

「情報システム学特別講義3」 エクサフロップスコンピューティング に向けて

東京大学情報基盤センター准教授 片桐孝洋

2014年8月5日(火) 14:40－16:10

情報システム学特別講義3

講義日程

(情報システム学特別講義 3)

レポートおよびコンテスト課題

(締切:

2014年8月11日(月)24時 厳守

~~1. 4月8日: ガイダンス~~

~~2. 4月15日~~

~~▶ プログラム高速化の基礎(その1)~~

~~3. 4月22日~~

~~▶ プログラム高速化の基礎(その2)~~

~~4. 5月13日~~

~~▶ MPIの基礎~~

~~5. 5月20日~~

~~▶ OpenMPの基礎~~

~~6. 5月27日~~

~~▶ Hybrid並列化技法
(MPIとOpenMPの応用編)~~

~~7. 6月3日~~

~~▶ プログラム高速化の応用~~

~~8. 6月10日~~

~~▶ 行列・ベクトル積の並列化~~

~~9. 6月17日~~

~~● ベキ乗法の並列化~~

~~10. 6月24日~~

~~● 行列・行列積の並列化~~

~~11. 7月8日~~

~~● LU分解の並列化~~

~~12. 7月15日~~

~~● 非同期通信~~

~~● 疎行列反復解法の並列化~~

~~13. 7月22日~~

~~● ソフトウェア自動チューニング~~

14. 8月5日(補講日)

● エクサフロップスコンピューティング
に向けて

スパコンの技術動向

エクサフロップスコンピューティング に向けて

- ▶ 2018年度ごろに実現するハードウェア性能をもとに、
エクサフロップス(Exa-flops, 1000PFLOPS)のスパコンを制作可能か、必要となる技術について検討する
- ▶ **HPCI技術ロードマップ白書、2012年3月**
 - ▶ <http://www.open-supercomputer.org/workshop/sdhpc/>
において、ハードウェア、システムソフトウェア、
プログラミング環境、数値計算ライブラリの観点
から検討
- ▶ 本講義では、その中身について紹介

2018年までのハードウェア予想緒元

▶ HPCI技術ロードマップ白書、2012年3月、pp.11から転載

表 2-1 プロセス世代別の性能予測

	単位	40nm-45nm	28nm-32nm	20nm-22nm	15nm-16nm	10nm-11nm	7nm-8nm
利用可能時期	西暦年	2009	2011	2013	2015	2017	2019
倍精度演算器サイズ	um ²	160000	90000	51000	29000	17000	13000
倍精度演算器消費電力	pJ/FLOP	20	10.5	6.0	3.5	2.0	1.1
SRAM サイズ	MB/mm ²	0.24	0.42	0.74	1.29	2.25	3.92
SRAM 電力	pJ/access	6.30	3.60	2.06	1.18	0.67	0.35
レジスタファイル電力	pJ/access	12.00	7.00	4.00	2.29	1.31	0.75
On-chip バス(64bit)電力	pJ/mm	7.00	4.00	2.29	1.31	0.75	0.42
TSV 転送電力(64bit)	pJ	-	0.64	0.37	0.21	0.12	0.08
Off-chip シリアル転送	pJ/bit	12.00	6.60	3.77	2.16	1.23	0.70
Off-chip バス	pJ/bit	50.00	30.00	17.14	9.80	5.60	3.20
シリアル転送チャンネルサイズ	mm ² /ch	0.60	0.60	0.60	0.60	0.60	0.60
On-chip バスサイズ	mm ² /ch	4.00	4.00	4.00	4.00	4.00	4.00
光ケーブル転送レイテンシ	ns/m	5.00	5.00	5.00	5.00	5.00	5.00

