Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

グリーンスパコン TSUBAME2.0 における電力危機対応運用

遠藤敏夫[†1,†2] 松岡聡[†1,†3,†2] 額田彰[†1,†2]
長坂真路[†4] 四津匡康[†4]

本稿では，本年三月の大震災に起因する電力危機状況における，東工大 TSUBAME2.0 スーパーコンピュータの運用について報告する．スパコン設計・運用において省エネルギーは第一級の課題であるが，電力供給能力の不足状況においてはさらに，ピーク電力の上限遵守が必須であるという課題が明らかになった．今夏，時間・資源が限られた中で TSUBAME2.0 においてほどこした対策とその将来課題について報告する．

Operation of TSUBAME 2.0 Green Supercomputer dealing with Power Crisis

Toshio Endo,[†1,†2] Satoshi Matsuoka,[†1,†3,†2]
Akira Nukada,[†1,†2] Masamichi Nagasaka[†4]
and Tadayasu Yotsu [†4]

We report the operation of TSUBAME2.0 supercomputer dealing with the power crisis caused by the poweful earthquake on March 11, 2011. While saving energy consumtion is and will be the most important issue in design and operation of supercomputers, capping 'peak power consumption' also becomes essential in the power crisis. We report measures taken on operation of TSUBAME2.0 in this summer within the limitation on time and resources, and issues to be solved.

1. はじめに

2011 年 3 月 11 日に発生した東日本大震災は特に東北地方において甚大な被害を引き起こした一方，それに起因する福島第一原子力発電所事故は関東地方に電力供給不足を引き起こした．3 月から 4 月にかけては計画停電 (輪番停電) が行われ，7 月から 9 月にかけては経済産業省が電力大口需要家に対して昨年比 15%減の電力使用制限を課した．この状況は産業界の混乱だけでなく，大学・研究機関にも大きな影響を及ぼした．スーパーコンピュータ TSUBAME2.0 を擁する東京工業大学学術国際情報 (GSIC) センターにおいても，最大利用電力は約 2MW と，キャンパスの 10%超の電力を消費するため，事故直後においてはシステムのほぼ完全な停止，その後も大幅縮退を余儀なくされた．社会的要請に応える必要性がある一方で，研究機関の責務としては科学技術研究を推進し続ける必要があり，ましてや電力性能比において世界トップクラスである TSUBAME2.0 を効率的に活用することはセンターの第一義的な課題である．

そのため，我々は次々に変動する電力事情下において，システム運用面の対策を継続的に講じてきた．満たすべき条件は，電力制限を満たしつつ，スパコンユーザのジョブをできる限り多く走行させる，および強制終了の可能性を最低限に抑制するという点である．ここで考慮する必要があるのは，TSUBAME2.0 は様々な分野のユーザにより利用され，走るジョブは特に並列度，走行時間において広い多様性を持つという点である．また TSUBAME2.0 計算ノードに GPU が搭載されるため，利用電力の変動も典型的な x86 クラスタよりも大きいという特性がある．そして無論，システムが持つペタバイト級ストレージの内容は失われてはならない．さらに対策決定においては，エネルギー削減とピーク電力遵守を明確に区別し，後者に注力するべきである．講じた対策としては以下が挙げられる：

- 計画停電対応のため短時間運用の検討
- 消費電力 (システム/内訳) のリアルタイム可視化
- 昼夜ピークシフト運用

†1 東京工業大学
Tokyo Institute of Technology
†2 JST, CREST
†3 国立情報学研究所
National Institute of Informatics
†4 日本電気
NEC Corporation

Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

表 1 TSUBAME2.0 の消費電力の概算．表の IT 機器には，計算ノード，ネットワーク，ストレージを，冷却機器には，チラーおよび MCS ラックを含む．「最大時」は，全 GPU において compute-intensive な行列積を走らせた場合の，事実上ほぼ最大消費電力となる場合を示す（カタログ値とは異なる）

	IT 機器	冷却機器	全体	PUE
通常運用時	750 (kW)	230 (kW)	980 (kW)	1.31
最大時	1610	410	2020	1.25

- 短時間ジョブキュー導入

この過程においては種々の問題が発覚し，それらは電力モニタリングシステムやジョブスケジューラの特性によるものから，大規模システム運用にとって本質的な課題まで多岐に渡る．対策のために割くことのできる時間や人員は限られており，その中で現実的もしくはアドホックな手法を取った場合もあれば，電力リアルタイム可視化のための新規ツールの開発も行った．また今後ポストペタスケール時代に向けて引き続き研究開発が必要な課題も存在するため，本稿において報告する．

