を31%削減

評価1と比較し想定以上に省電力効果が低下

■ 会話型ジョブへの影響

	JSCAPS適用前 2018年3月1日～30日	JSCAPS適用後 2018年4月3日～30日※
会話型ジョブの実行前キャンセル率	0.8%	5.8%

※4月10日除く

適用前と比較し想定以上に会話型ジョブの実行前キャンセル率が増加

16

評価2の考察

FUJITSU

■ 省電力効果低下の原因
- ■ 平均ノード稼働率が低い（約45%）
- ■ 最大停止ノード数が少ない（30ノード）

} 電源停止されない待機ノード数が増加

ノード稼働率の低い時間帯に最大停止ノード数を増加させる必要あり

■ 会話型ジョブの実行前キャンセル率増加の原因
- ■ 停止ノードに会話型ジョブが割り当たることで待ち時間が発生
- ■ 会話型ジョブのwait-time=0に設定

} 待ち時間の発生と同時にスケジューラによりキャンセル

wait-timeを増加させる必要あり

JSS2省電力化のパラメータチューニングを行い、再評価

17

FUJITSU

JSS2省電力化の パラメータチューニングによる再評価

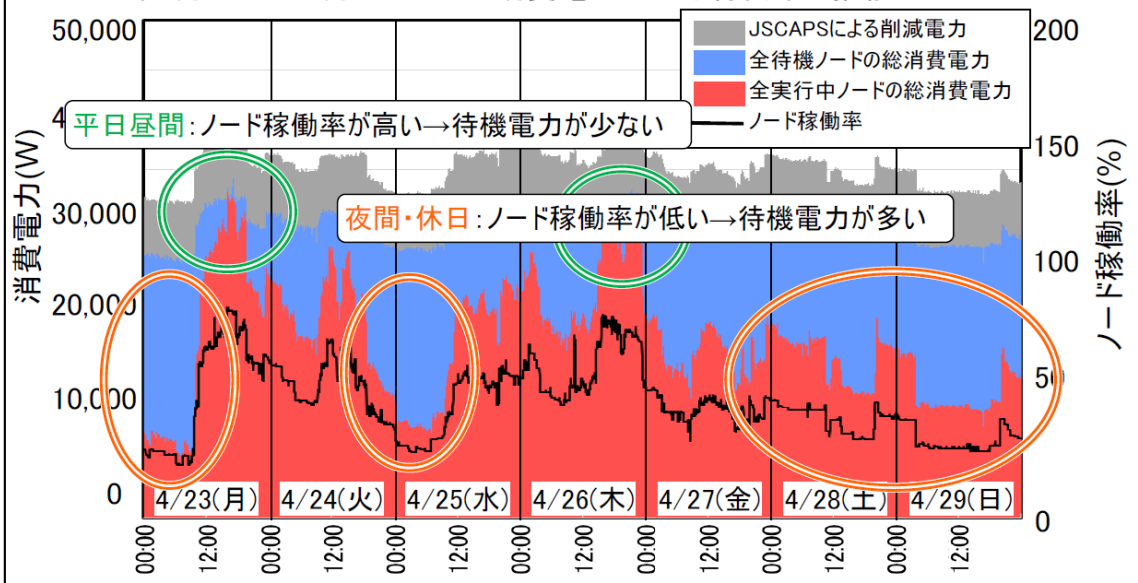
18

JSCAPS適用後の消費電力とノード稼働率の推移

FUJITSU

消費電力とノード稼働率の推移から最大停止ノード数を増加する時間帯を決定

■ 2018年4月23日～4月29日までの消費電力とノード稼働率の推移



夜間・休日に最大ノード数を増加することで待機電力をさらに削減

JSS2省電力化のパラメータチューニング

FUJITSU

■ 省電力効果の増加

ノード稼働率の低い夜間・休日に最大停止ノード数を増加

時間帯	最大停止ノード数
平日：5:00～20:00	30