3.2x64bit
=204.8
pJ/64bit

(1)データ転送の消費エネルギー: on-chipバスと、off-chipバスの転送電力の差は100倍。

⇒如何にoff-chipのデータ伝送電力を減らすかがカギ

(2)データ転送のレイテンシ: 光ケーブル転送レイテンシは、2019年まで削減されない

⇒通信は、ほとんどレイテンシ律速になる

2018年までのハードウェア予想緒元

▶ レイテンシ・プロセッサ

- ▶ 高いシングルスレッド性能を指向
- ▶ レイテンシ・プロセッサ上のコアをレイテンシ・コア
- ▶ 例) PC/ワークステーション向けプロセッサ
 - ▶ Intel Xeon、IBM Power、ARM Cortex-A シリーズ

▶ スループット・プロセッサ

- ▶ 高いスループット性能を指向
- ▶ スループットプロセッサ上のコアをスループット・コア
- ▶ 例) GPU やClear Speed など、アクセラレータ

レイテンシプロセッサの性能予測

▶ HPCI技術ロードマップ白書、2012年3月、pp.12から転載

表 2-2 レイテンシプロセッサの性能予測

		2012 年	2014 年	2016 年	2018 年
設計プロセス	nm	32	22	16	11
レイテンシコア数	個/chip	4～8	8～16	16～32	32～64
クロック周波数	GHz	3.0～4.0	3.0～4.0	3.0～4.0	3.0～4.0
演算器数	FMA/Core	4	4	8	8
演算レイテンシ	ns	1.0	1.0	1.0	1.0
ハードウェアスレッド数	Thread/Core	4	4	4	4
スレッドあたりレジスタ数	本/スレッド	16～64	16～64	16～64	16～64
キャッシュ容量	MB/core	2.0～4.0	2.0～4.0	2.0～4.0	2.0～4.0
コア間共有キャッシュ容量	MB	16.0	32.0	64.0	128.0
コア間同期・通信レイテンシ	ns	65.0	65.0	65.0	65.0
コア間同期・通信帯域	GB/sec	64.0	64.0	128.0	128.0
CPUチップあたりメモリ容量	GB	32.0	64.0	128.0	256.0
CPUチップあたりメモリ帯域	GB/sec	50.0	100.0	200.0	400.0
メモリレイテンシ	ns	40.0～80.0	40.0～80.0	40.0～80.0	40.0～80.0
ノード内 CPU チップ数	個/node	2	2	2	2
ノード内通信レイテンシ	ns	80.0～200.0	80.0～200.0	80.0～200.0	80.0～200.0
ノード内通信帯域	GB/sec	8.0	16.0	32.0	64.0

- クロック周波数はあまり変わらないが、コア数が増えていく(～64コア)
- コア当たりのキャッシュ容量は増えない
- メモリレイテンシも減らない

スループットプロセッサの性能予測

▶ HPCI技術ロードマップ白書、2012年3月、pp.12から転載

表 2-3 スループットプロセッサの性能予測

		2012 年	2014 年	2016 年	2018 年
設計プロセス	nm	32	22	16	11
レイテンシコア数	個/chip	0	2	4	8
スループットコア数	個/chip	32	64	128	256
クロック周波数	GHz	1.0～2.0	1.0～2.0	1.0～2.0	1.0～2.0
演算器数	FMA/Core	8	8	16	16
演算レイテンシ	ns	4.0	4.0	4.0	4.0
ハードウェアスレッド数	Thread/Core	4	8	16	32
スレッドあたりレジスタ数	本/スレッド	64	64	64	64
キャッシュ容量	KB/core	32.0～64.0	32.0～64.0	32.0～64.0	32.0～64.0
コア間共有キャッシュ容量	MB	16.0	32.0	64.0	128.0
コア間同期・通信レイテンシ	ns	100.0	100.0	100.0	100.0
コア間同期・通信帯域	GB/sec	64.0	64.0	128.0	128.0
CPU チップあたりメモリ容量	GB	8.0	16.0	32.0	64.0
CPU チップあたりメモリ帯域	GB/sec	150.0	250.0	500.0	1000.0
メモリレイテンシ	ns	40.0～80.0	40.0～80.0	40.0～80.0	40.0～80.0
ノード内 CPU チップ数	個/node	4	2	1	1
ノード内通信レイテンシ	ns	80.0～200.0	80.0～200.0	80.0～200.0	80.0～200.0
ノード内通信帯域	GB/sec	8.0	16.0	32.0	64.0