2. TSUBAME2.0 スーパーコンピュータ

TSUBAME2.0 スーパーコンピュータは 2010 年 11 月に運用が開始され，理論速度性能 2.4 ペタフロップスと，日本で初めて 1 ペタフロップスを超えたシステムである[8),9)]．代表的な特徴としては，最新世代の GPU アクセラレータである NVIDIA Tesla M2050 による浮動小数演算性能や電力効率の大幅向上，7.1PB の (raw) 容量の並列ファイルシステム，フルバイセクションファットツリー構造のネットワーク，ノードローカルストレージとしての SSD の採用，水冷の Modular cooling system (MCS) ラックによる高効率な冷却などが挙げられる．本節ではシステムの概要を示すとともに，電力危機対応に大きく影響を与える，電力測定機器とバッチキュー運用の概要について示す．

2.1 システム概要

TSUBAME 2.0 では，1400 ノード以上の計算ノードと，合計 7.1PBytes のストレージが QDR InifiniBand により接続されている (図 1)．計算ノードは 1408 台の Thin ノードと 34 台の Medium/Fat ノードから成るが，以下の説明ではノード数の少ない後者については省略する．

表 1 にはシステムの消費電力の概算について，IT 機器 (計算ノード，ネットワーク，ストレージ) と冷却機器に分類して示す．またデータ/計算機センターのエネルギー効率を示す指標の一つの PUE (power usage effectiveness, 全体電力/IT 機器電力) を示す．これは 10 月に行った分電盤によるデータ測定を基にしたものであり，本来は年間平均の PUE などを

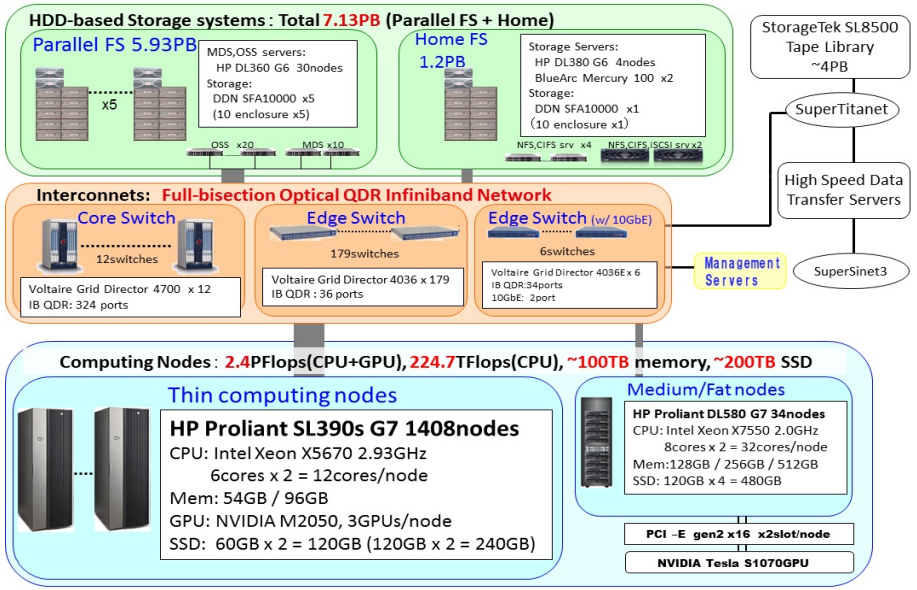

図 1 TSUBAME 2.0 の全体構成図

図 2 Thin 計算ノード HP Proliant SL390s の外観

Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

考慮すべきであるが，それでも典型的なデータセンター・計算機センターにおいて挙げられる 2.0 前後の値と比較すると非常に高効率であると言える．

以下では代表的なシステム要素について説明する．

計算ノード (図 1 下部): Thin 計算ノード Hewlett-Packard Proliant SL390s G7 (図 2) は 6 コアの Intel Xeon X5670 2.93GHz プロセッサを二つ，NVIDIA Tesla M2050 GPU を三つ搭載する．メインメモリ容量は 54GB(ただし 41 ノードは 96GB) であり，局所ストレージとしてはハードディスクの代わりに 120GB の SSD を持つ．インターコネクトに対しては 40Gbps QDR InfiniBand の host channel adapter (HCA) 二つを介して接続される．

ノードあたりの理論演算性能は，CPU 部分が 140.8GFlops，GPU 部分は三個合計で 1545GFlops であり，GPU を効率的に利用することにより高効率な演算が可能となる．このようなハイブリッド型のノードの特性として，典型的な x86 クラスタ等と比べると利用電力のダイナミックレンジが広い．簡易的な計測によれば，アイドル時のノードあたりの消費電力は 270W 程度なのに対して，CPU がビジーな場合は 400W 程度，GPU が全てビジーな場合は 850W 程度であった (この数値は後述の iLO3 を用いた DC レベルの計測であり，電源ユニットによる損失を含まない．ビジーにするために compute-intensive な行列演算を走行)．この特性が，上述の表 1 において通常運用時と最大時の電力の乖離の最大の原因ともなっている．

ペタバイト級ストレージ (図 1 上部): TSUBAME2.0 は，全計算ノードが共有するための，合計容量 7.2 ペタバイトのストレージを持つ．六つの 1.2PB のストレージシステムからなり，それぞれの中核となるのは DDN SFA10000 ストレージである．六つのシステムのうち一つはホーム領域として NFS/CIFS プロトコルに対応し，五つは大容量データ領域として Lustre プロトコルなどに対応する．

