54

NWTジョブにおける処理性能の改善

土屋雅子[*1]、吉田正廣[*1]、中村孝[*1]、藤岡晃[*2]、山口靖[*3]

The Improvement of the Performance on NWT Jobs

by

Masako TSUCHIYA, Masahiro YOSHIDA, Takashi NAKAMURA

National Aerospace Laboratory

Akira FUJIOKA

Sanko Software Development Co.

Yasushi YAMAGUCHI

Fujitsu Limited

ABSTRACT

NWT (Numerical Wind Tunnel) is a high performance parallel computer. It has served at the very high CPU utilization-rate. To make the best use of NWT we have investigated the behaviors of input/output(I/O) on NWT jobs, and have found that NWT jobs have a serious problem to deteriorate the performance. It is in the output process of Fortran unformatted sequential writing. We have found the scheme that improves the performance, and we recognize the importance of I/O performance on NWT jobs.

In this paper we show the statistical and actual surveyed data of the I/O performance on NWT jobs, and the details of the method to solve this problem.

1．はじめに

平成５年２月に導入した数値風洞（Numerical Wind Tunnel:ＮＷＴ）は要素計算機（Processing Element:ＰＥ）にベクトル計算機を配置する分散主記憶型並列計算機システムである。ＮＷＴは大規模数値シミュレーションの計算エンジンとして絶大な力を発揮し、航技研における先進的な航空宇宙技術の研究開発に重要な役割を果たしている。

現在、ＮＷＴへのジョブ処理要求は増大の一途をたどり、システムの処理待ちジョブキューは常時、満杯状態を呈している。ちなみに平成１０年では、システムのｃｐｕ稼働率は９０％に迫り、並列計算機システムのセンタ運用システムとしては比類のない高稼働率となっている。

ＮＷＴの運用では、日頃からユーザの個々のジョブについて実効性能を高め、ジョブ・スループットの増大を図る種々な運用方策に取り組んでいる。

その方策の具体例を列挙すると以下のとおりである。

（１）入出力時間が長いジョブのチューニング

（２）低効率並列ジョブのチューニング

（３）低ベクトル化ジョブのチューニング

ＮＷＴジョブに対するこれらの方策はまだ順次進行中のものもあり、その効果は明確な数字となって響いてきてないが、実運用の中で実質的な利益を生み出した事例を多く確認している。

以上に示すチューニングを要するジョブは主に日々の膨大に発生するプロセス課金レコード、各種ログ情報ならびにシステム稼働情報等を分析した結果から抽出している。また、システム運用担当者が日々のジョブ走行を監視する中で確認する場合も多々ある。さらに、ＮＷＴではシステムに網を掛け、積極的に効率の悪いジョブを発見し、ときには、このようなジョブの実行阻止を図っている。

ＮＷＴシステムの更なる高度有効利用を図るため、実行ジョブ（ＮＷＴジョブ）の入出力性能について着目し、システムの隘路ならびに改善すべき要因の調査を行った。この結果、ＮＷＴジョブの入出力において、最も利用の中心となる書式なしＦＯＲＴＲＡＮレコードの出力処理に性能低下を確認した。本報告では、ＮＷＴジョブの入出力性能についての調査結果を示すとともにその中から確認された性能低下の要因と考えられるシステムの隘路に対する改善