平日：20:00～5:00 土日：終日	178

■ 会話型ジョブ実行前キャンセル率の低下

会話型ジョブのwait-timeを 0秒 ⇒ 1,200秒 に増加

上記チューニングによる省電力効果と会話型ジョブへの影響の変化を確認

チューニング後の省電力効果

FUJITSU

■ チューニング前後の省電力効果の比較

	パラメータチューニング前 2018年4月3日～30日	パラメータチューニング後 2018年8月1日～31日
平均ノード稼働率	45.0%	71.9%
全ノードの総消費電力の平均値	28.2kW	32.1kW
全実行中ノードの総消費電力の平均値	16.4kW	30.5kW
全待機ノードの総消費電力の平均値	11.8kW	1.6kW
JSCAPS未適用時に想定される待機電力の平均値	17.0kW	8.2kW
待機電力の平均削減割合	待機電力を31%削減	待機電力を80%削減

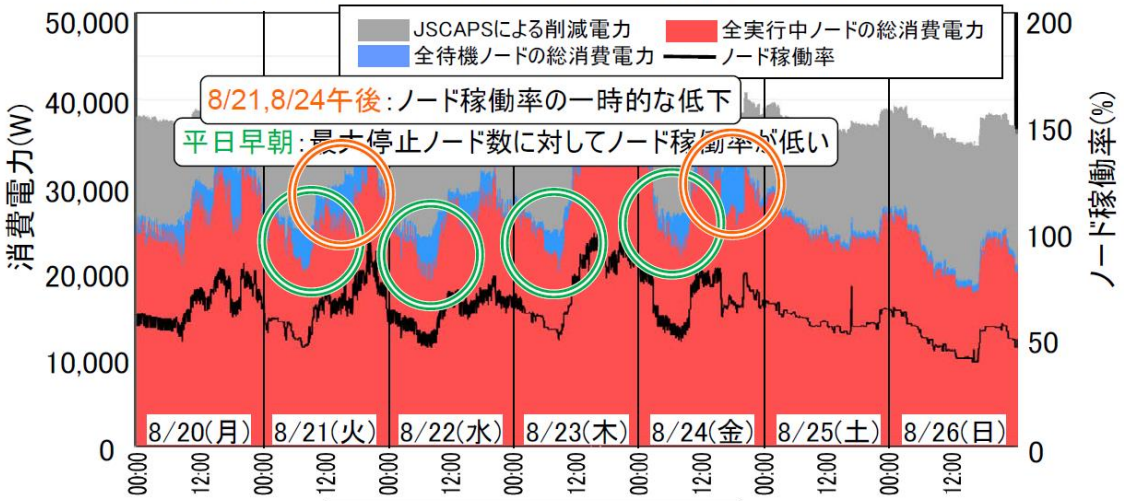
省電力効果の大幅な改善を確認

省電力化の改善状況を検証

チューニング後の消費電力とノード稼働率の推移

FUJITSU

■ 2018年8月20日～8月26までの消費電力とノード稼働率の推移



全期間で待機電力が減少

さらなる待機電力の削減に向けて

- 平日早朝　… 最大停止ノード数の切り替え時刻の変更
- 8/21,8/24午後 … ノード稼働率の変動に応じた最大停止ノード数の動的変更

チューニング後の会話型ジョブへの影響

FUJITSU

■ チューニング前後の会話型ジョブへの影響の比較

	パラメータチューニング前 2018年4月3日～30日※	パラメータチューニング後 2018年8月1日～31日
会話型ジョブの実行前キャンセル率	5.8%	14.7%

