Development of detection algorithm of bipolar waveforms around the moon using a parallel and distributed workflow “Pwrake”

第22回年次大会予稿

分散処理用タスクスケジューラ Pwrake を用いた月周辺電界波形データからのバイポーラ型波形の抽出アルゴリズムの開発

Development of detection algorithm of bipolar waveforms around the moon using a parallel and distributed workflow “Pwrake”

矢木大介[1*]，村田健史[2]，笠原禎也[1]，後藤由貴[1]
Daisuke YAGI[1*], Ken T. MURATA[2], Yoshiya KASAHARA[1], Yoshitaka GOTO[1]

1 金沢大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Kanazawa University
〒920-1192 石川県金沢市角間町
E-mail: yagi@cie.is.t.kanazawa-u.ac.jp
2 情報通信研究機構
National Institute of Information and Communications Technology
〒184-0015 東京都小金井市貫井北町4-2-1
E-mail: ken.murata@nict.go.jp
*連絡先著者 Corresponding Author

月探査衛星かぐやで観測した電界波形データには，特徴的なバイポーラ型波形が多数確認されている．このバイポーラ型波形を観測データから自動抽出し，波形の特性を求めるアルゴリズムを開発中であるが，観測データの総容量が約 190GB にも及ぶため，情報通信研究機構(NICT)のサイエンスクラウド上で Pwrake を用いて並列処理による高速化を図った．本論文では効率を評価した結果を報告する．

Characteristic bipolar waveforms were frequently observed by the electric waveform receiver onboard the lunar orbiter named KAGUYA. An algorithm to extract these bipolar waveforms is now under development, but the total amount of the waveform data is about 190GB and it is too huge to apply the algorithm on a general PC workstation. In the present study, we attempt to accelerate computation time by a parallel and distributed workflow named “Pwrake” implemented on NICT science cloud. We report the evaluation results of the efficiency of the data processing.

キーワード: 月探査衛星かぐや，波形捕捉器, NICT サイエンスクラウド，並列処理
KAGUYA spacecraft, Waveform Capture, NICT Science Cloud, parallel processing

1 はじめに

地球観測，宇宙環境の解明などの目的で毎年様々な科学衛星が打ち上げられている．科学衛星で得られるデータは観測価値が高く，可能な限りのデータを収集するため長期運用される衛星では，膨大なデータが地上に蓄積される．

2007年9月に打ち上げられ，2009年6月に月面に制御落下した月探査衛星かぐやには，当研究グループが開発・解析を担当する月周辺のプラズマ波動を観測する波形捕捉器（以下、WFC）が搭載されている[1]．特に100Hz～100kHzの電界波形を観測するWFC-Lでは，いくつかのパターンに分類できる特徴的なバイポーラ型の波形が多数確認されている．WFC-Lは250kHzのサンプリング周波数で波形データを取得し，運用期間中に取得されたデータ総量は約190Gにも達する．我々は現在，この波形データからバイポーラ型波形を自動抽出するアルゴリズムを開発し，網羅的にバイポーラ波形の特性を解明することを計画中である．しかし開発中の波形抽出アルゴリズムは，汎用PCワークステーションを用いて，全観測時間に対し適用すると，1週間近い処理時間を要する．観測データにはバイポーラ波形以外の様々な自然波動も重畳しており，これらを除去しつつ，抽出対象となるバイポーラ波形を精度よく検出・分類するために種々のアルゴリズム改善が試みられている．しかし上述のように1回の試行に必要な計算量が多いことから，アルゴリズム改善のターンアラウンドが悪いことが課題である．

情報通信研究機構（以下、NICT）が次世代の科学研究環境を提供するために構築したNICTサイエンスクラウドでは，大規模データ処理向けの分散データ処理サーバが用意されている．これまでに，22年にわたる長期観測を達成したGEOTAL衛星の観測データに対して，従来のデータ処理環境に比べて100倍以上の高速処理を実現した実績がある[2]．NICTサイエンスクラウド上で並列化処理のタスクスケジューラツールであるPwrakeは，元のプログラムを改変することなく簡便に並列化を実装できることが特徴である．そこで本研究では，NICTサイエンスクラウド上でワークフローシステムPwrakeを用いた並列分散処理による波形抽出処理の高速化を行った．

2 かぐやプラズマ波動データ処理

2.1 かぐやプラズマ波動データ

かぐや衛星に搭載されたWFC-Lで取得されたデータの形式について説明する．図1はWFC-Lで取得した波形データの例である．WFC-Lは，かぐや衛星が月の南北の極を結ぶ極軌道を高度100kmで，1周回2時間で周回していた定常運用時に，衛星に搭載された一本15mの直交2軸ダイポールアンテナを用いて，100kHz以下の電界波形を計測した[1]．全計測データを地上伝送できないことから，1回あたり最大750,000点，時間長にして1～3秒間の連続波形データを間欠的に取得し，地上ではCDF(Common Data Format)[3]と呼ばれるファイル形式で保存されている．CDFはNASAが開発した自己表現型式データフォーマットであり，ランダムアクセスや欠損部における容量節約など科学衛星の観測データを利用する上で利便性の高いデータ形式である．