策を提示し、その有効性を示す。

2．NWTの稼働実績

2．1ハードウェア構成概要

NWTを中核とする航技研の数値シミュレータシステム（NSシステム）のハードウェア構成概念図を図2．1に示す。

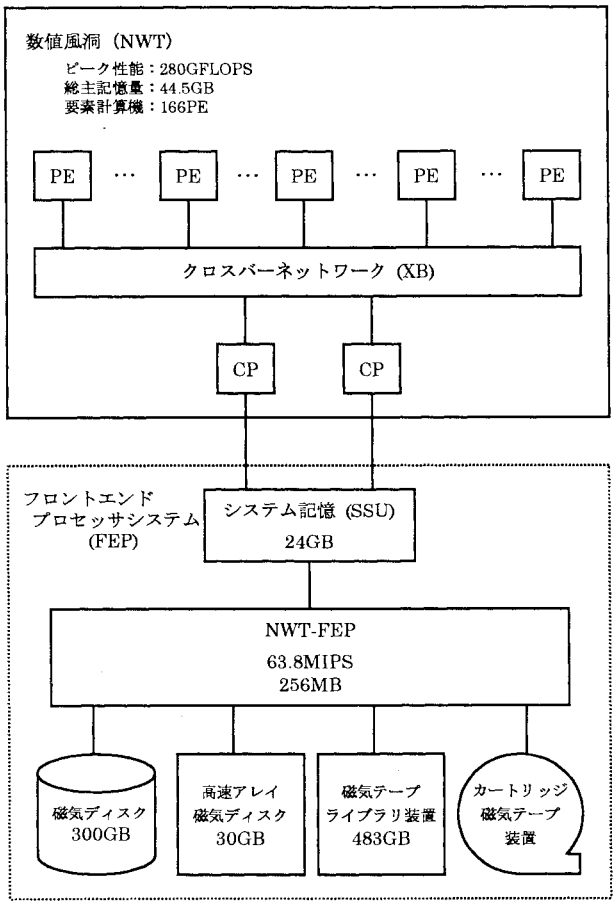


図2．1 数値シミュレータシステムのハードウェア構成概念図

同図に示すとおり、NWTの中核部は2台のコントロールプロセッサ（CP）、166台の要素計算機（PE）、およびそれらを相互接続するクロスバネットワーク（XB）のハードウェアから構成されており、フロントエンドプロセッサ（FEP）のNWT－FEPとシステム記憶装置（SSU）を介して接続されている。PEの構成台数のうち162台は256MBのメモリ容量を有し、残りの4台は1GBのメモリ容量を実装している。NWTはピーク性能280GFLOPSの超高速処理性能を有する並列計算機システムであり、そのハードウェアの処理性能を十二分に引き出すために、大規模ジョブの数値シミュレーション処理に専念し、ジョブ入出力等のフロントエンド処理はFEPが分担する。また、FEPには入出力装置および補助記憶装置として磁気ディスク装置(転送速度：4.5MB/秒、同時アクセスパス数：20)、高速アレイ磁気ディスク装置(転送速度：36MB/秒)、磁気テープライブラリ装置（転送速度：3MB/秒、同時アクセスパス数:4）ならびにカートリッジ磁気テープ装置（転送速度：3MB/秒）が接続している。

2．2NWTの稼働実績

平成5年度のNWT導入当初より現在までのNWTジョブの稼働実績を以下に示す。表2．1はNWTジョブの処理件数をジョブが使用するPE台数ごとに示したものである。表2．2はNWTジョブの処理に要した総CPU使用時間を表したものである。

表2．1 NWTジョブの処理件数

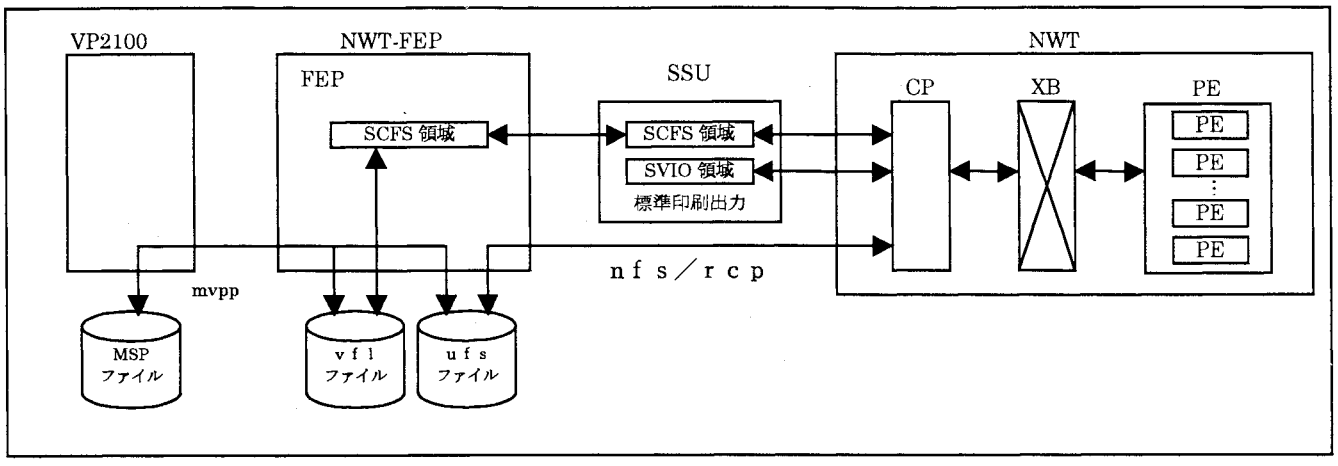
	平成5年度	平成6年度	平成7年度	平成8年度	平成9年度
1PE	42,121	51,109	51,194	75,384	81,599
2-7PE	8,136	7,030	13,414	16,209	35,098
8-15PE	3,723	4,017	4,336	11,520	19,110
16-31PE	4,226	7,099	8,580	17,937	12,632
32-63PE	167	754	2,215	2,727	3,016
64-166PE	2,028	2,110	1,722	9,962	1,499
総処理件数	60,401	72,119	81,461	133,739	152,954