※4月10日除く

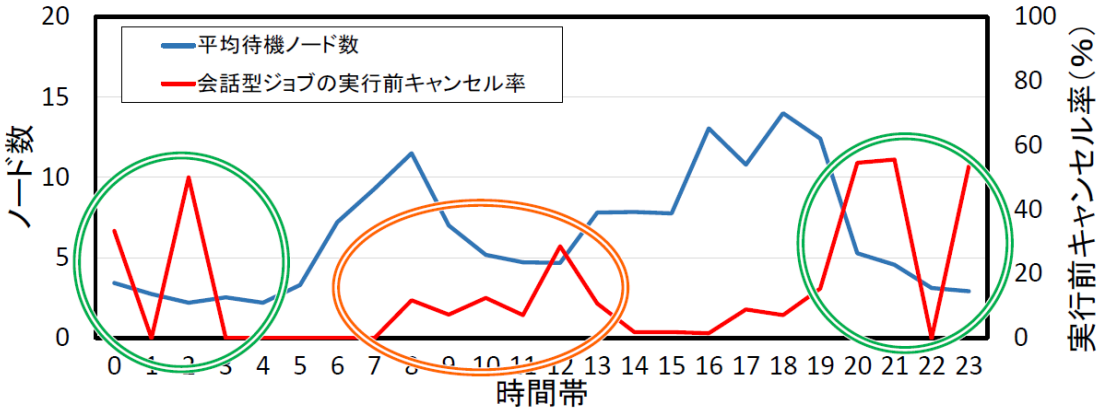
会話型ジョブへの影響は想定以上に大きくなった

会話型ジョブが実行前にキャンセルされる原因を確認

チューニング後の会話型ジョブと待機ノード数の関係

FUJITSU

■ 待機ノード数と会話型ジョブの実行前キャンセル状況（2018年8月20日～8月26日）



待機ノード数の減少により会話型ジョブの実行前キャンセル率が増加

- 夜間 … バッチジョブのスケジューリングを優先
- 午前 … 会話型ジョブに対して即時割り当て可能な待機ノード数が少ない

会話型ジョブの投入に応じた待機ノードの事前確保が必要

JSCAPSによるJSS2省電力化の課題と今後の取り組み

FUJITSU

- JSCAPSによるJSS2省電力化の課題
  - 一時的なノード稼働率の変動への対応
    時間帯によるパラメータの変更だけでは不十分
  - 会話型ジョブの実行前キャンセル率増加への対応
    必要な待機ノードを事前確保するため会話型ジョブの投入状況の予測が必要
  - 実運用下でのパラメータチューニング
    多くのユーザが利用しているためパラメータ変更は慎重さが必要
- 今後の取り組み
  - JSCAPSのパラメータをリアルタイムにフィードバックできる仕組みの追加
  - 会話型ジョブの投入履歴をもとにした投入予測機構の追加
  - より実環境に近いシミュレータを導入してのパラメータチューニング

まとめ

FUJITSU

- まとめ
  JSCAPSをJSS2実行環境PPに適用し、省電力効果を評価
  - シミュレーション結果との省電力効果の比較
    待機電力の平均削減割合 95% → 実環境においても高い省電力効果を確認
  - 長期間運用時の省電力効果と運用への影響の確認
    パラメータチューニング後
    ・待機電力の平均削減割合　　80%
    ・会話型ジョブの実行前キャンセル率　14.7%
    → 高い省電力効果が確認できたが、会話型ジョブへの影響が増加
- 今後の取り組み
  - ノード稼働率の一時的な変動への対応
  - 会話型ジョブの実行前キャンセル率増加への対応
  - 実運用かでの効率的なパラメータチューニング

謝辞

本研究における省電力の検証結果は、
JAXAスーパーコンピュータ”JSS2”を利用して得られました。

Appendix2　冷却設備電力の削減見込みに向けた施策

冷却設備電力の削減見込みに向けた施策

理化学研究所
黒川 原佳

はじめに

- 状況整理
- データセンターの冷却システム
- ある程度想像と前提条件をいれつつ、何が出来るのか。

状況整理

- IT機器には冷却システムは不可欠であり、データセンターの構成要素として必要なものである。
- 冷却方式には様々な方式と様々な条件が存在し、一意に何が最適という結論は出せない。
  - 条件とは、IT機器とその運用、冷却方式、立地、資金、環境などなど
- 一般的にはIT機器のワークロードは定量的に定義できないし、時間的にも不確定であるが、スパコンのような事前にジョブ特性が推測出来るものもある。
  - IT機器の負荷変動が予知できない普通のデータセンターには適用できない。
  - スパコンのワークロードにしても、予測による運用は実運用に耐えられるのか？