- クロック周波数はあまり変わらないが、コア数が増えていく(～256コア)
- コア当たりのキャッシュ容量は増えない
- メモリレイテンシも減らない

2018 年のシステム性能予想

- ▶ HPCI技術ロードマップ白書、2012年3月、pp.16から転載
- ▶ 2011 年から2012 年にかけてアプリケーション開発者を交えた合同作業部会を、複数回にわたって開催
- ▶ サイエンスロードマップを達成する為に最適と考えられる計算機構成を議論
- ▶ 京とほぼ同程度の消費電力・設置面積で実現できること(20MW~30MW、2000m²~3000m²)を制約条件

2018 年のシステム性能予想

▶ 以下の4種に分類

1. 汎用型:

京のように汎用的に様々な問題に適用可能

2. 容量・帯域重視型:

汎用型から演算性能を落として、メモリ性能により多くの資源を割く

3. 演算重視型:

メモリ性能を落とし、演算性能により多くの資源を割く

4. メモリ容量削減型:

メモリ容量を極限まで削減し、オンチップメモリですべての計算を完結

2018 年のシステム性能予想

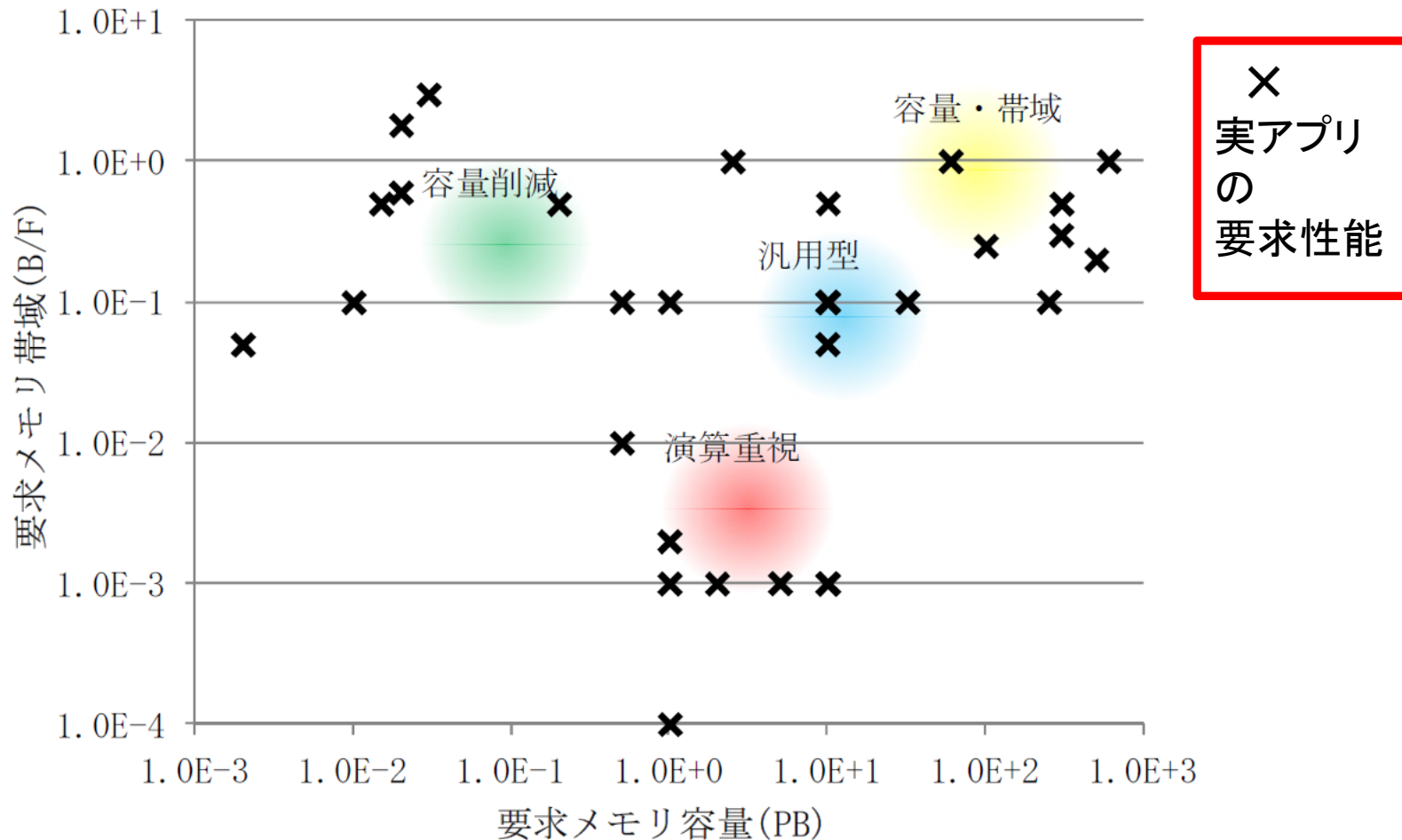
▶ HPCI技術ロードマップ白書、2012年3月、pp.16から転載

表 2-5 20MW で実現できるシステム性能予測（ただし，京のみ消費電力 12.66MW）

	総演算性能	総メモリ帯域	総メモリ容量
汎用（従来型）	200～400 PFLOPS	20～40 PB/s	20～40 PB
容量・帯域重視	50～100 PFLOPS	50～100 PB/s	50～100 PB
演算重視	1000～2000 PFLOPS	5～10 PB/s	5～10 PB
メモリ容量削減	500～1000 PFLOPS	250～500 PB/s	0.1～0.2 PB
京（参考）	10 PFLOPS	5 PB/s	1.2 PB

2018 年のシステム性能予想

▶ HPCI技術ロードマップ白書、2012年3月、pp.17から転載



2018 年のシステム性能予想

▶ HPCI技術ロードマップ白書、2012年3月、pp.17から転載

