京チェックスイートによる HOKUSAI-GreatWave/FX100の性能確認

理研シンポジウム

2015年6月19日(金)

国立研究開発法人 理化学研究所
計算科学研究機構 運用技術部門
ソフトウェア技術チーム
北澤 好人
yoshito.kitazawa@riken.jp

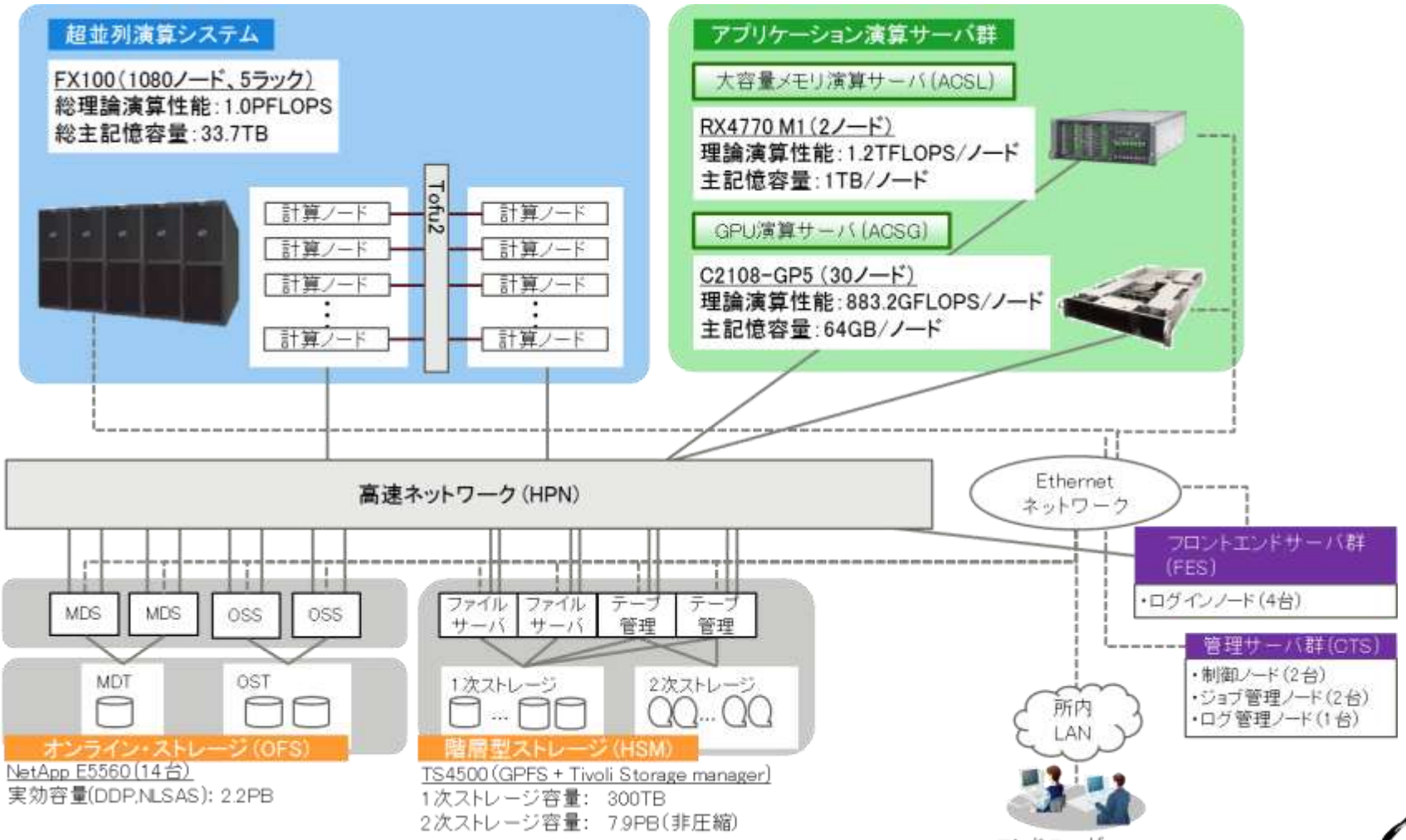
K computer

概要

- ベンチマークの目的
  - 新たに導入されたシステムFX100にて、京コンピュータのシステム更新時に性能健全性を確認するために利用している「チェックスイート」を実行して、演算性能と通信性能を確認した結果と、性能を引き出すためのポイント（実行例・注意点）を紹介する。