ストレージシステムの消費電力は，設置個所の都合でストレージサーバおよびバッチジョブサーバ等と合算の計測となるが，それらを含めて約 80kW となる．

フルバイセクションネットワーク (図 1 中央部): 全ノードとストレージを接続するインターコネクトはファットツリートポロジーの QDR InfiniBand ネットワークであり，フルバイセクション構成であることが大きな特徴である．Dual rail 構成であり，各 rail が二段のスイッチから成るファットツリーを構成する．エッジスイッチとして 36 ポートの Voltaire GridDirector 4036 を 185 台持ち，それぞれのポートのうち 18 は上流のコアスイッチと，残り 18 は下流のノードと接続される．コアスイッチとしては 12 台の 324 ポートの GridDirector 4700 である．

図 3 (左) 屋外に設置されたチラーの様子．(右)Thin 計算ノードを 30 台格納する MCS ラック．

コアスイッチの消費電力は合計で約 35kW である．エッジスイッチについてはノードと同じラックに格納されているために計測が困難であるが，カタログの標準消費電力から積算すると 106W×185=19.6kW 程度となる．

冷却機構: システム中で最も電力を消費し熱を発生するのは計算ノードであり，典型的な運用時には 0.6∼0.7MW 程度，最大では 1.4MW に達する．これを冷却する冷却機構は主に，屋外に設置されたチラー (図 3 左図) と，計算ノードを 30 ノードずつ格納する modular cooling system (MCS) ラック (図 3 右図) から成る．MCS ラックはチラーとパイプでつながれ，チラーから送られてきた冷水と空気の間で熱交換を行う．その冷気を計算ノードの前面に当てることにより冷却を行う．MCS ラックは (冷蔵庫のように) 密閉されており，冷却対象を部屋全体ではなくラック内の空間のみとすることによって効率を向上させている．

チラーの消費電力は，システムの消費電力や外気温などにより大きく変動するが，典型的には 150∼250kW の間を変動する．MCS ラックは 42 ラック存在し[⋆1]，ラック自身の消費電力はそれぞれ 1.6kW×42=67kW 程度である．

⋆1 設置場所の都合上，MCS ラックに格納されないノードが 100 ノード強存在するが，それらは空調で冷却されている

Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

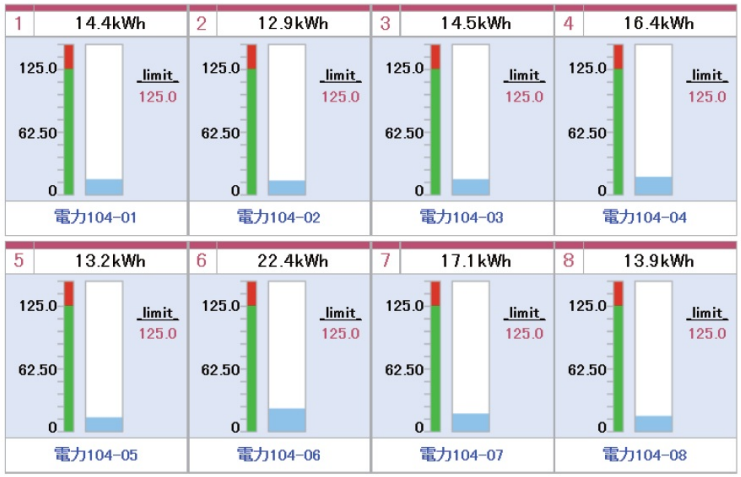

図 4 分電盤で計測された消費電力の 1 時間毎の積算値を Web 閲覧する様子.

2.2 TSUBAME2.0 の電力測定機器

我々は TSUBAME2.0 の設計当初から，電力効率を高めることと同時に，運用中のリアルタイム電力モニタリングが重要という認識に立っており，その点は調達仕様書にも記述されていた．それに基づき TSUBAME2.0 は導入時から 3 種類の電力測定装置を備えており，それぞれ異なる単位での消費電力情報の取得が可能である．本年の電力危機対応運用においては，このうち分電盤の測定機能がシステムの電力動向を把握するのに向いていると考え，それを主要に用いた．ただし，それぞれ計測の粒度や情報可視化などにおいて課題も発見された．

分電盤の電力測定機能: TSUBAME 2.0 への電力供給は幾つかの電力系統に分割されており，それぞれの消費電力を分電盤内で計測している．計算ノードの場合は 3 ラック 90 ノードが一つの分電盤に割り当てられている．

これらの分電盤からの電力情報は自動的に可視化され，GSIC の Web サイト http://mon.g.gsic.titech.ac.jp で公開されている (図 4)．システムの消費電力を把握するためには 3 種類の測定装置のうち最も適しており，電力性能の世界ランキング Green500(後述) に登録する際にもこの分電盤のデータを用いた．

