社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－ARC－162 （22）
2005－HPC－101 （22）
2005／3／8

クラスタの温度分布について

清水 敏行[†] 佐藤 聡[††] 児玉 祐悦[†††] 工藤 知宏[†††] 横川 三津夫[†††]

停電等の際に大規模クラスタを安全に停止させる為には，シャットダウン中の空調も問題となる．本稿では室温の2次元分布と時間的推移を実測し，空調機をバックアップすることなくシステムをシャットダウンさせる方式の安全性を確認した．

Temperature Distribution in the Cluster

Toshiyuki Shimizu[†] Satoshi Satoh[††] Yuetsu Kodama[†††] Tomohiro Kudoh[†††] Mitsuo Yokokawa[†††]

In order to stop a large-scale cluster safely in the cases, such as a power failure, air-conditioning under shutdown also poses a problem. In this paper, a 2-dimensional distribution and trends of room temperature is surveyed, and the safety of way which shuts a system down without backing up an air-conditioning machine is confirmed.

1. はじめに

産業技術総合研究所では，AISTスーパークラスタ（以下ASCと略記）と呼ぶ大規模クラスタを運用している（図1参照）．ASCはOpteron™2.0GHz×2構成のP-32ノード群1,072台，Itanium®2 1.3GHz ×4構成のM-64ノード群132台，Xeon™3.06GHz×2構成のF-32ノード群268台等からなり，定格総計で約800kWの電力を消費する．

一方ASCを設置したクラスタ室は37m×14m×2.7m，約1,400m³の容積で，定格1,000kWのUPS電源（1次電源遮断時の出力維持時間5分）と，総排熱容量1,320kWの空調機を備える．またこのクラスタ室内には，常に室温を監視し過熱等の異常事態に備える為に，温度センサを多数配置し室内の2次元温度分布を測定できるようにしている．

昨年9月，ASCの一般ユーザへの開放と本格運用を前にし，落雷等による1次電源遮断時のシステムの挙動と緊急事態下の熱設計を確認する為に，人為的に停電と等価な状況を作る実験（以下，停電実験と称する）を行った．その際，想定を越えた室温の上昇等に対処する為，前記の温度センサを利用し室内の2次元温度分布を測定した．

本稿ではこの温度センサ，及び各ノード内の温度センサ（CPUのコア温度，及び筐体内温度の測定が可能）について簡単に紹介し（2節），次にそれらを用いて測定した停電実験時の室温の分布やその時間的推移について報告する（3節）．また同じ装置を用いて，通常運転時の負荷による室温への影響について測定する（4節）．最後に5節では，停電等緊急事態におけるシステム停止の重要性について述べる．

2. 温度測定システムについて

2.1 半導体温度センサを用いた測定装置

停電実験の際の温度の測定では，多数の点を同時・継続的に測定でき，かつ後から結果を解析し易いように計算機可読な形式で保存できることが望ましい．また同実験の際にはASCを構成する計算機はすべてシャットダウンされるので，ASCとは独立したシステムでなければならない．

そこで汎用の半導体温度センサ[1]とA/Dコンバータ[2]，及びPCを組み合わせた温度測定システムを製作し，測定を行った（図2参照）．

温度センサは筐体間の通路の床上2.2mに計56個（7本の各通路ごとに1.4m間隔で8個）を配置した（図3及び4参照）．測定用のPC及び接続ユニットはASCのほぼ中央に配置し，センサはシールド・ケーブルで接続する．