システムの概要

- FX100の構成



*「HOKUSAI-GreatWaveシステム 利用手引書 Version 1.3」より

システムの概要

- ハードウェア諸元

	京コンピュータ	HOKUSAI/FX100
CPU	SPARC64™VIIIfx	SPARC64™XIfx
動作周波数	2.0GHz	1.975GHz
コア数	8コア/ノード	16コア/CMG, 2CMG/ノード
キャッシュ容量	L1I$:32KB/コア(2WAY) L1D$:32KB/コア(2WAY) L2$:6MB/ノード	L1I$:64KB/コア(4WAY) L1D$:64KB/コア(4WAY) L2$:12MB/CMG
SIMD幅	128bit(倍精度SIMDx2)	256bit(倍精度SIMDx4)
理論演算性能(倍精度)	128GFLOPS/ノード	505.6GFLOPS/CMG 1,011.2GFLOPS/ノード
メモリ容量	16GB/ノード	16GB/CMG, 32GB/ノード
メモリバンド幅	64GB/s	240GB/s/CMG
ノード間通信性能	5.0GB/s×4方向同時	12.5GB/s×4方向同時
ノード数	82,944ノード	1,080ノード

*京コンピュータ:「チューニングチュートリアル V1.6.4」より
*FX100:「White paper FUJITSU Supercomputer PRIMEHPC FX100 次世代技術への進化」
「HOKUSAI-GreatWaveシステム 利用手引書 Version 1.0」より

チェックスイートとは？

- チェックする内容について

京コンピュータの言語版数アップやシステム変更を行う時に、性能低下や問題が発生していないか、チェックするツール「チェックスイート」を用意

- 性能の妥当性　　：　理論性能または基準となる性能と比較する
- ロードバランス　　：　ランク間の経過時間のバラツキを比較する
- 性能ブレ　　：　システム起因と想定される経過時間のブレを確認する
- 計算結果の妥当性：　計算が正しく行われているか確認する

0) 環境情報
1) 演算性能
2) 通信性能
3) ロードバランス
4) 性能ブレ
5) ステージング性能
6) 計算結果妥当性

チェックスイートとは？

- プログラムについて

実行するプログラムは数値計算の実アプリケーションを利用

プログラム名	アプリケーション概要	並列数	問題規模
FrontFlow/blue	Large Eddy Simulation(LES)に基づく非定常流体解析プログラム	1，024	（weak　scale）
NICAM	全球雲解像大気大循環モデル	20，480	G13データ
Lattice QCD	格子QCDシミュレーションプログラム	1，296	48x24x24x48
PHASE	擬ポテンシャルと密度汎関数法によるナノ材料第一原理分子動力学プログラム	3，072	screw__48x24
RSDFT	実空間第一原理動力学プログラム	9，216	SiNWd20x4__18Ry
Seism3D	地震波伝搬・強震動シミュレーションプログラム	24，576	128x192

K computer

チェックスイートとは？

- **利用状況について**

チェックスイートは、京コンピュータの言語版数アップやシステム変更を行う時に実際に実行して、性能低下や問題が発生していないかチェックを実施している。
実際にチェックした例を以下に示す。

- 1.2.0-15版　：問題なく、公開・デフォルト化へ
- 1.2.0-16-2, -3版　：RSDFTの通信に一部問題がありデフォルト化を見送り、言語環境としては制限付きで公開へ
- 1.2.0-17版　：PHASEの通信の異常終了を検知して公開せず
  →1.2.0-17-2版で正常終了を確認して公開へ