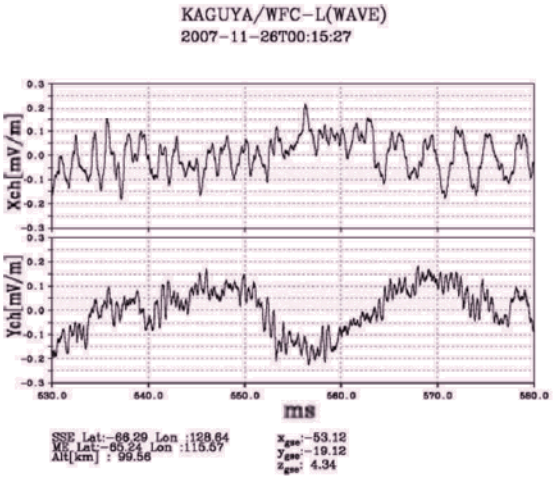

図1 WFC-L波動データ描画例

2.2 バイポーラ波形抽出法

波形抽出の概要を以下に示す．WFC-Lで計測された波形データには，様々な種類の自然波動が重畳している．一般にバイポーラ波形は，周波数領域に変換すると広帯域にわたるスペクトルを持つため，これと周波数帯が重畳する別種の自然波動との分離が困難である．現行の抽出法では，前処理として，対象とするバイポーラ波形より十分高い高周波成分と，バイポーラ波形にオフセットを与える低周波成分を除去するバンドパスフィルターを通す．次に，信号の振幅に閾値を設け，前後の時間帯の波形の振幅に対して卓越した点をバイポーラ波形の候補点とした．最後に，前項で決定した候補点を基準に，バイポーラ波形の始点と終点を決定し，抽出したバイポーラ波形の時刻，時間幅，振幅値などのパラメータを記録する．

このような手順で実際に抽出した波形の例を図2，3に示す．図2に示す波形は，パルスの形状が前半と後半で対称な事例で，一般に静電孤立波形(ESW)と呼ばれる波動現象である[4]．これまでの解析では，波形の時間幅は約2ms程度のものが支配的である．これに対し，図3に示す波形は，パルスの形状が前半と後半部分で異なる事例である．前半部分に比べ後半部分の時間幅が長いという特徴があり，波形の時間幅も約5ms～20msが多い．この波形は，宇宙空間で励起した自然波動ではなく，何らかの原因でアンテナと衛星間で電位差が生じる現象を捉えたものと考えられ，その物理過程が現在も研究対象となっている．

このようにWFC-Lから抽出されるバイポーラ波形は，月周辺の宇宙空間のプラズマ物理の解明のための貴重な手がかりであるが，冒頭で述べたように，網羅的な解析を実現するには，個々の波形の高精度な抽出が必須である．しかし汎用のPCワークステーションでは，1回の試行に約1週間を要し，振幅や時間幅が多岐にわたるバイポーラ波形を精度よく抽出するためのアルゴリズムや抽出パラメータのサーベイには適さない．そのため，より豊富な計算リソースを用いて波形抽出処理を高速化できる環境が必要である．