しかしながら，Web で閲覧可能なのは電力そのものではなく，(1) 毎時ゼロ分からの積算電力量と，(2) 時間ごと電力積算量のログ (過去数日分) である．このインタフェースはシステムの動向を把握する上で使いやすいものではない．例えばシステム負荷の変動やチラー電力の変動をリアルタイムで知りたいときに，(2) の一時間毎のデータでは情報が不足している．この点については，次節に述べるように本年の運用において改良した．

MCS ラックの電力測定機能: HP の Modular Cooling System (MCS) ラックにも電力管理機能が搭載されている．ラック内に搭載される計算ノードの消費電力情報を一括して管理しており，リモートから管理者権限でログインすることにより情報閲覧することができる．しかし，閲覧可能なのはノードの最大電力と平均電力であり，現在の電力の閲覧機能が (なぜか) ない．

ノードの電力測定機能: Thin 計算ノードである HP SL390s G7 には HP の Integrated Lights-Out 3(iLO3) マネジメントプロセッサが搭載されている．リモートからのサーバ管理に利用するもので 温度などの各種モニタ機能も有し，リモートから管理者権限でログインし現在の消費電力情報を取得することが可能である (図 5)．

ノード毎の現在の電力取得機能は，本稿の範囲からは外れるものの，各ジョブにおける電力最適化を行う上で有用と考える．ただしこのためには以下の課題がある：

- 情報の更新頻度が約 10 秒毎である．例えば関数ごとの電力を測定しそれをフィードバックさせた最適化を行うためには，はるかに細かい時間粒度が必要となると考えられる．10 秒の粒度でどの程度まで最適化が可能となるか，今後の研究が必要である．
- 現在のところ情報取得には管理者権限が必要である．一般ユーザが軽量に情報取得可能な API やツールが必要であろう．

他の iLO3 の機能として，電力値に上限 (power cap) を設定し，これを超えた場合にプロセッサの動作周波数を下げることも可能である．

2.3 速度性能と電力性能比

TSUBAME2.0 は世界スパコンランキングの上位にランクされ，速度性能，電力効率ともにトップクラスとなっている．最も知られたランキングは Top500[2] であり，Linpack と呼ばれる密行列計算のベンチマークの速度性能によりランクを決定する．我々は TSUBAME2.0 の 4,000 枚以上の GPU を効率的に利用するために Heterogeneous Linpack ソフトウェアを開発し[10]，1,357 台の計算ノードにより 1.192PFlops を記録した．これにより 2010 年 11 月の Top500 において世界 4 位となった．

Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

```
</>hpiLO-> show /system1/oemhp_power1

status=0
status_tag=COMMAND COMPLETED


/system1/oemhp_power1
  Targets
  Properties
    oemhp_powerreg=os
    oemhp_pwrcap=unavailable
    oemhp_PresentPower=352 Watts
    oemhp_AvgPower=355 Watts
    oemhp_MaxPower=965 Watts
    oemhp_MinPower=266 Watts
    warning_type=disabled
    warning_threshold=0 Watts
    warning_duration=0 Minutes
    oemhp_power_micro_ver=3.7
  Verbs
    cd version exit show set
```

図 5 iLO3 の電力情報の出力例

Green500 ランキング[1] は電力性能比のランキングであり，Linpack 速度性能と，計算中の消費電力の比によってランクされる．Linpack 実行中の平均消費電力について分電盤の計測値から求めたところ，計算ノードとネットワーク部分において 1440kW であった．Green500 のルールによると，Linpack 実行中の 10%以上の時間を計測すればよい，コアスイッチ分は除いてもよいと確認されたので，それに基づく値 1243.8kW を提出時には用いた．電力性能比は 958MFlops/W となり，2010 年 11 月 Green500 において世界二位，さらに Greenest Production Supercomputer 賞を受賞した[⋆1]．

2.4 通常運用時のバッチキュー構成

TSUBAME2.0 および以前のシステムである TSUBAME1 は大学内外の研究者や企業を含めたユーザベースを持っており，その利用法は様々である．特に各ジョブの並列度や実行時間，GPU 利用の有無などの多様性に対応するために，TSUBAME1 時代から引き続き，複数の異なる特性を持つバッチキューを用意することとした．このキュー構成は，本年必要となった節電対応にも強く関連するため，概要を説明する．

TSUBAME2.0 の計算ノード群はバッチキューシステムである PBSPro により管理されており，通常運用時のキュー構成の概要を表 2 に示す．計算ノード群は大きく 3 つのグループに分割されている．これ以外に，インタラクティブノード，96GB ノード，Medium/Fat ノードから成るキューがそれぞれ存在するが省略する．

- S キューは並列ジョブを想定したキューであり，ユーザは任意のノード数を指定してジョブを投入することができる．複数ジョブがノード内で混在することは無く，ノード占有型である．このキューにおいては，ジョブが 12CPU コアと 3GPU のほぼ全てを用いる場合に効率が良いと言える．300 ノードがこの目的に用意されている．
- G/V ノードグループにおいては，1 ノードを仮想計算機技術 KVM によって 2 ノードに見せている．