京の安定稼働に貢献

ベンチマークの内容

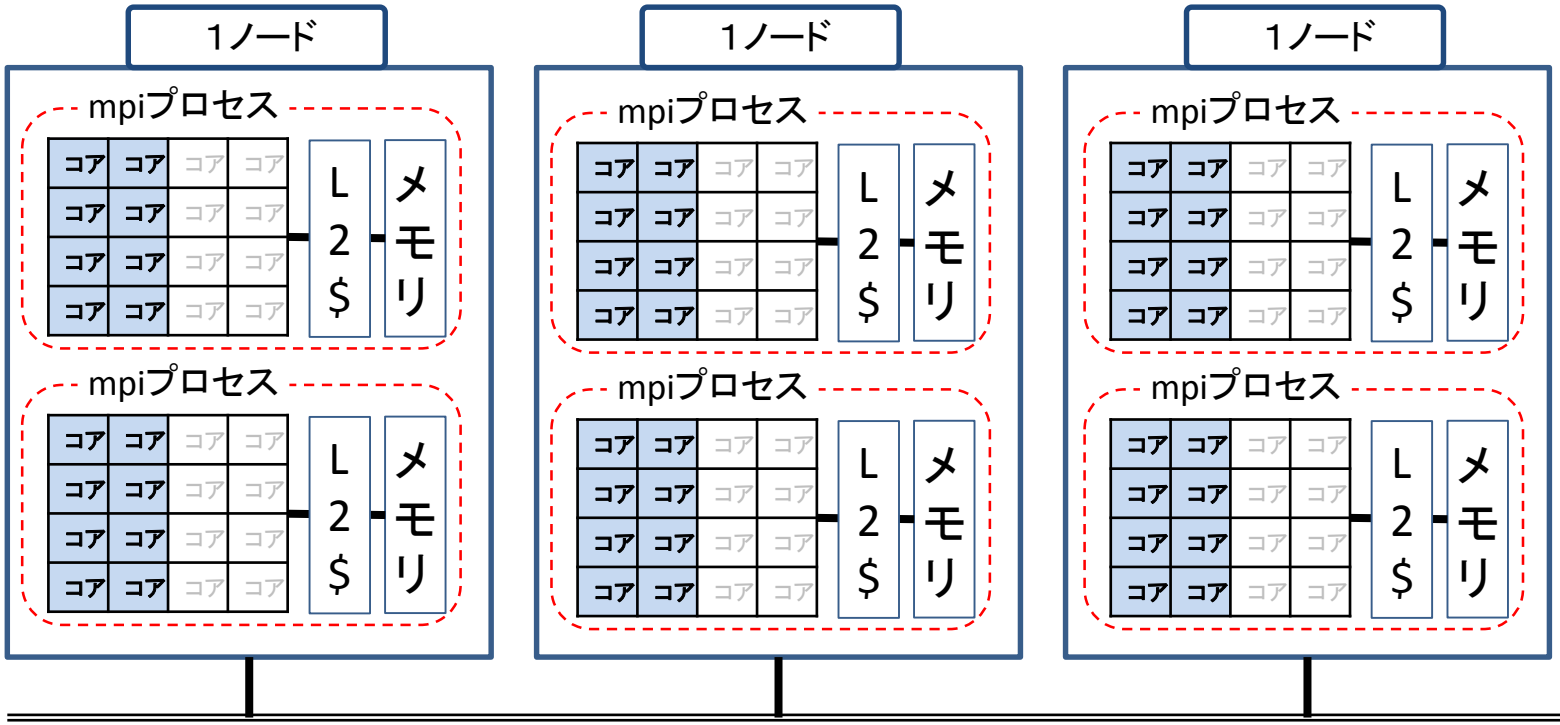
- **今回の設定**

  並列数の制限があるのでFX100では規模を縮小して実行

プログラム名	アプリケーション概要	並列数（ノード数）	問題規模
FrontFlow/blue	Large Eddy Simulation(LES)に基づく非定常流体解析プログラム	1，024（512）	（weak　scale）
NICAM	全球雲解像大気大循環モデル	1，280（640）	g11データ
Lattice QCD	格子QCDシミュレーションプログラム	1，296（648）	48x24x24x48
PHASE	擬ポテンシャルと密度汎関数法によるナノ材料第一原理分子動力学プログラム	2，160（1，080）	screw__48x24
RSDFT	実空間第一原理動力学プログラム	1，728（864）	d10x3__18Ry
Seism3D	地震波伝搬・強震動シミュレーションプログラム	1，728（864）	48x36

ベンチマークの内容

- 並列数の構成について
  - 京の8スレッド実行の結果と比較するため、FX100も8スレッド並列により実行
  - 1ノードに2CMG（Core Memory Group）があるので、1CMGあたり1mpiプロセスにて実行（1ノードに2mpiプロセスを割り当てる）



K computer

実行結果

- チェックスイートの各アプリの実行結果(全体)