表2．2 NWTジョブのCPU使用時間

	平成5年度	平成6年度	平成7年度	平成8年度	平成9年度
1PE	42,315	69,059	75,461	101,498	98,564
2-7PE	6,832	13,083	34,853	35,423	62,757
8-15PE	14,126	20,846	37,451	133,446	243,035
16-31PE	40,866	76,102	156,076	253,544	384,802
32-63PE	97	10,988	74,130	53,507	124,974
64-166PE	89,157	133,618	69,793	255,417	104,611
総CPU使用時間	193,393	323,695	447,763	832,836	1,018,743
CPU稼働率(%)	33.89	43.97	53.36	79.00	83.50

CPU稼働率=総CPU使用時間/電源投入時間*100

表2．1と表2．2から、年度が進むごとに並列ジョブが増大していることが読みとれる。航技研において、NWTは最初に運用した並列計算機システムではあったが、導入当初から並列ジョブが徐々に増加している。NWTのユーザは不特定多数とはいえ、30人程度のユーザがシステムリソースのおよそ9割を使用するという特殊性をもっている。表中、64台以上の並列ジョブについては、そのジョブ処理量の推移に表れているように、これらのユーザの研究リズムに伴ったシステム利用サイクルに呼応し、NWTジョブの処理量は増減している。NWTの運用ではこのような特殊性を活かし、ビッグユーザをパイロットにして、新システム構築やシステムチューニング等の検討の基礎と



している。なお、平成８年１月のＦＥＰの更新によりＮＷＴの処理性能が２割近く向上した理由にも呼応して、ＮＷＴジョブ処理量の伸び率も飛躍的に増大した。しかし、１ＰＥの処理件数が増大した理由には、並列ジョブの増大とは異なるものがある。すなわち、更新前のＦＥＰで処理されていたバッチジョブ（非並列ジョブ）が、更新後にＮＷＴの１ＰＥジョブに回ったためと考えられる。ちなみに、従来ＦＥＰでは平成７年度においては約３万件のジョブを処理した。このことから、むしろ１ＰＥジョブは減少の傾向にあることが明白であり、本来の並列計算機システムとして並列ジョブがＣＰＵ寄与率を高めつつあるといえる。また、１ＰＥジョブの中には、並列ジョブの結果の検証やパーフォーマンスを１ＰＥジョブで確認するといった並列ジョブ処理に必要なジョブも含まれている。

３．ＮＷＴジョブの入出力データの流れ

ＮＷＴジョブはそのほとんどのものがＦＯＲＴＲＡＮプログラムで記述されている。表３．１はＮＷＴジョブが使用する入出力データの種類を示す。また、ＮＷＴジョブが使用する入出力データの流れの概念図を図３．１に示す。ＮＷＴは上記のファイルを直接的に格納する記憶装置を持たないシステムである。ジョブ実行時に入出力するデータの保存場所はＦＥＰまたはＶＰ２１００の磁気ディスク装置上のファイルである。図３．１に示されるｕｆｓファイルシステムは一般のＵＮＩＸ形式のファイルである。また、ｖｆｌファイルシステムは大容量データの高速入出力を実現するファイル形式であり、磁気ディスクの入出力性能を上げるためにストライピング機能を使用し、１データは 32 分割され複数ボリュームに同時転送される。