  V キュー: ゲストノードは，Thin ノードの資源のうち 8CPU コア (hyperthreading により 16 スレッド) を占有する．V キューは，このゲストノードから成るキューである．ユーザはノード単位ではなくコア/スレッド単位でジョブを投入する．想定用途としては，逐次 CPU ジョブを多数用いるパラメータスイープ型ジョブが主に挙げられる．

  G キュー: ホストノードは，ノード資源のうち 3GPU と 4CPU コア (8 スレッド) を占有する．G キューはこのホストノードから成るキューであり，GPU を利用するが，CPU コアはそれほど必要ないようなジョブを想定している．
- H/X ノードグループは以下のように利用される．

  H キュー: S キューでは事実上実行が困難な (実行順が回ってこないなど)，数百ノード級の大規模並列ジョブの実行を日常的に可能とするために，予約制のノードグループを用意し，H キューと呼んでいる．ユーザは Web から希望予約日とノード数を (ホテルの予約のように) 一日単位で予約する．予約日の間はそのユーザおよび同グループのユーザは，当該ノード群を自由に利用できる．

  X キュー: 上記の予約が入らない場合は，数百ノード単位の空きが出てしまうこととなる．この資源の有効活用のために，S キューに投げられたジョブの一部を，H/X ノードグループの空きノードにおいて実行する機能を，システム導入とほぼ同時に運用開始した．この機能を便宜的に X キューと呼んでいる．

さらに，H キューですら実現できない超大規模計算により，科学技術的にも計算科学的に

⋆1 その次の 2011 年 6 月のランキングでは，Top500 5 位，Green500 4 位とともに，Greenest Production Supercomputer を再度受賞

Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

表 2 TSUBAME2.0 の通常運用時のバッチキュー構成

ノードグループ	台数	キュー名	特徴	想定用途
S グループ	300	S キュー	ノード占有	並列ジョブ
G/V グループ	480	G キュー	3GPU 利用可	GPU 利用ジョブ
		V キュー	VM, 8 コア	パラメータスイープなど
H/X グループ	420	H キュー	日単位予約	大規模並列ジョブ
		X キュー	H 未使用時	S キューと同様

も顕著な成果を実現するために，年二回の審査制の「グランドチャレンジ制度」を発足した．これにより，TSBUAME2.0 ほぼ全体の資源を用いた超大規模計算が可能となる．第一回は 2011 年 4 月に予定していたが，震災のために実施可否についての判断がセンターとして必要となった．この点も含め次節に述べる．

3. 実施した電力危機対応運用

震災に起因する電力供給不足が発生して以来，TSUBAME2.0 の運用は次々と変更を余儀なくされた．表 3 に運用内容を時系列で示す．特に震災直後においては，東京電力からの日々の連絡対応に追われ，混乱が大きくなっていることが表からも分かる．それ以後も縮退運用などが続いたが，おおよそ一貫して下記のような方針であった．

- ユーザの利便性の損失はある程度発生するが，最小限とする．多様なユーザが存在することに注意し，キュー毎の利用率をかんがみつつも不公平をなるべく抑える．それはグランドチャレンジユーザも含む．
- 社会情勢として節電は必須であるが，エネルギーの節約と，ピーク電力制限の遵守は異なる目的関数であることに注意する必要がある[⋆1]．今回優先すべきは東京電力管内のピーク電力制限の遵守であり，管内の電力需要が減少する夜間・休日にまで一律的に節電を行うのは，スパコンの社会的意義からは無意味である．その概念に基づき，ピークシフト運用を初期から行った．

3.1 初期の対応

震災および原発事故を受け，東京電力管内では輪番停電が実施されることとなり．東京都目黒区に位置する東工大大岡山キャンパスもその対象となった．スーパーコンピュータの運用においては，完全に電力供給が停止する (かもしれない) というのは，当然のことながら致命的である．計算ノードの減少だけでなく，ログインノードやストレージも，停電の可能性がある以上，停電の可能性がある時間までにシャットダウンしておかなければならない．特にストレージのデータ損失の可能性を避けるためには，確実にシャットダウンしておく必要があった．