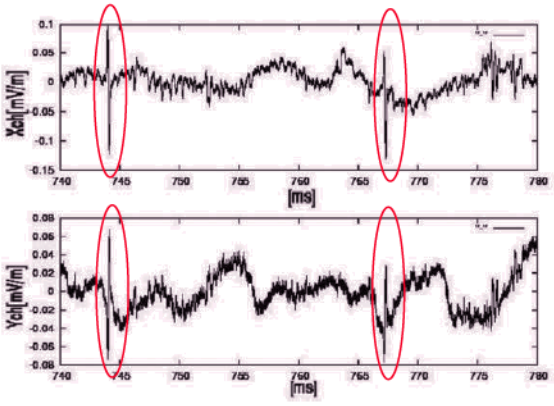

図2 ESW抽出例

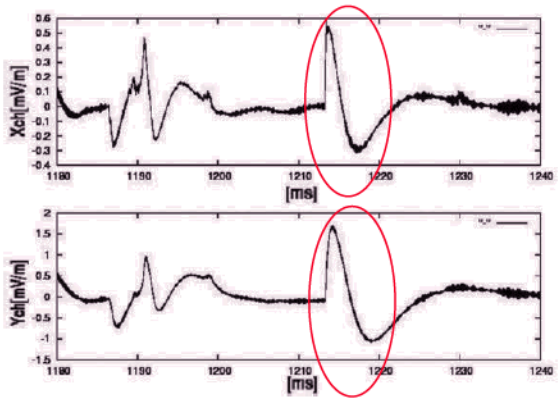

図3 非対称波形抽出例

3 クラウドによる波形抽出処理環境

3.1 波形抽出対象データ

今回波形抽出処理の対象となる波動データは，かぐや衛星の定常運用期間で得られた2106ファイル(142GB)である．WFC-Lの観測は前述のように間欠的に行われるため，個々のCDFファイルの大きさは，小さいもので3MB，大きいもので300MBと，容量が大きく異なる．そのため，ファイルによって処理タスク時間が大きくばらつく．これに対応するため，異なるサイズのデータ処理を効率的に行うためのヘテロ並列処理技術が必要となる．

3.2 NICTサイエンスクラウド

NICT サイエンスクラウドは NICT が科学研究者向けに，次世代の科学研究環境を提供するために開発されたクラウドである[5]．高速ネットワークにより各拠点を結び，大容量ストレージや分散データ処理サーバなどの様々なリソースを使用することができ，高度な科学研究を可能とするクラウド基盤である.NICT サイエンスクラウドの高速ストレージと大規模並列データ処理サーバを複数使用し，波形抽出処理の並列処理を行った．使用するサーバのスペックを表 1 に示す．本研究で利用する NICT サイエンスクラウドのサブセットである計算システムを図 4 に示す．ユーザはゲートウェイサーバにネットワーク経由でアクセスし，大規模データ処理サーバに ssh 接続することでサービスを利用できる．本研究で使用したサーバは n100～n109 までの 10 台である．それぞれ 12 コア CPU から構成されており仮想的に 24 コアまでを利用できる。またメモリは 71GB/node であり，HDD は 1.4PB である．各サーバは並列ストレージシステムに GPFS (General Parallel File System) プロトコルで接続されており，ストレージ内に格納されている波形抽出処理のソースファイル，WFC-L データが記録された CDF ファイルを読み取ることができる．GFPS は IBM 社が開発したクラスタコンピュータ向けの分散ファイルシステムである．ヘテロ並列処理を行うには，これらのサーバ群にワークフローを与え，タスクを割り振る必要がある．本研究で用いたのは Pwrake と呼ばれるワークフローツールである．Pwrake については次章で説明を行う．

3.3 Pwrake

Pwrake(Parallel Workflow extension for Rake)は，Ruby 言語で記述されたビルドツールである Rake を，複数マシンを用いた並列分散処理向けに拡張したものである[6]．主な拡張点は，SSH によるリモート実行，Gfarm ファイルシステムへの自動マウント，ローカリティを考慮したタスク配置(Affinity scheduling)などである．Rake との互換性が図られており，使用するノードやコア数を指定する

と Rakefile で記述したワークフローをそのまま並列分散実行が可能である.今回は，1 つの CDF ファイルと波形処理命令を 1 つのタスクとし各ノードに割り振り，処理が終わったコアに順次タスクを与えるヘテロなワークフローを与え，並列処理を実装した.

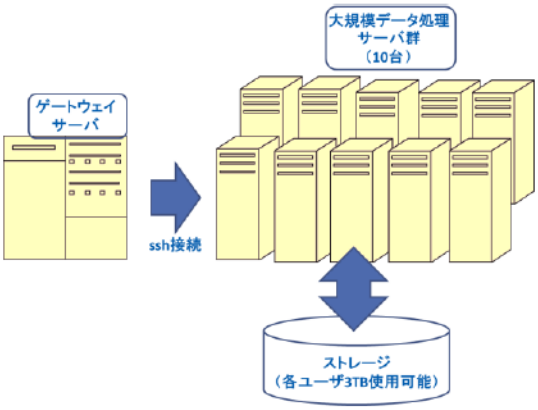

図4　並列処理システム構成図

4 波形抽出処理の高速化結果

Pwrakeを実行する際に使用するノード数・コア数，また処理するファイル数を変化させながら処理時間を測定した．まず，試験的にCDFファイル20個分(2339MB)について並列処理環境で抽出処理を行った結果を示す．各ノードのコア数を固定し，ノード数を変化させた．測定環境及び測定結果を表1に示す．この結果から，ノード数を増やしていくことで高速化が実現できていることが確認できた．また，使用するノード数に比例して処理時間が高速になっていることが読み取れる.