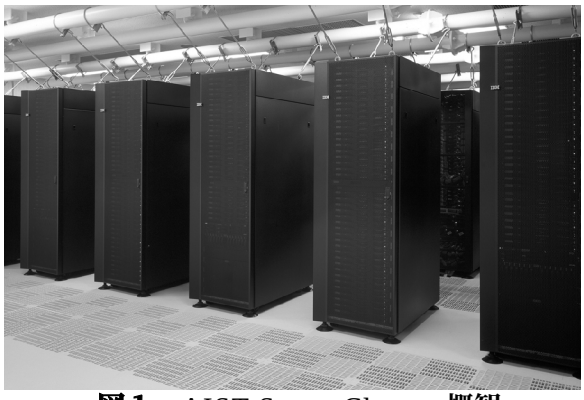

図1　AIST Super Cluster 概観

測定用のソフトウェアは1秒ごとにすべてのセ

†　(株)シナジェテック　SynergeTech, Inc.
††　(株)創夢　SOUM Corporation.
†††　産総研　AIST

[1] ナショナル・セミコンダクタ社製 LM35DZ
[2] コンテック社 製 AD16-16U(PCI)EH

図 2　温度測定システム外観

ンサを走査し，物理的位置に対応した温度表示を行う．各センサの測定値は温度に応じて色分けして一目で分布が分かるようにした．また測定結果のログ・ファイルはCSV形式で保存するので，Excel等での解析・グラフ化が容易である．

更に別途ログ・ファイルを読み出してネットワーク経由で表示するWebサーバを用意した．これにより，ブラウザさえ用意すれば任意のＰＣから現時点での温度をモニタすることが出来る．

2.2 ノード内蔵の温度測定機構

P-32ノード群を構成するIBM eServer325はベースボード･マネージメント･コントローラ[1]（以下BMCと略記する）を搭載し，2つのCPUのコア温度，及び筐体内の温度をモニタすることができる．BMCはノード本体とは独立に動作可能で，管理用のネットワークを介してスペア･ノード（64台の計算ノードからなるクラスタ･ユニット(CU)ごとに1台用意された予備ノード）と交信する．スペア・ノード上の測定プログラムは，サーバ運用システムxCATの機能を利用してCU内の64台の計算ノードに並列に要求を発行し，温度測定データを取得･蓄積する．測定プログラムが全BMCと交信するのに約10秒かかるので，ノード1台あたりの測定頻度は約10秒に1回となる．

2.3 空調機の配置

クラスタ室内の空調機は，1台あたり41.3kWの冷却能力の装置を16台ずつ北ゾーンと南ゾーン2群に分けて配置する．（図3参照）冷気はフリーアクセス床下を通り，適宜配置したメッシュ床板から吹き上げる．すべてのラックは前面から吸気して後方へ排出する方式のため，メッシュ床板はラックの前後面に配置し，床板に内蔵されたフィンの角度を調整して冷気の流量を調整している．また隣接するラック内で加熱された空気を吸い込まないよう，互いに向かい合わせとなる配置を採っている（空気の流れの方向は図3を参照のこと）．空調機の制御は，南北ゾーンごとに還流する空気の温度を測定して独立に行っている．

2.4 シャットダウン・シーケンスについて

何らかの原因で商用1次電源が遮断されると，UPSは蓄電池運転に切替ると共に停電したことを示す電気信号を発生する．この信号は1秒の遅延を経て停電信号分配器に至る（1秒未満の停電信号の場合は分配器に信号は伝搬しない）．同様に空調機に至る停電信号には2秒の遅延を挿入している．（図5参照）この設定により，短時間の停電に対する動作は以下のようになる：