FFB7	id	name	K	FX100
(全体)	1000	Loop	435.06	323.896
		(peak比)	(3.81%)	(2.59%)
	1100	Loop-blk	435.059	323.895
		(peak比)	(3.81%)	(2.59%)
(通信)	1112	ddcom2 (集団通信:Allreduce)	0.814	3.793
	1115	ddcomx (一対一通信:isend/irecv)	51.246	26.732
NICAM	id	name	K	FX100
(全体)	1000	dynstep	46.473	25.295
		(peak比)	(6.76%)	(6.29%)
(通信)	200	COMM_data_transfer (隣接通信)	3.222	3.798
PHASE	id	name	K	FX100
(他ライブラリ)	101	FFT_DIRECT	151.905	378.341
		(peak比)	(1.54%)	(0.31%)
	102	FFT_INVERSE	356.143	772.855
		(peak比)	(2.45%)	(0.57%)
(演算)	103	GS/W1SW2/dgemm1	2.435	3.298
		(peak比)	(27.99%)	(10.46%)
	104	GS/W1SW2/dgemm2	31.186	40.602
		(peak比)	(61.12%)	(23.77%)
	105	GS/MODBPPSI/dgemm1	1.916	1.372
		(peak比)	(40.22%)	(28.45%)
	106	GS/MODBPPSI/dgemm2	31.412	28.248
		(peak比)	(60.85%)	(34.26%)
(他ライブラリ)	107	EIGEN	15.179	12.102
		(peak比)	(0.13%)	(0.09%)
(通信)	202	FFT_D/alltoall1	16.709	**153.887**
	205	FFT_D/alltoall2	11.188	11.083
	208	FFT_I/alltoall1	40.139	38.792
	211	FFT_I/alltoall2	60.203	**575.356**
	222	GS/WSW/allreduce	0.396	3.234
	225	GS/W1SW2/allreduce	1.835	3.543
	228	GS/bcast1	2.744	1.674
	231	GS/bcast2	0.701	0.443
	234	GS/allreduce	1.474	5.375

QCD	id	name	K	FX100
(演算)	2101	kernel1	0.041	0.036
		(peak比)	(26.23%)	(15.11%)
	2102	kernel2	0.039	0.092
		(peak比)	(28.00%)	(5.92%)
	2103	kernel3	1.275	1.217
		(peak比)	(35.69%)	(18.94%)
	2108	drbicgstab_dd	10.093	9.422
		(peak比)	(16.93%)	(9.18%)
(通信)	2105	send	0.554	0.612
	2106	recv	2.551	2.510
RSDFT	id	name	K	FX100
(全体)	1000	Gram_Schmidt	61.481	73.852
		(peak比)	(44.80%)	(18.89%)
	1100	GS/MM	35.892	32.533
		(peak比)	(77.08%)	(43.06%)
(演算)	1101	GS/MM/dgemm1	17.771	14.549
		(peak比)	(77.28%)	(47.80%)
	1102	GS/MM/dgemm2	16.688	12.723
		(peak比)	(82.30%)	(54.65%)
(通信)	1104	GS/MM/allreduce1	0.819	2.280
	1402	GS/Allgatherv	1.404	1.849
	1502	GS/Bcast	1.733	3.582
Seism3D	id	name	K	FX100
(演算)	1101	kernel_stressderiv	3.963	2.390
		(peak比)	(17.03%)	(14.30%)
	1102	kernel_velderiv	3.513	2.003
		(peak比)	(19.22%)	(17.06%)
	1103	kernel_update_vel	2.996	1.199
		(peak比)	(21.85%)	(27.65%)
	1104	absorb_update_vel	2.819	1.116
		(peak比)	(0.68%)	(0.87%)
	1111	kernel_update_stress	10.742	8.007
		(peak比)	(20.18%)	(13.71%)
	1112	absorb_update_stress	4.051	2.374
		(peak比)	(0.39%)	(0.34%)
(通信)	150	mpproc_passing_velocity	1.056	1.010
	160	mpproc_passing_stress	1.061	0.996

結果の分析

- 演算性能(RSDFT／DGEMM実行部分)
  - FX100を8スレッド実行して京と比較
    (FX100理論演算性能：252.8GFLOPS/1CMG/8core)
  - 実行時間は京 16.6秒→FX100 12.7秒へ短縮(区間1102)
  - 浮動小数点演算性能ピーク比は京 82.3%→FX100 54.6%へ低下
    DGEMMが16スレッド用にチューニングされているため効率が低下

RSDFT	id	name	K	FX100
(演算)	1101	GS/MM/dgemm1	17.771	14.549
		(peak比)	(77.28%)	(47.80%)
	1102	GS/MM/dgemm2	16.688	12.723
		(peak比)	(82.30%)	(54.65%)

結果の分析

- 演算性能（DGEMM／16スレッドの単体性能）
  - FX100にて1ノード・1CMG単体の8スレッド／16スレッドの浮動小数点演算性能を精密PAにより確認した
  - 8スレッド：ピーク比 57．3%　16スレッド：ピーク比89．4%
    → 16スレッド用にチューニングされている