データセンターの
冷却システムの制約条件

- 一般論として語る場合、従来のIT機器や冷却システムを最適化し、電力削減を行うことは非常に困難。
  - IT機器のワークロードや時系列の稼働状況を予測できないため、入れているIT機器をいつでもそれなりに冷やす必要となる。
- スパコンとして語る場合、ジョブのプリ計測によって、IT機器の作動状況から、冷却システムを最適化することもできなくはないかもしれない。
  - いつ、どのぐらい冷却しなければならないか、予測できれば冷却電力を削減できる可能性がある。しかし、外した時のことを考えると、運用として際どいところまでは踏み込めないだろう。
  - ただし、ストレージ機器については、一定以上の冷却を考えておくしかない。
- そもそも、なぜ本体系以外でこんなに困らないといけないのか？
  - 昔ほど冷やさねばならないという状況にはなっていないように思えるが、5年で10倍の性能向上になっている。
  - ただ、実装密度は上がっているが、消費電力がそれほど上がっていると思えない。
    - トップエンドは別にして。

冷却システムから考える データセンター構築の施策

- ITシステムとして冷却システムは最小限のみを考えておくだけにしたい。
  - どこまで、どのように冷やすかのガイドラインが必要ではないでしょうか。
  - センター側からすると、IT機器提供側の（過剰な）要求に従って（乱暴に言えば）、施設環境を用意して、それを吸収しているに過ぎない。
    - これは電源設備（ブレーカー数など）でも同様ですが、その条件設定は安定のためのコミットメントになるで、誰も責任を負わず、一番安全条件のみが提示される。
- 最小限の冷却とするなら、IT機器は可能な限り自然冷却（空冷、水冷込みで）できればいい。
- 日本のデータセンターにおいて、－10度～＋45度の外気温度範囲、湿度20%～100%で稼働できるIT機器（ストレージも含む）があればいいのでは？
- ファンやポンプ動力もかなり大きな消費電力になるので、出来るだけ自然に空気や水が移動するような冷却システムが考えられないのか。

問題点

- IT機器の動作環境条件
  - サーバ系はいけるのではないか。
  - ストレージはどうする？
  - 湿度に対する条件が厳しい？
  - スパコン系で高密度の実装が本当に必要なのか？
  - 機器の種別（サーバとかストレージなど）で動作条件がことなるのは大変困る。
- 既存の建物
  - 既存建物を利用する限り、データセンターの全体最適化は非常に困難と思われる。
  - きちんと冷やせるように密閉度が高いものが多い。消火も考えると、、、
  - 本当に45度になった時の保守作業どうする？
- 設置条件
  - 空気の熱対流だけでは冷やしきれないか。
  - 温水冷却などを検討しないと難しいか。

今後、データセンターを構築する場合の方向性

- IT機器が本当にどの程度の冷却が必要なのかをまずは考える。
  - 以下の理由で将来のことは考えない。
  - IT関連機器の動作環境条件を一定にできる機器を選ぶ。
- データセンターに立派な建物は立てない。ホットとコールドをエリアで分けるのではなく、ホットエリアはほぼ屋外となるように分ける。
- IT機器は動作環境を幅広なものを検討し、高密度の実装にはしない。
- 必要最低限のファン・ポンプを環境条件（IT機器の稼働状況、天候など）に応じて動作させる。
  - このファンやポンプ電力は、再エネかグリーン電力からの利用。

以上

==========

情報システムの効率的エネルギー活用検討 WG 成果報告書

【発行者】：サイエンティフィック・システム研究会
【編　集】：サイエンティフィック・システム研究会　情報システムの効率的エネルギー活用 WG
【発行日】：2020 年 3 月 17 日
【連絡先】：サイエンティフィック・システム研究会 事務局
〒105-7123 東京都港区東新橋 1-5-2 汐留シティセンター
Email: ssken_office@ml.css.fujitsu.com
http://www.ssken.gr.jp/MAINSITE/

==========