1秒未満：　空調機，ASC共に停止しない
1秒以上2秒未満：ASCはシャットダウン，空調機は運転継続
2秒以上：　空調機，ASC共に停止

2秒以上の停電信号を受けて一旦空調機が停止すると手動により再起動させるまで停止したままになる為，上記のような方式で「空調機が停止す

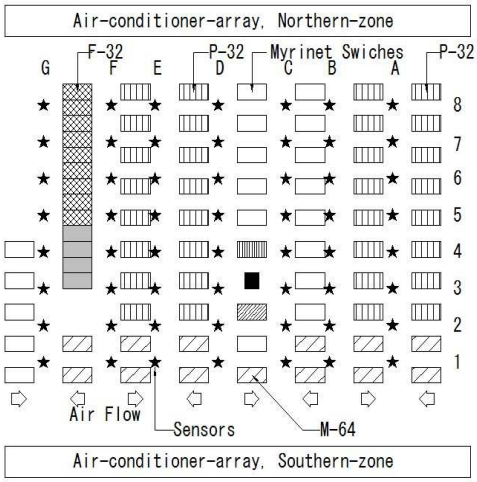

図 3　センサの配置

図 4　温度センサの取付け

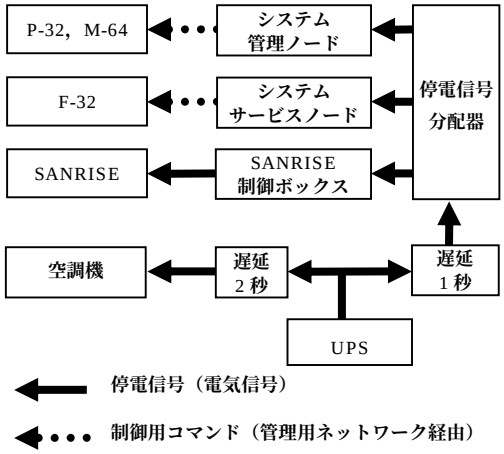

図 5　シャットダウン制御

るのにASCはシャットダウンしない」と言う最悪の事態を避けることができる．また2つの遅延回路の設定が異なるのは，誤差を考慮した為である．

停電信号分配器によって3つに分岐した停電信号は，システム管理ノード（P-32及びM-64），システムサービスノード（F-32及びストレージ），及び二次記憶装置（SANRISE）に接続されている．前2者はシステム管理用のノードで常に停電信号を監視しており，同信号を受け取ると自動的に配下の計算機群に対してシャットダウン用のコマンドを発行する．これにより各計算機群はそれぞれの主電源のOFFまでを自動的に行う．

これに対してSANRISEは外部から電源の遠隔制御ができない仕様のため，別途電源をON/OFFする為の装置を追加し，停電信号を受けてSANRISEの電源を強制的にOFFにするようになっている．

3. 停電実験とその結果について

3.1 実験の目的と手順

停電実験の目的は以下の3つである：

(1) 停電時動作の仕様と実際の動作の確認
(2) 停電直後の室温の変化の測定
(3) シャットダウン後に残る機器の消費電力と室温の推移の測定

また実験に際しては，フロア全体を停電させると影響が広範囲に及び過ぎるため，以下のような方法で停電の状況を模して行う：

A) UPSの1次電源入力を切放
B) 空調機を全数停止

上記操作はすべて作業員を配置して手動で行う．A)により図5に示した経路を経て停電信号とシャットダウン･コマンドが伝搬し，ASCを構成するすべての計算機は最終的に自らの主電源を切って待機状態になる．この動作に要する時間とその間の室温の変化を実測し，温度が危険なレベルに達しないか確認する（上記目的(1)及び(2)）．

また主としてネットワーク関連の装置群はシャットダウン及び電源制御の機能を持たないため，UPSからの給電が続く限り動作を続ける．そこで復電までに時間がかかる場合を想定して，これらの機器が消費する電力と，それによる室温の変化を実測する（上記目的(3)）．

3.2 温度上昇見積

ASCの設計段階では，クラスタ室において仮に空調が全停止状態でASCの運転を続けた場合に，温度がΔT℃上昇するのに要する時間hを以下のように見積もった：

$$h = \frac{M \times k \times C \times \Delta T}{860 \times W \times e}$$

室の容積M：	37m×14m×2.7m=約1400m^3
空気の比重k：	1.16kg/ m^3
空気の比熱C：	0.24kcal/kg℃
温度上昇ΔT：	10℃
W/Cal変換係数:	860
熱量W：	800kW
熱効率e：	0.5

10℃の温度上昇に要する時間hは約41秒である．室内に占める機器の体積や壁や天井を通しての熱の交換を無視している点や熱効率の見積り等近似が多いが，オーダとしては妥当であると考える．また室内での強制的な空気の流れが止まることを考慮すれば，機器近傍の温度上昇はこれを上回る可能性がある．したがって停電に際して温度がどのように推移するのかを確認することはきわめて重要である．

3.3 無負荷時の温度変化

停電を模した最終的な実験に先立ち，ASCを無負荷状態にして空調機の半数（図3の上辺に相当する位置の16台）のみを停止する実験を行った．これは空調機の停止と再起動にかかる時間を確認するための実験を兼ねている．その結果を図6～7に示す．

図7は空調機の停止（11:09:35）を含む10分間の温度変化を捉えたグラフで，温度が4度以上変化した測定点（55[3]点中18点）のみをプロットした．その中でも特に期間中他と比較して高温を示したA4，E8（何れもセンサ位置，以下同様），最大の温度変化（8.6℃）を記録したD8，及び単位時間あたりの上昇率が最も大きかった（0.5℃/s）D6を太い線で表す．尚，以上の4つの測定点は必ずしも停止した空調機に近い側ではないが，何れ