△8スレッド実行

	実行時間（sec）	浮動小数点演算ピーク比	MFLOPS	MIPS	浮動小数点演算数
Process	12.70	57.33%	144941	28545	1.84E+12

L1ビジー率	L2ビジー率	メモリビジー率	L2スループット（GB/sec）	メモリスループット（GB/sec）
49%	37%	28%	166.66	39.71

◎16スレッド実行

	実行時間（sec）	浮動小数点演算ピーク比	MFLOPS	MIPS	浮動小数点演算数
Process	4.06	89.45%	452284	87011	1.84E+12

L1ビジー率	L2ビジー率	メモリビジー率	L2スループット（GB/sec）	メモリスループット（GB/sec）
64%	81%	11%	465.64	15.19

○ 32スレッド(2CMG利用)実行（参考）

	実行時間（sec）	浮動小数点演算ピーク比	MFLOPS	MIPS	浮動小数点演算数
Process	2.23	81.21%	821220	158003	1.84E+12

L1ビジー率	L2ビジー率	メモリビジー率	L2スループット（GB/sec）	メモリスループット（GB/sec）
59%	74%	12%	845.49	31.52

結果の分析

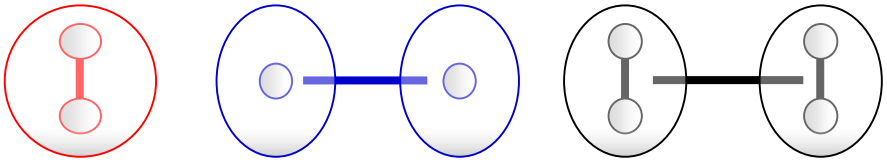
- 通信性能(RSDFT/集団通信部分)
  - 一部のノードで通信縮退が発生→性能1/2
    (FX100理論性能: 12.5GB/s×4方向／2CMG／2＝12.5GB/s
    京実効性能値 ： 5.0GB/s×4方向×効率75% ＝15.0GB/s)
  - 3次元形状(12x12x6:torus)を指定して実行
    1CMGあたり1mpiプロセス(1ノードあたり2mpiプロセス)にて実行
  - 理論性能が同程度のはずだが通信時間が2倍
    1CMGあたり1mpiプロセス(ノード内2プロセス)で実行すると、
    ノード内通信の時間がかかる

RSDFT	id	name	K	FX100
(通信)	1104	GS/MM/allreduce1	0.819	2.280
	1402	GS/Allgatherv	1.404	1.849
	1502	GS/Bcast	1.733	3.582

結果の分析

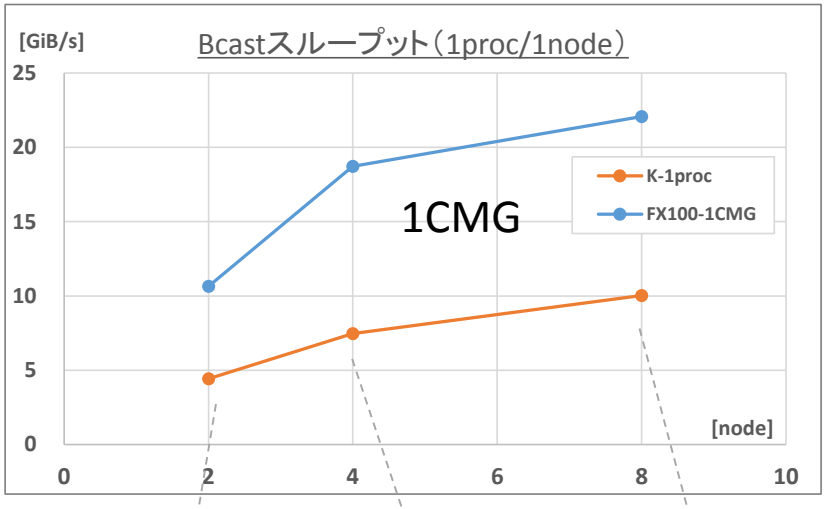
- 通信性能(MPI_Bcast/ノード内の通信性能)
  - MPI_Bcast通信により256MBのデータを転送する性能を確認した
  - ノード内通信: 京 5.0GB/s, FX100 8.1GB/s → 性能が低い
  - ノード間通信: 理論性能×90%の性能 → 妥当な性能
  - ノード内・ノード間混合: 低い方の性能に依存する
    → FX100ではノード内通信の性能の影響が大きい