表３．１　ＮＷＴジョブの入出力データの種類

処理フェーズ	入出力データの種類	転送方式
翻訳処理	ＦＯＲＴＲＡＮソースプログラム	ＮＦＳ
	コンパイラの印刷出力情報	ＮＦＳ
	オブジェクトモジュール	ＮＦＳ
	リンケージエディタの印刷出力情報	ＮＦＳ
実行処理	ロードモジュール	ＮＦＳ
	実行時標準入力ファイル	ＮＦＳ
	実行時標準印刷出力ファイル	ＳＶＩＯ
	実行時使用入出力ファイル(大規模)	ＳＣＦＳ
	実行時使用入出力ファイル（小規模）	ＮＦＳ

表３．１および図３．１に示されるＮＦＳ(Network File System) はＬＡＮ接続されたＵＮＩＸシステム間においてよく使用されるファイル転送プロトコルである。ＮＷＴジョブ実行時のＮＦＳによるデータ転送では、データはＬＡＮを経由してＦＥＰとＣＰ間を移動する。このため、その入出力の性能はＬＡＮの通信負荷に大きく依存するので、ＮＷＴジョブでは小規模データの転送に限って利用する運用としている。

ＳＣＦＳ(Ssu Cashe File System) はＳＳＵをＦＥＰファイルのキャッシュとして使用し、ＮＷＴジョブからの大量の入出力を高速に処理することを目的に作られた機能である。さらに、ＮＷＴではジョブ実行時に大規模な入力データの読み込み時間を短縮するために、ジョブが実行起動される前に予め入力データをＳＳＵに格納するプレステージングという機能を使用している。プレステージングされた入力データがジョブ実行時にもＳＳＵに存在すれば（ＳＳＵキャッシュがヒットするという)、改めてＦＥＰのファイルからＳＳＵへデータ転送するための入出力動作は発生しないので、データ入力時間が非常に短縮できる。ＮＷＴでは、ＳＣＦＳを利用した大規模順編成ファイル、すなわ

ち、書式なしＦＯＲＴＲＡＮレコードの利用がＮＷＴジョブの実行における入出力処理の中心となる。

標準印刷出力データは直接には磁気ディスクファイルに出力しないで、いったん仮想的な出力イメージでＳＳＵに出力される。このときに使用されるＳＳＵ記憶域をＳＶＩＯ領域という。ジョブ実行終了時にＳＶＩＯ領域の印刷出力データはｒｃｐコマンドでＦＥＰのｕｆｓファイルに格納される。なお、ｕｆｓは一般のＵＮＩＸシステムで利用されるファイル形式である。標準印刷出力としてＳＳＵのＳＶＩＯ領域に出力する方式はＮＦＳに比べて非常に入出力性能が高いので、ジョブ実行時間の短縮を図ることができる。

なお、ＶＰ２１００のＭＳＰシステムより投入されるＮＷＴジョブの入出力データは、ジョブがＮＷＴで実行される前または後の処理フェーズでＦＥＰのファイルにいったん格納される。この格納処理はＭＳＰシステムとＦＥＰの連携機能を実現するＭＶＰＰプログラムが行い、ファイルはＬＡＮを経由して両システム間を移動する。

以上のことから、ＮＷＴジョブの実行時における入出力データ転送の性能はＳＣＦＳおよびＮＦＳデータ転送の処理性能に依存するということが理解できる。このうち、本報告では最も利用の中心となるＳＣＦＳのデータ転送における書式なしＦＯＲＴＲＡＮレコードの入出力性能について述べる。

４．書式なしＦＯＲＴＲＡＮレコードの入出力性能

本章では、先ず実運用におけるＮＷＴジョブの入出力性能がどのような状況であるかを示す。次に、テストジョブにより書式なしＦＯＲＴＲＡＮレコードの入出力性能を実測した結果を示す。