次に，クラウド環境で利用できる最大計算リソースである 10 ノードを用い，コア数を変化させ測定を行った．測定環境及び測定結果を表 2 に示す．最大計算リソースである各 20 コア 10 ノードで測定を行う際は，全観測データの 2106 ファイルについて処理を行ったが，計算リソースを減らした際に全観測データに対して処理を行うと時間がかかりすぎるため，段階的にファイル数を減らし，測定を行った．全計算リソースを用いて全観測データを処理した結果は，1 時間 38 分 41 秒であった．各 2 コア 10 ノードの測定結果から，1 ノード 1 コアでの全観測データに対する処理時間は約 230 時間かかると推測されるため，20 コア 10 ノードを用いることで約 1/150 まで短縮できることが確認できた.

また各 2 コア 10 ノードでの処理時間を基準として，その他のコア数の処理時間を比較したグラフを図 5 に示す．12 コア 10 ノードまでは並列化効率はほぼ一定であり，12 コア 10 ノードにおける効率は 86%である．一方，12 コア 10 ノード以上に計算リソースを増やすと，高速化の効率が悪くなることが確認できる．24 コア 10 ノードにおける並列化効率は 60%である．これは，1 つのノードが 12 コア以上のリソースを使うとハイパースレッドの影響が現れ始めるためだと考えられる.

表1 コア数を固定しノード数を変化させた際の測定環境及び測定結果

ノード数	各ノードのコア数	処理時間
1	1	0:52:42
2	1	0:26:59
3	1	0:18:54
4	1	0:14:50
5	1	0:12:42

*CDFファイル20個分(2339MB)が処理対象

表2 10ノードでコア数を変化させた際の測定環境及び測定結果

各ノードのコア数	ファイル数	ファイルサイズ(MB)	処理時間
2	960	70,918	5:22:36
4	960	70,918	2:43:13
8	960	70,918	1:27:46
12	960	70,918	1:02:51
16	1280	89,315	1:15:55
20	1600	111,920	1:26:16
24	2106	144,526	1:38:41

*10ノードで固定して測定

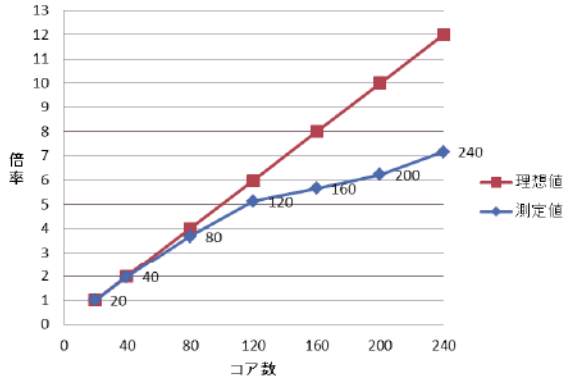

図5 10ノード2コアを基準とした各コア数の処理時間と理想値

5 おわりに

今回はNICTサイエンスクラウドの並列処理環境を用いて，かぐや衛星のプラズマ波動データから特徴的なバイポーラ型の波形を抽出する処理の高速化を行った．汎用PCワークステーションでは膨大な処理時間を要するため，抽出アルゴリズムの改良・試行や，抽出に最適なパラメータサーベイを行う効率が問題であった．本研究では，波形抽出効率の大幅な改善を目指して，NICTサイエンスクラウドを利用し，その効率について評価した．波形抽出を行う波形データを格納したCDFファイルは個々ファイルサイズが大きく異なるため，タスクスケジューラPwrakeを使ってヘテロ並列処理を行った場合の処理時間を測定した．その結果，最大計算リソースである10ノード各24コアを用いて並列処理を行った場合，汎用PCワークステーションによる処理に比べて約1/150の時間で処理を終えることが実証できた．

今後の並列処理による高速化の展望について述べる．今回は，並列ファイルシステムとしてGPFSを用いたが，GPFSはI/Oがボトルネックになり，スケーラビリティに限界がある．そのため，I/Oを並列化することで高速なデータI/Oが行うことができ，Pwrakeと連動しているGfarmを用いることで更なる高速化及び，より大規模なデータへの適用が期待できる．

参考文献

[1] Y. Kasahara et al., Earth, Planets and Space, 60(4), 341-351, 2008.

[2] Ken T.Murata et al., IEICE Communications Express, 3(2), 74-79, doi:10.1587/comex.3.74, 2014.

[3] CDF, http://cdf.gsfc.nasa.gov/

[4] K. Hashimoto et al., Geophys. Res. Lett., doi:10.1029/2010GL044529, 2010.

[5] Ken. T. Murata et al., Data Science Journal, 12, WDS139-WDS146, 2013.

[6] 田中昌宏，建部修見，宇宙航空研究開発機構研究開発報告：宇宙科学情報解析論文誌, Vol. 1, JAXA-RR-11-007, 67-75, 2012.