▶ ネットワーク

1. **Low-Radix ネットワーク**(Dragonfly網):
隣接するデータ同士が情報を交換することに適する
2. **High-Radix ネットワーク**(4D-Torus網):
Bisection 帯域やシステム直径の短縮に注力

▶ ストレージ

1. **ローカルストレージ**: 計算ノードに付加する
2. **グローバルストレージ**:
任意の計算ノードからアクセス可能

2018 年のシステム性能予想

- ▶ HPCI技術ロードマップ白書、2012年3月、pp.17から転載

表 2-6 ネットワーク諸元

	Injection	P-to-P BW	Bisection	最短遅延	最長遅延
High Radix (Dragonfly)	32 GB/s	32 GB/s	2.0 PB/sec	200 ns	1000 ns
Low-Radix (4D-Torus)	128 GB/s	16 GB/s	0.13 PB/sec	100 ns	5000 ns

表 2-7 ストレージ諸元

総容量	1EB	およそメモリ容量の 100 倍程度
総帯域	10 TB/s	メモリ容量を約 1000 秒で退避する程度

克服すべき課題

エクサフロップスコンピューティングに向けた技術的な5障壁

1. スケーリング
2. 消費電力
3. プログラミング
4. 信頼性
5. 入出力

スケーリングの問題

スケーリングの問題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.27
- ▶ エクサスケールの計算機では、コア数は
1000万コア～10億コアにまで上がっていく
- ▶ 各ノードのコア数は、1000コア～1万コアのオーダーになるとすると、ノード数も1万ノード～10万ノードになる
- ▶ 以上から、以下を実現しないといけない
 - ▶ (最悪は) 10億コア 対 10億コア の効率の良い通信方式
 - ▶ ノード内の通信時間 と ノード外の通信時間 の差を考慮した効率の良い通信方式
 - ▶ 1000コアを用いた1000スレッド実行(共有メモリ)の高効率実行