４．１ＮＷＴジョブの入出力性能実績

実運用における過去２ヶ月間の全ＮＷＴジョブについて、ジョブが使用したＰＥの台数ごとにまとめた入出力性能実績を表４．１に示す。表中のデータは各ＮＷＴジョブが使用した全てのＰＥに関するプロセス課金レコードから抽出したＣＰＵ使用時間、入出力データ量ならびにプロセス経過時間に関する情報である。データ転送時間については正確な情報がプロセス課金レコードから抽出できないので、厳密には正確とは言えないが、プロセス経過時間とＣＰＵ使用時間の差をデータ転送時間と定義した。各ジョブの全てのＰＥが入出力したデータ量の合計値をデータ転送時間で割った値を当該ジョブの平均データ転送速度とした。並列ジョブについては最初に並列プログラムをロードし、マスタプロセスとなるプロセスを実行するＰＥ（マスタＰＥという）のＰＥ使用時間とプロセス経過時間を基に同様に算出した。このときの入出力データ量はマスタＰＥおよび他の全ＰＥについての入出力データ量を合計した値とした。

表４．１　ＮＷＴジョブの入出力性能実績

ジョブPE	マスタPE	ファイルサイズ		入出力性能	
使用台数	CPU-TIME	最大値	平均値	最大値	平均値
(台)	(秒)	(MB)		(MB/S)	
1	5047	7012.8	91.3	215.3	5.7
2	1228	349.0	60.1	35.2	7.3
3	7244	325.8	78.8	17.4	3.1
4	143	13209.0	351.6	310.7	47.4
5	6355	197.7	97.7	18.4	5.2
6	6980	328.0	97.7	47.3	3.2
7	8422	102.5	60.9	4.5	2.5
8	4311	1197.9	133.4	100.8	7.5
10	8735	406.9	130.1	48.5	5.4
12	5854	475.8	117.1	12.5	3.1
13	1119	226.6	123.1	49.3	14.1
14	10462	1339.0	84.0	26.2	3.9
16	5356	4375.4	143.1	60.0	8.3
17	11837	421.3	394.2	17.8	13.1
18	4180	226.6	140.0	38.6	10.5
19	383	226.6	147.3	38.9	11.6
20	5593	132.1	62.8	13.3	1.8
24	4839	701.6	659.8	39.1	20.8
28	7774	715.4	319.4	89.2	12.7
32	2997	1561.1	520.9	42.0	13.4
40	7394	2257.8	2032.2	93.1	26.3
64	7140	7752.7	1153.0	86.7	14.9
128	160	951.7	86.7	22.0	2.0
代表性能	4847	13209.0	199.3	310.7	10.0

プロセス課金レコードでは、ＮＷＴジョブが実際に入出力したデータ量については表３．１に示すように入出力の種類ごとに詳細に分類できないので、ＮＷＴジョブの入出力データがどれだけのデータ量をどういった方式で入出力したかは全く不明である。このため、表４．１は実績として参考となるだけで、入出力性能を正当に判断・評価する材料として採用できない。しかし、表中の平均データ転送速度を一覧した場合、ＮＷＴジョブの入出力性能が非常に低いという実感を得る。

以上の理由から、ＮＷＴジョブの入出力性能に関して、特に、書式なしＦＯＲＴＲＡＮレコードの入出力性能について調査するため、データ量や入出力方式を明確にしたテストジョブを実行した。以下にその実測結果を示す。

４．２テストジョブによる実測結果

テストジョブによる書式なしＦＯＲＴＲＡＮレコードの入力性能および出力性能を実測した結果をそれぞれ表４．２（１）と表４．２（２）に示す。この性能実測では、ＰＥを１台使用するジョブ（単一ＰＥジョブ）とＰＥを４台使用する並列ジョブを実行し、サイズ５１２ＭＢのファイルに書式なしＦＯＲＴＲＡＮレコードを入出力した。なお、実行パラメータとしてＦＯＲＴＲＡＮバッファサイズを６００ＫＢと８ＭＢの場合で実測した。