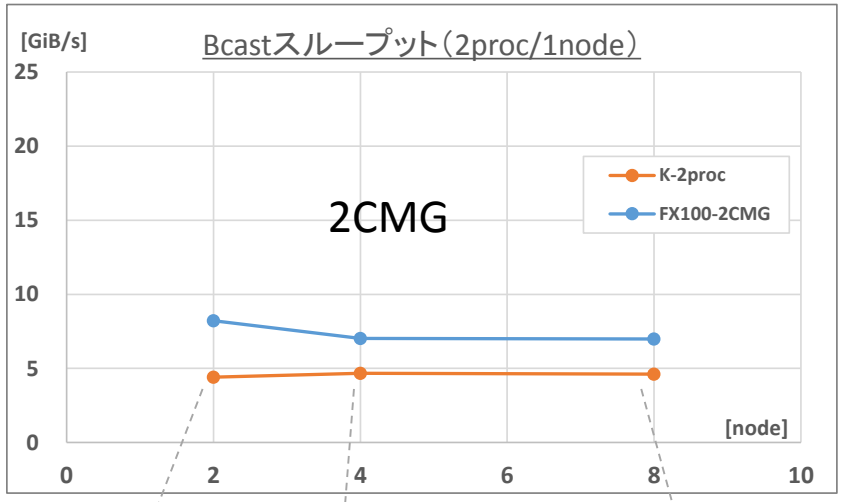
通信方向	ノード数	並列数	CMG数	スループット 京[GiB/s] ﾉｰﾄﾞ間理論性能:5.0GB/s	スループット FX100[GiB/s] ﾉｰﾄﾞ間理論性能:12.5GB/s
ノード内	node=1	proc=2	2	5.037	8.105
ノード間	node=2	proc=2	1	4.430	10.653
混合	node=2	proc=4	2	4.420	8.224

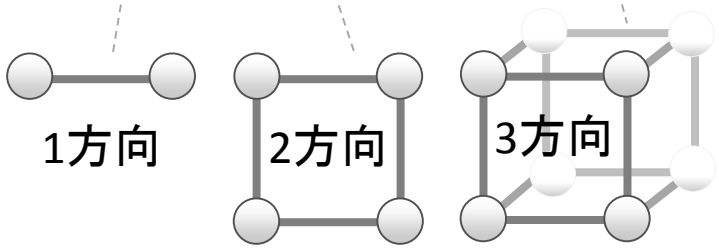


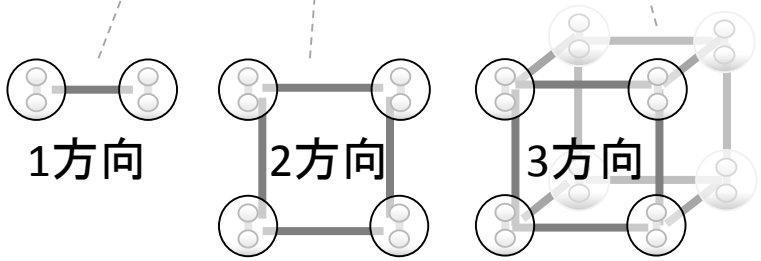
結果の分析

- 通信性能(MPI_Bcast/ノード間の通信性能)
  - ノード内1mpiプロセス: 京 10.0GB/s, FX100 22.0GB/s
    (同時3方向通信でバンド幅拡大)
  - ノード内2mpiプロセス: 京 4.6GB/s, FX100 最大8.2GB/s
    →ノード内通信に性能が引きずられる









K computer

結果の分析

- 演算性能(Seism3D/演算カーネル部分)
  - FX100を8スレッド実行して京と比較
    (FX100理論演算性能: 252.8GFLOPS/1CMG/8core)
  - 実行時間は京→FX100で全て短縮
  - 浮動小数点演算性能ピーク比は処理によって向上・低下
    (京/SIMD:2要素、FX100/SIMD:単精度4or8要素→SIMD化状況で変化)

Seism3D	id	name	K	FX100
(演算)	1101	kernel_stressderiv	3.963	2.390
		(peak比)	(17.03%)	(14.30%)
	1102	kernel_velderiv	3.513	2.003
		(peak比)	(19.22%)	(17.06%)
	1103	kernel_update_vel	2.996	1.199
		(peak比)	(21.85%)	(27.65%)
	1104	absorb_update_vel	2.819	1.116
		(peak比)	(0.68%)	(0.87%)
	1111	kernel_update_stress	10.742	8.007
		(peak比)	(20.18%)	(13.71%)
	1112	absorb_update_stress	4.051	2.374
		(peak比)	(0.39%)	(0.34%)