[3] A/D コンバータに障害が発生した F6 センサを除く．

もP-32の近傍である．

また図6は恒常的に最も高い温度を示すことが多いE8センサでの測定値が最大となる時刻(空調停止から約2.5分後の11:12:02)での温度分布である．やはり停止した空調機に近い方（同図中右側に相当）で比較的高い温度の偏りが観測されるが，余り顕著ではない．

尚，空調機の操作試験では，OFF/ON操作にそれぞれ約10秒，また再起動後の全台数稼動までには約3.5分かかることを確認した．

3.4 全負荷時の温度変化

図8は，最終目標の実験，即ちUPSの1次電源入力を切ると同時に空調機を全停止させた時の温度の変化の様子である．A/Dコンバータの入力回路に障害が発生し測定値が安定しないF6センサ以外のすべての測定値をプロットしている，すべての計算ノード群には予めLinpack等最も負荷の高い（発熱量の大きい）プログラムを実行させておき，13:46:20に模擬的停電の開始を指示した．その約30秒後の13:46:50付近を境に殆どの測定点で急激な温度上昇を記録しているが，この冷気の供給源が失われたことによる温度上昇は30秒程度で鈍り始める．これは最も大きな熱源であるP-32の計算ノード群のシャットダウンと主電源OFFによる発熱の減少が原因であると考えられる．ただしこの30秒間での温度上昇の最大値はA3センサにおいて6.0℃に達した．またこの期間での温度上昇値を外挿して，仮に空調が停止したにもかかわらず自動シャットダウンが発動しなかった場合に最も早く気温が45℃を突破する測定点は，A4において約86秒後であった．これはUPSの容量に比べて十分短い時間である．

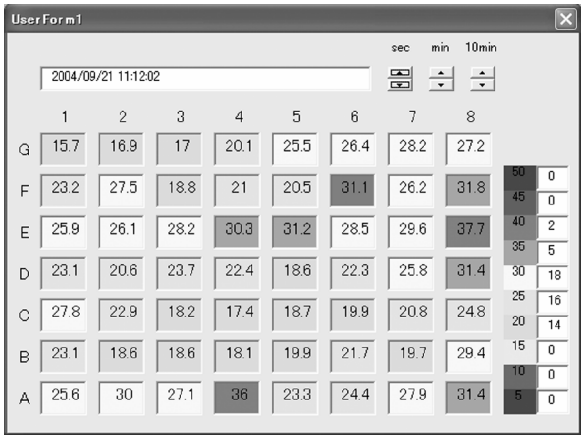

図 6　無負荷時に空調を短時間停止した場合の温度分布

更に図8中の他のセンサ番号は，測定期間中に一定期間最も高い（C5については最も低い）値を観測した点のプロットを示した．

また実験開始後に温度が30℃を超える測定点の数が最も多くなる13:47:40時点（停電開始指示の1分20秒後）での温度分布を図9に掲げる．やはりP-32計算ノードの排気が集中するA及びE列に高温を示す測定点が多い．また図10は空調機の再起動を指示した瞬間の13:58:20時点での温度分