表 3 TSUBAME2.0 運用変更の時系列一覧

日程	内容	運用ノード数
2011/3/11	震災発生	
3/12	不要不急のジョブの実行の遠慮のお願い	
3/14	輪番停電の対応のためにシステム停止	
3/16	ストレージ・ログインノードの運用再開	
3/17	一部キュー再開	S 160, G/V 100, H/X 150
3/18	再度輪番停電対応のため全システム停止	
3/24	一部キュー再開	S 300, G/V 300
3/25–4/6	運用終了，年度末メンテナンス (ピークシフト運用準備，グランドチャレンジ部分実行含む)	
4/6–4/7	春季グランドチャレンジ本実行 (夜間)	
4/8	新年度サービス開始	S 280, G/V 100
4/8–11	第一回ピークシフト試験	4/8 の時点 + X 400
4/15–18	第二回ピークシフト試験	4/8 の時点 + X 900
4/21	東工大震災対策本部より削減指針が 50%から 75%に緩和	
4/22–25	第三回ピークシフト試験	4/8 の時点 + X 900
4/25–6/8	運用ノード増加，ピークシフト運用	S 300, G/V 480, 夜間 X 420
6/9–6/30	大学の許可の下，100%運用	S 300, G/V 480, H/X 420
7/1–7/24	夏季のため縮退運用	S 300, G/V 310, H 80, 夜間 X 480
7/25–9/25	100%稼働 (変則)．短時間ジョブキュー (Y) 運用	S 300, G/V 280, H/X 420, Y 200
(8/10–15)	法令停電による全システム停止	
9/26–10/3	秋季グランドチャレンジ・700 ノードチーム	S 300, G/V 280
10/3–10/6	秋季グランドチャレンジ・全ノードチーム	
10/6–	通常運用再開	S 300, G/V 480, H/X 420

当時の輪番停電計画は，一日のうち三時間停電する可能性があり，実際に起こるかは前日夜か当日に判明するというもので，この点もスパコン運用においては致命的となった．システムの起動および安定化までには約 7 時間[⋆2]，シャットダウンには 2〜3 時間かかる．確実に連続的に電力供給されると分かる時間が仮に 18 時間だとすると，数時間の運用のために起動・シャットダウンを繰り返すのは現実的ではなく，三月はほぼ運用できない事態となった．なお結局大岡山キャンパスでは輪番停電は起こらなかった．

三月下旬には，本来予定されていた年度末システムメンテナンスが行われたが，この時点

⋆1 当時はマスコミ等においても両者を混同する傾向にあった

⋆2 この点は後に，ネットワークスイッチがすでに起動していれば，3 時間程度に短縮できると判明

Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

より，上述のピークシフト運用に向けた準備と技術評価を開始した．

震災前より，四月上旬にはグランドチャレンジ制度を実施する予定となっており，システム全体を用いたプログラム実行を行う四チームも決定していた．実施の可否について大学からの節電要請，当時の電力事情などを鑑み検討した結果，時間の縮小により実施が決定した．当初予定では一チーム 24 時間の持ち時間だったところを，6–8 時間とし，かつ一般的に電力需要の減少する夜間に実施を行った．このような悪条件の下ではあったが，四チームの実験は終了し，それぞれ学術的な成果を上げた．なかでも，結晶の成長解析シミュレーションのチーム[5] と，生体流体シミュレーションのチーム[4] の結果はそれぞれ SuperComputing11 会議において，高性能計算分野の最も著名な賞である Gordon Bell Award の最終候補 (全てで五件) となるという成果を上げた．

3.2 ピークシフト運用

四月上旬には，大学側よりセンターの消費電力を前年度比 50%に制限するように指示が下りた．センターにはスパコンだけでなく全学ネットワーク基盤やホスティングマシンなど，停止のできないシステムも設置されている．さらにスパコンシステムの中でも，インターコネクトスイッチやストレージについては原則的に停止できない．以上のことから，計算ノード数についてはより厳しい数値にまで縮退する必要がある．予備測定などを通し，計算ノードの運用数は約 30%と決定した．

一方，上記に述べたように夜間や休日に同じ条件を課すのは合理的ではない．そのため夜間に運用ノードを増加させるピークシフト運用の可能性を探ったが，以下の課題があった．

- 毎日数百ノードの起動・停止を人手で行うのは現実的ではないため，自動化ツールを作成する必要があった．
- さらに，確率的に起動に失敗するノードが発生することが問題となった．OS が起動しないケースは対応が楽だが，起動しても一部ファイルシステムマウントが外れていたり，デバイスのリンク速度に問題が起こるケースが見られた．起動時にマシンの正常性を確認するツールを段階的に改良し対応を行った．
- 以上のような対応について，正式サービスとする前に，テストサービスと銘打ってユーザにジョブ投入を協力依頼した．

以上の対応により，30%運用と 100%運用を自動で切り替えるシステムを構築した．なお夜間・週末に増加させるノードは X キューとして，つまり S キューの混雑を緩和するキューとして稼働させた．その後大学からの節電要請が緩和されたために，四月下旬より昼間運用ノードは 60%に増加させた．

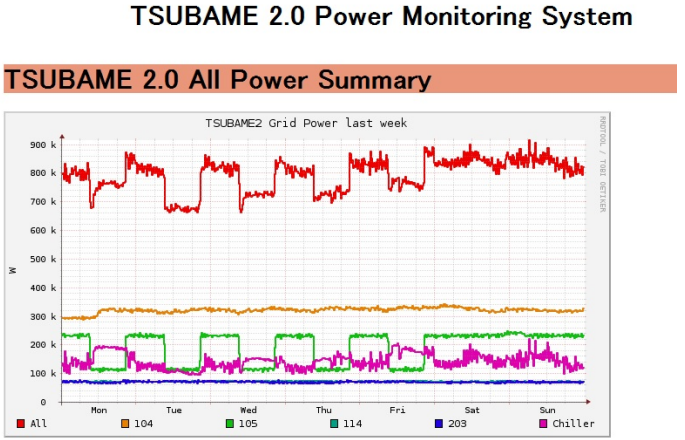

図 6 新たな電力モニタリングツールを用いた TSUBAME2.0 のリアルタイム電力表示の様子．104/105/114/203 はマシン室番号を表し，さらに細かい分電盤レベルの表示も可能