結果の分析

- 通信性能(Seism3D/隣接通信部分)
  - 一部のノードで通信縮退が発生→性能1/2
    (FX100理論性能: 12.5GB/s×4方向／2CMG／2＝12.5GB/s
    京実効性能値 ：5.0GB/s×4方向×効率75% ＝15.0GB/s)
  - 2次元形状(24x36:torus)を指定して実行
    1CMGあたり1mpiプロセス(1ノードあたり2mpiプロセス)にて実行
  - 隣接2方向×送受信(=4方向)の同時通信
  - 京とFX100で同様の実行時間 → 理論性能の傾向と同様

Seism3D	id	name	K	FX100
(通信)	150	mpproc_passing_velocity	1.056	1.010
	160	mpproc_passing_stress	1.061	0.996

結果の分析

- 演算性能(FFB/ループ全体)
  - FX100を8スレッド実行して京と比較
    (FX100理論演算性能: 252.8GFLOPS/1CMG/8core)
  - 実行時間は京 435.0秒→FX100 323.8秒で短縮
  - 浮動小数点演算性能ピーク比は3.8%→2.6%へ変化
    (京/SIMD:2要素、FX100/SIMD:単精度4or8要素→SIMD化状況で変化)

FFB7	id	name	K	FX100
(全体)	1000	Loop	435.06	323.896
		(peak比)	(3.81%)	(2.59%)

結果の分析

- 通信性能(FFB/集団通信・一対一通信)
  - 一部のノードで通信縮退が発生→性能1/2
    (FX100理論性能: 12.5GB/s×4方向／2CMG／2＝12.5GB/s
    京実効性能値 ：5.0GB/s×4方向×効率75% ＝15.0GB/s)
  - 1CMGあたり1mpiプロセス(1ノードあたり2mpiプロセス)にて実行
  - 一対一通信(ddcomx)区間はデータのPACK/UNPACK処理がコストの90%のため実質的に演算処理が高速化
  - 集団通信Allreduceはノード内2mpiプロセスの実行のため性能低下
    更に、1要素のためTofu高機能バリア通信が適用されるが
    測定時にバリア通信の問題があり性能低下した可能性がある

FFB7	id	name	K	FX100
(通信)	1112	ddcom2 (集団通信:Allreduce)	0.814	3.793
	1115	ddcomx (一対一通信:isend/irecv)	51.246	26.732

結果の分析

- 通信性能(MPI_Allreduce 1要素によるTofu高機能バリア通信)
  - MPI_Allreduce 1要素(4byte)の性能を確認した
  - チェックスイート実行時(2015/4)では 13.1マイクロ秒だが、
    その後(2015/6)に確認すると 8.7マイクロ秒と下がり、
    測定時にTofu高機能バリア通信に問題があったと考えられる
  - MPI_Barrierの時間も 14.5マイクロ秒から 4.0マイクロ秒へ改善した

→2015/4ではバリア通信に問題があり、その後2015/6には改善

通信種類	ノード数	並列数	CMG数	通信時間 京[μsec]	通信時間 FX100[μsec] (2015/4)	通信時間 FX100[μsec] (2015/6)
MPI_Allreduce	node=8	proc=8	1	3.855	10.118	7.875
MPI_Allreduce	**node=8**	**proc=16**	**2**	8.886	13.145	**8.697**
MPI_Barrier	node=8	proc=8	1	2.202	10.078	3.392
MPI_Barrier	**node=8**	**proc=16**	**2**	2.908	14.486	**3.989**

結果の分析

- 演算性能(NICAM/ループ全体)
  - FX100を8スレッド実行して京と比較
    (FX100理論演算性能: 252.8GFLOPS/1CMG/8core)
  - 実行時間は京 46.4秒→FX100 25.2秒で短縮
  - ループ部分の浮動小数点演算性能ピーク比は同程度

NICAM	id	name	K	FX100
(全体)	1000	dynstep	46.473	25.295
		(peak比)	(6.76%)	(6.29%)

結果の分析

- 通信性能(NICAM/隣接通信)
  - 一部のノードで通信縮退が発生→性能1/2
    (FX100理論性能: 12.5GB/s×4方向／2CMG／2＝12.5GB/s
    京実効性能値 ：5.0GB/s×4方向×効率75% ＝15.0GB/s)
  - 1CMGあたり1mpiプロセス(1ノードあたり2mpiプロセス)にて実行
  - 隣接通信(COMM_data_transfer)区間は理論性能の傾向と同様