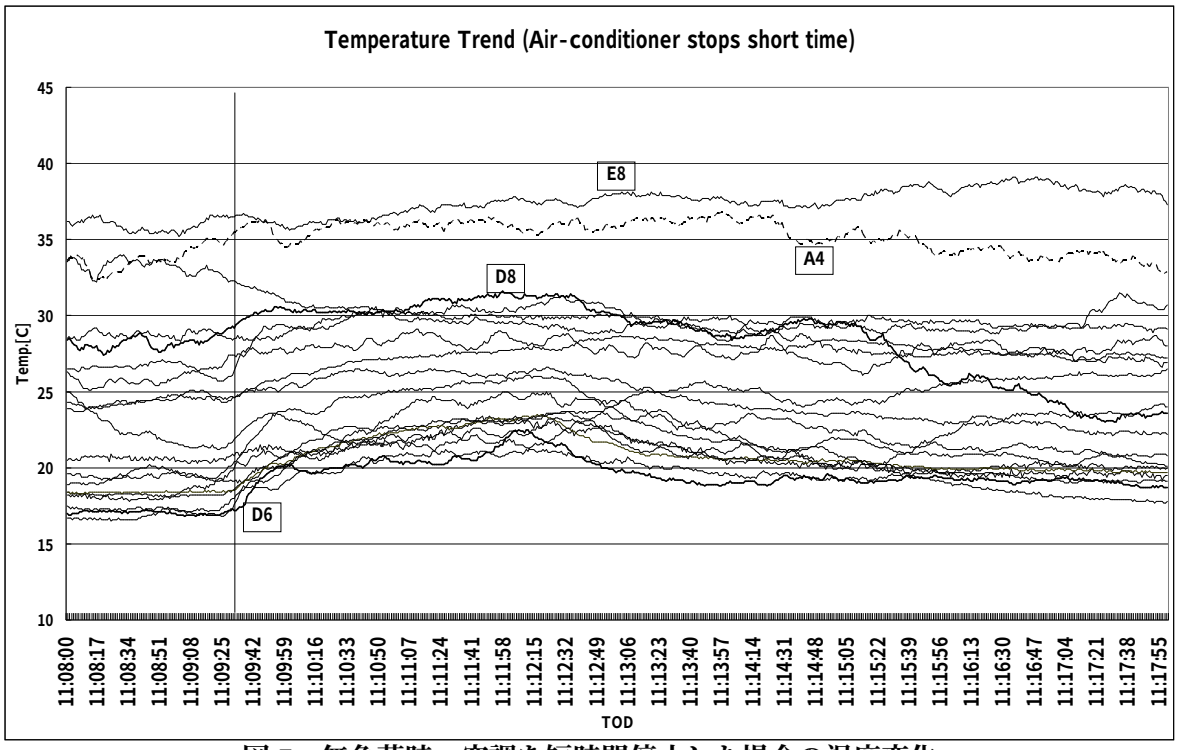

図 7　無負荷時，空調を短時間停止した場合の温度変化

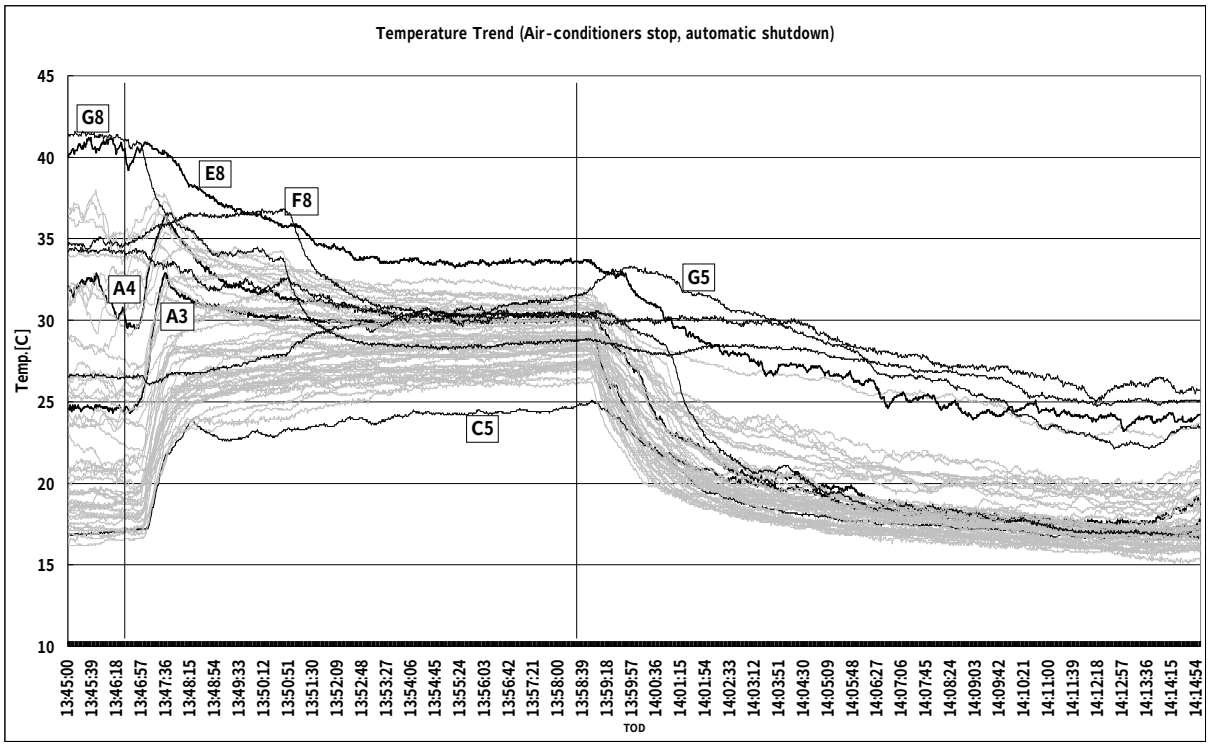

図 8　空調全停止，シャットダウン実行時の温度変化

布である．35℃を超えるような高温の点がなくなった代りに，ほぼすべての測定点で25℃以上を示し，分布が平均化している様子が見られる．

3.5 シャットダウンに要した時間

停電信号発生からP-32が電源OFFに至るまでの所要時間は30秒，M-64については36秒であった．またF-32については3分43秒かかった．F-32の所要時間が他に比べて大きいのは，配下のファイルサーバをマウントしている他のノードがシャットダウンを完了するのを待つため，無条件に3分の待ち時間を入れていることによる．

3.6 停電後の温度変化

すべてのノードが主電源をOFFにした後も，Myrinet switchをはじめとするネットワーク機器は（UPSによる給電が続く限り）稼動し熱を放出し続ける．実際，すべてのノードの電源がOFFした後の14:00時点でのクラスタ室内の分電盤上の電力計の読みは121.4kWであった（ASCに含まれない小型クラスタの消費電力20kWを含む）．