表4.2(1) 書式なしＦＯＲＴＲＡＮレコードの入力性能

テストジョブ	単一PEジョブ		並列ジョブ	
ファイルサイズ	512MB		512MB	
FORTRAN バッファサイズ	600KB	8MB	600KB	8MB
入力性能結果	124.8	240.3	81.0	120.0
入力性能結果	121.3	240.0	81.0	120.0
入力性能結果	122.3	238.3	82.0	119.3
入力性能結果	118.3	235.5	81.5	119.3
入力性能結果	121.0	235.3	80.0	118.3
入力性能結果	121.3	240.3	82.3	120.0
入力性能結果	116.5	229.5	79.5	118.0
入力性能結果	124.8	237.8	81.5	115.7
入力性能結果	121.3	239.5	79.3	118.3
平均値	121.3	237.4	80.9	118.8

入力性能：MB/秒

表4.2(2) 書式なしＦＯＲＴＲＡＮレコードの出力性能

テストジョブ	単一PEジョブ		並列ジョブ	
ファイルサイズ	512MB		512MB	
FORTRAN バッファサイズ	600KB	8MB	600KB	8MB
出力性能結果	13.9	10.5	15.60	9.90
出力性能結果	14.7	10.1	10.30	9.80
出力性能結果	15.8	11.6	14.30	11.70
出力性能結果	15.0	9.7	10.10	10.00
出力性能結果	15.6	9.5	10.60	9.90
出力性能結果	15.7	11.1	14.40	11.60
出力性能結果	14.8	11.5	13.00	11.50
出力性能結果	15.9	9.6	13.70	11.60
出力性能結果	15.9	10.9	11.90	11.90
平均値	15.3	10.5	12.66	10.88

出力性能：MB/秒

なお、ＦＯＲＴＲＡＮバッファとは、ジョブの入出力処理に使用されるデータの一時格納場所であり、ジョブ固有の特定のメモリ領域に確保される。ジョブ実行時にプログラムから発生する入出力命令は即時には、実際の入出力動作を行わず、いったん入出力データをＦＯＲＴＲＡＮバッファ領域に蓄える。バッファがいっぱいになった時点で実際の入出力処理を発生させる。ＮＷＴジョブでは、ＦＯＲＴＲＡＮバッファサイズはユーザが任意に指定できる運用としている。また、システムの標準値は６００ＫＢとしているので、ユーザが陽に指定しない場合には、省略値としてこの値が設定される。

５．考察および改善策とその効果

本章では、先ず、表４．２（１）と表４．２（２）の実測結果から、ＮＷＴジョブにおける書式なしＦＯＲＴＲＡＮレコードの入出力性能を評価し、性能劣化の要因とその改善策を示す。次に、改善策を講じた入出力方式によるテストジョブの実測結果からその有効性を検証する。

５．１　書式なしＦＯＲＴＲＡＮレコードの入出力性能評価

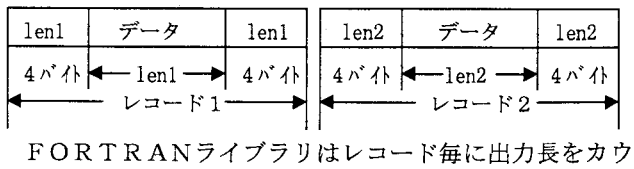
入力性能については表４．２（１）から、以下のとおり評価できる。第１に、入力はＳＳＵのキャッシュ機能が有効に働き、どの実測結果の場合にも一様に高い性能が得られている。第２に、ＦＯＲＴＲＡＮバッファサイズは８ＭＢの場合の方が６００ＫＢよりも高い性能が得られている。このことから、バッファはキャッシュの管理単位である８ＭＢの倍数が論理的によい理由が実測と符合している。第３に並列ジョブについてはどの実測結果も単一ＰＥジョブのそれより一様に性能が低い。これは、ＰＥ間データ転送等の並列ジョブのオーバヘッドによる遅延があるためである。

表４．２（２）から、出力性能については以下のとおり評価できる。第１に、出力はＳＳＵのキャッシュ機能が働かず、どの実測結果の場合にも入力に比較すると一様に性能が低い。また、ＦＯＲＴＲＡＮバッファサイズを大きくしても有効でないことが結論できる。