NICAM	id	name	K	FX100
(通信)	200	COMM_data_transfer (隣接通信)	3.222	3.798

結果の分析

- 演算性能(PHASE/DGEMM実行部分)
  - FX100を8スレッド実行して京と比較
    (FX100理論演算性能: 252.8GFLOPS/1CMG/8core)
  - 実行時間は京 31.4秒→FX100 28.2秒へ短縮(区間106)
  - 浮動小数点演算性能ピーク比は京 60.8%→FX100 34.2%へ低下
    DGEMMが16コア用にチューニングされているため効率が低下
    (RSDFTと同様)

PHASE	id	name	K	FX100
(演算)	103	GS/W1SW2/dgemm1	2.435	3.298
		(peak比)	(27.99%)	(10.46%)
	104	GS/W1SW2/dgemm2	31.186	40.602
		(peak比)	(61.12%)	(23.77%)
	105	GS/MODBPPSI/dgemm1	1.916	1.372
		(peak比)	(40.22%)	(28.45%)
	106	GS/MODBPPSI/dgemm2	31.412	28.248
		(peak比)	(60.85%)	(34.26%)

結果の分析

- 通信性能(PHASE/集団通信部分)
  - 一部のノードで通信縮退が発生→性能1/2
    (FX100理論性能: 12.5GB/s×4方向／2CMG／2＝12.5GB/s
    京実効性能値 ：5.0GB/s×4方向×効率75% ＝15.0GB/s)
  - 3次元形状(12x15x6:torus)をランクマップファイルにより実行
    1CMGあたり1mpiプロセス、1ノードあたり2mpiプロセスにて実行
  - Allreduce通信はノード内2mpiプロセスの実行のため性能低下
  - Bcast通信はランクマップファイルにより通信方向が連続化して実行時間が短縮
  - Alltoall通信は特定のコミュニケータにて通信性能が異常なノードがあり、
    異常な結果となる。正常範囲で確認すると他の区間と同様の値

PHASE	id	name	K	FX100	正常ノードのmax
(通信)	202	FFT_D/alltoall1	16.709	153.887	→15.512
	205	FFT_D/alltoall2	11.188	11.083	
	208	FFT_I/alltoall1	40.139	38.792	
	211	FFT_I/alltoall2	60.203	575.356	→57.433
	222	GS/WSW/allreduce	0.396	3.234	
	225	GS/W1SW2/allreduce	1.835	3.543	
	228	GS/bcast1	2.744	1.674	
	231	GS/bcast2	0.701	0.443	
	234	GS/allreduce	1.474	5.375	

結果の分析

- 演算性能(QCD/カーネル部分)
  - FX100を8スレッド実行して京と比較
    (FX100理論演算性能: 252.8GFLOPS/1CMG/8core)
  - 実行時間はkernel2区間の除いて短縮の傾向
  - 浮動小数点演算性能ピーク比は京と比較していずれも低下

QCD	id	name	K	FX100
(演算)	2101	kernel1	0.041	0.036
		(peak比)	(26.23%)	(15.11%)
	2102	kernel2	0.039	0.092
		(peak比)	(28.00%)	(5.92%)
	2103	kernel3	1.275	1.217
		(peak比)	(35.69%)	(18.94%)
	2108	drbicgstab_dd	10.093	9.422
		(peak比)	(16.93%)	(9.18%)

結果の分析

- 通信性能(QCD/隣接部分)
  - 一部のノードで通信縮退が発生→性能1/2
    (FX100理論性能: 12.5GB/s×4方向／2CMG／2＝12.5GB/s
    京実効性能値 ：5.0GB/s×4方向×効率75% ＝15.0GB/s)
  - 1CMGあたり1mpiプロセス、1ノードあたり2mpiプロセスにて実行
  - 隣接通信(send, recv)区間は理論性能の傾向と同様

QCD	id	name	K	FX100
(通信)	2105	send	0.554	0.612
	2106	recv	2.551	2.510

性能を引き出すためのポイント(実行例・注意点)

- 演算
  - 16コア/CMG, 32コア/ノードをフルに使う
  - 数値計算ライブラリの高性能となるケースの設定(DGEMM:16スレッド)
  - コンパイラによるSIMD化の適用状況の確認(京:SIMDx2→FX100:SIMDx4)

- 通信
  - 1ノードあたり1mpiプロセスによる実行が効果的
  - 集団通信のTofu専用アルゴリズムの利用
  - 通信縮退(通信バンド幅)・バリア通信の状態に注意