この影響を見積もる為に，平均温度が斬増し始める13:53:40から5分間の上昇を見ると0.35℃であった．これより約3.8時間後に平均温度が45℃を超える計算になるが，これは前述の消費電力から算出されるUPSの残存容量の4倍に相当する時間であるので，実際にそこまで温度が上昇することはないと考えられる．

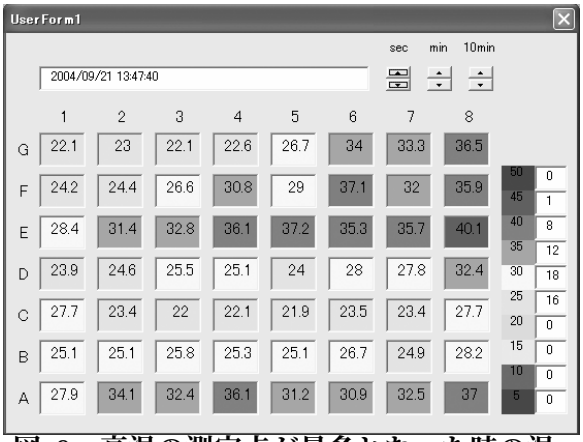

図 9　高温の測定点が最多となった時の温度分布

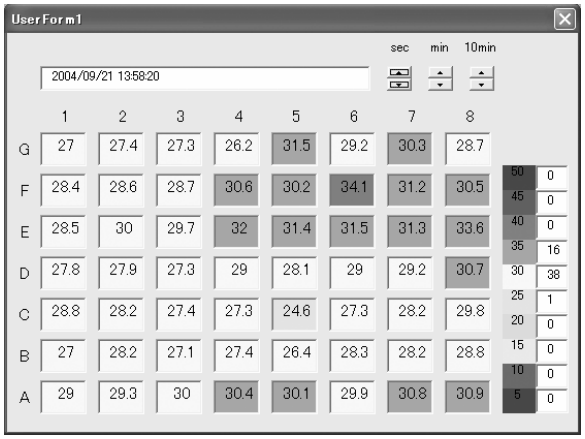

図10　空調機再起動直前の温度分布

4. 平常時の温度分布について

4.1 負荷と温度の変化

3.3で述べた無負荷時の空調停止実験の直前のF6センサを除く55点の単純平均値は22.4℃であった. これに対して3.4の全負荷時のそれは24.2℃であった. したがって負荷に伴う平均気温の上昇分は1.8℃となる.

これに対して別途実施したM-64の全ノードを用いたLinpackの実行では, 気温に有意な変化は観測されなかった.