数値計算アルゴリズムと スケーリングの問題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.27
- ▶ 現在のスパコンでは、1コア当たりのメモリ容量はほぼ一定で構築されてきた
- ▶ 今後のトレンドとして、演算重視型のプロセッサでは、1コア当たりのメモリ容量は、現在のものよりも削減される
- ▶ このとき、数値計算アルゴリズムにおいて、問題サイズを N とすると
 - ▶ 演算量が $O(N \log N)$ のアルゴリズム(例)FFTなど)では、弱スケーリングでも、並列効率を維持するための、1ノード当たりの問題サイズがメモリ利用量から足りず、並列性能が下がっていく(通信時間の占める割合が多くなっていく)

スケーリング問題の技術課題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.27

I. 通信時間の削減

- 特に強スケーリングでは、メッセージサイズを減らすより、メッセージ回数(通信レイテンシ)を減らすことが重要。
- そのため、集団通信を極力避けるアルゴリズム再構築、演算量が増えても通信回数を減らすアルゴリズムの開発が必要
- 通信と計算をオーバラップするため、通信を分割すると通信回数が多くなっていく。通信量が増えても、通信回数を減らす通信方式を検討。

スケーリング問題の技術課題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.27

2. システムソフトウェアのスケラビリティ

- コア数1億を想定するシステムでは、各コアで利用できるメモリ量が従来より少なくなっていく。
- この少ないメモリ量でも稼働するシステムソフトウェアをつくることが必要。
 - ランタイム、OSによりスレッド管理
(通常はスレッド数に比例するメモリが要る)
- **OSジッタの問題**。OSの不均一な負荷が集団通信性能を劣化させる。

スケーリング問題の技術課題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.27

3. スケーラブルなプログラミング

- **プログラミングの困難性**: ハイブリッドMPI実行 (MPI+OpenMP) が性能的に有望だが、全ての問題に対応できるわけではない。より汎用なプログラミング言語と処理系の開発が必要である。
- **アルゴリズムのサポート**: 数値計算アルゴリズムレベルで自明な並列性を確保できるアルゴリズムが開発された際には、階層化された並列処理をなるべく隠蔽し、かつ、用意に性能が出せるようなプログラミング環境を提供する必要がある。

スケーリング問題の技術課題

▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.27

4. 数値計算ライブラリにおける高並列性の確保

- ノード外、ノード内、の階層的な並列性を利用できるように、数値計算ライブラリを構成する要素を部品化する必要がある。
- データ分散の方法は、これまでは一次元分散が主流であったが並列性が足りないため、二次元や三次元分散を活用するように変更する。

省電力の問題

省電力の問題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.35
- ▶ 2018年稼働を目指すエクサスケールのスパコンでは、20～30MW、かつ、京コンピュータの100 倍の性能要求
- ▶ **ハードウェアの要求**
 - ▶ 最小加工寸法の縮小
 - ▶ 超低電圧回路技術
 - ▶ 3次元積層メモリデバイス
 - ▶ 性能の観点から適切な低電力化技術の適用
- ▶ **運用時の要求**
 - ▶ スパコンの平常稼働時はピークの半分の電力
 - ▶ スパコンのピーク消費電力が電源系や空調系の限界を超えるように設置し、ソフトウェアやハードウェア制御機構を導入し、設備的限界の範囲内に消費電力・温度をおさえる工夫の検討

省電力のためのハードウェア技術

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.35、36
- ▶ ハードウェアの省電力技術
 - ▶ プロセッサの消費電力は電源電圧の二乗に比例
 - ▶ いままでは、電源電圧を0.7倍に下げ、トランジスタ数倍増と、周波数向上分の消費電力増、を打ち消し
 - ▶ 90nmプロセス以降のプロセッサでは電源電圧が1.0V 近辺で下げ止まり、周波数向上ができない
- ▶ 微細化による消費電力削減:
 - ▶ トランジスタ当たりの実装面積を下げることで低電力化