3.3 リアルタイム電力可視化の改良

2.2 節で述べた電力測定機器のうち分電盤レベルにおいてはすでに Web 上で情報を閲覧できたが，リアルタイムでシステム電力の挙動が分かるものではなかった．そのため新たな電力モニタリングツールを作成し，電力可視化を大幅に改良し，四月下旬より Web ページ (http://mon.g.gsic.titech.ac.jp/powermon) でリアルタイム電力の様子を公開した．

ツールに入力するデータ源としては，すでに分電盤が提供していた情報である「毎時ゼロ分からの積算電力量」を用いた．これを定期的に読み込み，前回との差分から電力を算出するデーモンプログラムを作成した．さらにその電力データは Ganglia モニタリングシステムに提供し，分電盤ごとおよびシステム全体の電力もリアルタイムで表示することが可能となった．その様子を図 3.3 に示す．図のグラフはシステム全体電力とマシン室毎の電力が示されているが，さらに分電盤毎 (原則 3 ラック毎) のグラフも閲覧可能である．なお各マシン室には以下の機器が設置されている：104 号室には，24 ラック 720 台の Thin ノード．105 号室には，18 ラック 540 ノードとコアスイッチ全て．114 号室にはストレージと管理サーバ群．203 号室には Thin ノード 148 ノード (非 MCS ラック) と Medium/Fat ノード．

なお図 3.3 はピークシフト運用中の五月の一週間の電力推移を示す．この中で五回電力が低下している部分があるが，これが平日の昼間に対応していることが容易に見てとれる．

Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

ツールのアイデア自身は非常に単純なものであるが，時々刻々と移り変わる電力を分かりやすく表示する，世界のスパコンの中でも稀有な例となった．

3.4 短時間ジョブキューの運用

七月下旬に，大学よりシステム 100%運用の許可が下りた．これは気温の上昇に伴う節電強化の情勢とは一見逆行するようであるが，大学所属であることの事情による；つまり，授業期間が終了したため，キャンパスの教室などの冷房需要が大幅に減ったためである．一方同時に，キャンパスの電力消費が上昇した場合には，早急に稼働ノード数を削減可能とすること，という条件も課された．これに対応しつつ，かつ走行中ジョブをできる限り中途終了させない運用を行う必要が生じた．

一見，仮想マシンを用いる V キューを，マイグレーション技術によりその目的に使えるかとも思われるが，仮想マシンの oversubscription を前提としているデータセンターと異なり，CPU 負荷の高いスパコンではそれはふさわしくない．さらに，仮に仮想マシンを移動しても G キューがホスト OS を利用しているために物理マシン全体を空にすることができない．

その代わりに，投入時に宣言される実行時間の短い (makespan が一時間以下) ジョブだけが走行可能なキューを設置し，新規に Y キューとした．このような制限を設けることで，節電警報が出たときにジョブの新規走行開始を禁止すれば，原則一時間以内に走行中のジョブは終了するため，安全に運用ノードを削減することができる．

一方でこのキューの設置においては，実現上の障害は少なかったが，設置効果においては今後検証が必要と言える．Y キュー設置前のジョブ統計においては，makespan 一時間以下のジョブというのは，ジョブ数の割合においては 95%と圧倒的多数であるが，計算機占有時間の割合においては約 8%と，比較的少ないことが分かった．Y キュー設置期間には，通常運用時よりも短時間ジョブの投入がより motivate されていた (利用単価も安く設定したため) 面はあるが，より詳細な検証が必要であろう．

4. おわりに

TSUBAME2.0 運用開始時には明示的には予期しなかった災害および電力危機によって，刻々と運用を変化させる必要が生じ，それについて報告した．当初から最低限の稼働だけ行う conservativa な方針であれば運用ははるかに容易であったが，電力効率が世界トップクラスのスパコンを擁するセンターの責任として，可能な限り最大限の運用を行う方針を採ったがゆえに，ピークシフト運用をはじめチャレンジングな課題が生じ，乗り越えてきた．さらには，この悪条件下で実施されたグランドチャレンジ制度の結果として，Gordon Bell Award 最終候補の五件中二件を占めるなどの成果を上げることもできた (さらに他の候補の一つは京コンピュータ上の成果であり，日本のスパコンを利用したものが三件を占める)．