そのため2.2で述べたノード内蔵の温度測定機構を用い, 負荷（浮動小数点演算を多用する量子化学分野の計算）をかけた時のCPUコアと筐体内部の温度変化を観測した. （図11参照）測定対象はP-32クラスタ内の連続した32ノードで, コア温度の上昇の最大値は12℃であった. これに対して筐体内温度の上昇は最大でも2℃であり, 更に温度センサを用いた観測では, 当該ノード付近の気温に有意な変化は見られなかった. また観測対象ノードは同一ラック内にあって互いに隣接しているが, ラック内の位置と温度に相関は見られなかった.

5. 結論

3.2で述べた事前の温度上昇見積は41秒間で10℃の上昇であった. これに対して実測は30秒間で6℃の上昇であった. 前者を30秒当りに換算すると7.3℃となり, 機器の体積等を勘案すると妥当な範囲の誤差であると考える.

また実測では実験開始後約1分で計算ノード群の電源がOFFになったことの効果が現れ, 急激な温度の上昇が抑えられることを確認した. このことからも, 停電（とそれに伴う空調機器の停止）に際しては, いかに早く・確実にノード群をシャットダウン及び電源OFFに導くかが重要である. もちろん各ノードには, 過熱を防止する為に自身の温度をモニタして必要に応じて電源をOFFする為の装備がなされているが, これらは実際に温度が所定のレベルに達しない限り発動しないことに留意されたい. システムを安全かつ確実に停止させるには, あくまで停電信号に基づく自動シャットダウン･シーケンスが確実に動作することが重要である.

一方, シャットダウンが完了した後に残る機器群については, 前述したようにその消費電力によってクラスタ室内の温度が危険なレベルに達するには3〜4時間かかると考えられる. これに対してUPSの容量は定格出力（1MW）で5分間であるので, 測定された消費量では1時間弱で電源断となり, 過熱の心配はないと考えられる.

一般に計算機システムでは停電等の事態に備えてUPSによるバックアップを行う. この際, 理想的には空調機も含めてバックアップを行えば, シャットダウンの間も空調が維持され, システムを危険に晒す恐れはない. しかしそのためにはUPSで賄う電力が増加し, 膨大なコストの増加が見込まれる. 特にクラスタのような単位床面積あたりの電力密度の大きいシステムの場合, この傾向は顕著となる.

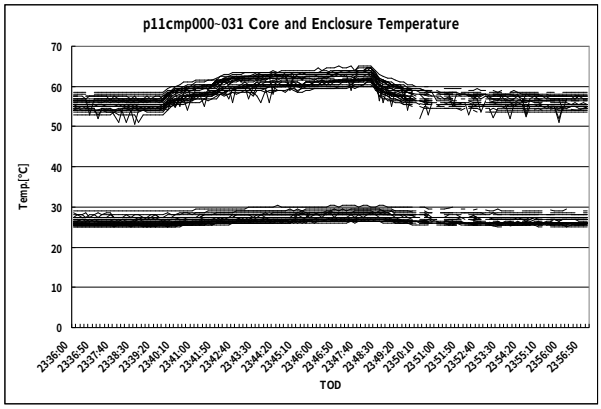

図 11　BMC 経由で測定した P-32 ノードのコア(上)と筐体内(下)温度の変化

そこでASCの場合は予め室温上昇を見積もった上で, 室温が危険なレベルに上昇する前にシャットダウンを完了する方式を採用した. そして今回の実験でその効果を確認し, 停電による空調停止の際には, 可及的速やかにシステムをシャットダウンに導くことで安全に対処できることを示した.

謝辞

最後になりましたが, 本実験の準備と実施に当っては, 以下の部署・会社の方々の全面的なご協力を得ました. あまりに多数の方々ゆえここでは本実験担当の代表者のお名前のみ記しますが, ご協力いただいたすべての方に, この場を借りて改めて御礼申し上げます（順不同）:

日本IBM（株）ITS事業部　　吉田 未樹 氏
日本SGI（株）カスタマー・サービス事業本部
　　福世 和雅 氏
産総研 施設管理室　　田村 治男 氏

参考文献

[1] http://www.redbooks.ibm.com/redpieces/pdfs/sg246495.pdf