５．２　書式なしＦＯＲＴＲＡＮレコードの出力性能劣化の要因と改善策

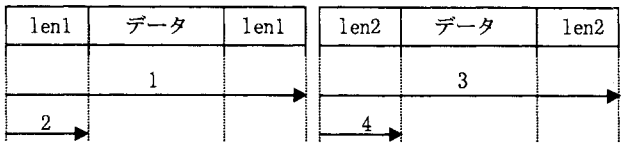
テストジョブの実測結果から、書式なしＦＯＲＴＲＡＮレコードの出力について性能を劣化させる以下のような重大問題があることを確認した。

順編成ファイルのＦＯＲＴＲＡＮレコード形式は、以下のように実際のデータの前後にレコード長(len1, len2)が付加された形となる。



ＦＯＲＴＲＡＮライブラリはレコード毎に出力長をカウ

ントし、レコードを最後まで出力した後、先頭の長さを出力する。この先頭のレコード長の出力処理が磁気ディスクへの書き戻し処理を待つため、実磁気ディスクの入出力性能になってしまう。



5．3入出力性能改善策とその検証

改善策は、出力性能低下を招くレコード長(len1, len2)の書き戻し処理を極力減らすために、出力時に実データを出力する前に当該レコードの直前に出力されたレコードのレコード長を予測値として予め書き込んでおく方式である。改善策を講じた出力処理を行う場合は、レコード長の書き戻しはこの予測値が外れた場合にのみ行えばよいことになる。

現状のシステム方式と改善策を講じた入出力方式による書式なしＦＯＲＴＲＡＮレコードの出力性能を実測した。その実測結果は表４．３に示すとおりである。同表に示すとおり、単一ＰＥジョブのテストジョブを実行し、サイズ１２０ＭＢのファイルに書式なしＦＯＲＴＲＡＮレコードの出力処理を行った。なお、実行パラメータとしてＦＯＲＴＲＡＮバッファサイズを６００ＫＢと８ＭＢの場合で実測した。

表4.3 書式なしＦＯＲＴＲＡＮレコードの出力性能

テストジョブ	単一PEジョブ			
ファイルサイズ	120MB			
FORTRAN バッファサイズ	600KB		8MB	
入出力方式	現状	改善策	現状	改善策
出力性能結果	14.7	61.0	15.4	89.1
出力性能結果	12.6	60.2	16.1	82.7
出力性能結果	15.0	52.9	14.7	78.9
出力性能結果	12.8	60.8	15.4	84.2
出力性能結果	14.8	57.8	14.8	75.3
出力性能結果	12.8	58.5	14.2	84.3
出力性能結果	14.9	58.8	15.3	80.7
出力性能結果	13.7	60.4	15.5	72.0
平均値	13.9	58.8	15.2	80.9

出力性能：MB/秒

表４．３において、列（現状）は現状の入出力方式による実測結果である。列（改善策）は改善策を講じた入出力方式を使用してテストした実測結果である。列（改善策）は、いずれも現状方式より高い性能が得られ、改善策が有効であることが検証できる。

6．おわりに

ＮＷＴジョブの入出力性能を正確に調査する上では、先ずジョブから発生する各種の入出力に対するデータ量、入出力回数、入出力処理時間ならびに入出力の方式等、詳細な入出力処理情報が必要となる。しかし、現在のＵＮＩＸシステムでは、第４．１項に述べたとおり、システム機能やツールからは必要とする十分な情報が得られない。ＮＷＴのような並列計算機システムで、かつＵＮＩＸシステムにおいては不特定多数のユーザにサービスするセンタマシーンとしての歴史はまだ浅く、システムを運用管理する上で有用なツール、ソフトウェアモニタ等はまだ十分整備されていない。このようなシステム環境のもとでＮＷＴジョブの入出力処理性能を調査し、ジョブの実行時に最も利用の中心となる大規模順編成ファイルの出力処理性能が非常に低いことを確認した。また、出力性能改善策を提示し、その有効性を検証し得た。この改善策を実運用に供し、その有効性を調査する予定である。本テーマから、今後の計算機システム構築の検討に際しては、ソフトウェアおよびハードウェアの両観点から高効率な入出力方式の設計が非常に重要であると考える。

おわりに当たり、計算機システムのハードウェアおよびソフトウェアに関して多くの情報提供と貴重な討論を頂いた富士通（株）の松岡玄一氏、岡田信氏、坂本喜則氏に対して感謝の意を表する。