本稿で述べた対策には，時間・人員の制限などにより，泥臭くアドホックな対応も多かった．生じた課題の中には本質的な解決が必要なものが含まれており，それにより今後のスパコン運用だけでなく，エクサスケールのシステム設計にもフィードバックさせることが可能となると考えており，以下でいくつか議論する．

- ピーク電力制限下における運用ノード数の決定においては，ある程度の猶予は持たせていたものの，原則的にノード数と消費電力が比例することを前提に置いていた．しかし 2.1 節で見たとおり，特にハイブリッド型の計算ノードにおいてはジョブ特性によって二倍以上の消費電力の開きが生じる場合があるため，単純な仮定では危険性が生じたり，逆に保守的すぎて資源に無駄が生じうる．ピーク電力遵守のために，システム利用状況やジョブ特性を動的にフィードバックさせ運用する，いわば「power capping スケジューラ」の必要性が高まる．既存のスケジューリングに関する研究としては，利用エネルギーの最小化[7] や冷却効率化[6] に関するものが挙げられるが，上記の目的においては満たすべき条件をピーク電力遵守とすることが必要である．
- 走行中のジョブをできる限り中途終了させない，という方針が運用を困難にしていた側面も見られた．チェックポインティングなどの耐故障技術，仮想マシン移送技術については多くの研究成果がすでに存在し，それらの適用範囲を早急に拡大することが，ユーザへの教育とともに求められる．
- 一点目に関連するが，エクサフロップスのシステムを 20MW 程度以下の電力で 2018 年前後に実現することが，高性能計算分野において課題となっている．その実現は現状のプロセスルール縮小などだけでは困難と見られているため，「戦略的高性能計算システム開発に関するワークショップ[3]」の作成するロードマップにおいては，power cap 可能なアークテクチャを当初から設計する方向が議論されている．つまり，プロセッサ・メモリなどシステム要素の最大電力の単純合計は電力制限を超えうるが (たとえば 20MW の制限に対して 30MW)，同時に稼働する電力が制限を超えないように，ジョブ特性に応じた制御を行う．このためにはアーキテクチャ，システムソフトウェア，アプリケーションの分野をまたいだ設計が必要となる．

謝辞 TSUBAME2.0 の運用は日本電気，ヒューレット・パッカード，NVIDIA，マイクロソフト，Mellanox，DDN をはじめとするベンダー連合との密な連携なしには成り立た

Vol.2011-ARC-197 No.12
Vol.2011-HPC-132 No.12
2011/11/28

ない．本研究の一部は科学技術振興機構戦略的創造研究推進事業「Ultra-Low-Powr HPC: 次世代テクノロジのモデル化・最適化による超低消費電力ハイパフォーマンスコンピューティング」，学術国際情報センター概算要求「スパコン・クラウド情報基盤におけるウルトラグリーン化技術の研究推進」，NVIDIA CUDA Center of Excellence の援助による．

参 考 文 献

1) The GREEN500 list.
`http://www.green500.org/`.
2) TOP500 supercomputer sites.
`http://www.top500.org/`.
3) 戦略的高性能計算システム開発に関するワークショップ.
`http://www.open-supercomputer.org/workshop`.
4) Massimo Bernaschi, Mauro Bisson, Toshio Endo, Massimiliano Fatica, Satoshi Matsuoka, Simone Melchionna, and Sauro Succi. Petaflop biofluidics simulations on a two million-core system. In *Proceedings of IEEE/ACM Supercomputing 11 (SC11) (accepted)*, page 11pages, 2011.
5) Takashi Shimokawabe, Takayuki Aoki, Tomohiro Takaki, Akinori Yamanaka, Akira Nukada, Toshio Endo, Naoya Maruyama, and Satoshi Matsuoka. Peta-scale phase-field simulation for dendritic solidification on the tsubame 2.0 supercomputer. In *Proceedings of IEEE/ACM Supercomputing 11 (SC11) (accepted)*, page 11pages, 2011.
6) Qinghui Tang, Sandeep K.S. Gupta, and Georgios Varsamopoulos. Thermal-aware task scheduling for data centers through minimizing heat recirculation. In *Proceedings of IEEE International Conference on Cluster Computing (Cluster 07)*, pages 129–138, 2007.
7) Qian Zhu, Jiedan Zhu, and Gagan Agrawal. Power-aware consolidation of scientific workflows in virtualized environments. In *Proceedings of IEEE/ACM Supercomputing 10 (SC10)*, pages 1–12, 2010.
8) 松岡 聡. TSUBAME 2.0 始まる．*TSUBAME e-Science Journal*, (2):2–8, 2010.
9) 松岡 聡, 遠藤 敏夫, 丸山 直也, 佐藤 仁, and 滝澤 真一朗. TSUBAME 2.0 の全貌. *TSUBAME e-Science Journal*, (1):2–4, 2010.
10) 遠藤 敏夫, 額田 彰, and 松岡 聡. スーパーコンピュータ TSUBAME 2.0 における linpack 性能 1 ペタフロップス超の達成. **情報処理学会論文誌コンピューティングシステム**, 4(4):169–179, 2011.