省電力のためのハードウェア技術

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.35、36
- ▶ パワーゲーティング:
 - ▶ プロセッサコア、キャッシュメモリなどのアーキテクチャ上のブロックごとにパワースイッチを実装し、状況に応じ電源を遮断する。ミリ秒単位でON, OFFするため、OS、ミドルウェア、アプリケーションの連携が必須。
- ▶ DVFS (Dynamic Voltage and Frequency Scaling):
 - ▶ 動作周波数と電源電力を下げることにより、消費電力を削減する。 $(\text{電源電圧})^2 \times (\text{動作周波数})$ に比例する。
 - ▶ 動作周波数とともに電源電圧も下げると、動作周波数の3 乗に比例する消費電力削減が実現可能。
- ▶ クロックゲーティング:
 - ▶ 状態を変える必要ないトランジスタへのクロック供給を遮断。

エクサフロップス実現のための技術

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.36
- ▶ 3次元構造トランジスタ(Fin FET, トライゲート・トランジスタ):
チャネル領域を立体構造化し、トランジスタを小型化
- ▶ 不揮発メモリ(MRAM、PRAMなど):
SRAMで生じるリーク電流を削減することで、低電力化
- ▶ 3次元積層デバイス:
複数のダイを3次元方向に積層し、これらの間を貫通ビアで接続する。低レイテンシ化、1チップ化による低電力化、微細化に頼らない大容量化ができるので、最も注目されている技術。
- ▶ シリコンフォトニクス:
光デバイスをシリコンで実現する。光通信の特徴から、通信性能を飛躍的に向上させ、かつ、通信に要する低消費電力の削減

ソフトウェア電力制御のための アーキテクチャ機構

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.37
- ▶ Power knobの必要性
 - ▶ アプリケーション単位での消費電力特性を設定するPower knobが必要
 - ▶ Power knobはアプリケーションから制御できることが望ましい
 - ▶ そのため、アプリケーション上の要求に対して遅延がある
 - ▶ 遅延があっても、ハードウェア上の消費電力の限界を超えてはいけない
 - ▶ オンライン消費電力モニタリング、リアルタイム化、の研究が重要

ソフトウェアによる省電力技術

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.37
- ▶ システムソフトウェアレイヤ
 - ▶ 数百万ノード級の電力情報をリアルタイムに取得しないといけない
 - ▶ 大域的な電力スケジューリングを実現しないといけない
 - ▶ 階層化されたI/Oシステムソフトウェアにより、データ移動を効率化することにより、低電力化する

ソフトウェアによる省電力技術

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.37

▶ プログラミングレイヤ

- ▶ システムソフトウェアおよびハードウェアと連携するためのプログラミングインターフェース
- ▶ コンパイル時情報の利用、各コアや各ノードの性能ばらつき拡大への対策、不揮発メモリ等があるプログラミング／コンパイル技術、電力制約下で最大性能を得るための性能予測モデル、など

▶ アプリケーションレイヤ

- ▶ アプリケーションアルゴリズムの特性からの低電力化
- ▶ 問題の定式化(決定的手法 対 モンテカルロ法など)、連続系の離散化(FEM 対 Spectral法)、アルゴリズムとデータ構造の選択による低電力化の方法

プログラミングの問題

プログラミングの問題

- ▶ 参考:
[小柳ほか、2012]小柳義夫、中村宏、佐藤三久、松岡聡:
計算科学 別冊:スーパーコンピュータ、岩波書店、2012年、pp.190-195
- ▶ プログラミングできないと、エクサフロップスのハードウェアが存在しても意味がない
- ▶ ソフトウェアの生産性も考慮しないといけない
 - ▶ 生産性の観点から、デバック環境、性能モニタ、数値計算ライブラリの利用、などの考慮されないといけない

プログラミングにおける技術要求

- ▶ 参考:[小柳ほか、2012]小柳義夫、中村宏、佐藤三久、松岡聡:計算科学 別冊:スーパーコンピュータ、岩波書店、2012年、pp.190-195
- ▶ **プログラミングモデル**: MPI+OpenMPによるハイブリッドMPIモデルなど、ノード内、および、ノード外の並列性を使えるモデルの実現と、そのモデルをどのように記述するのかの方法
- ▶ **データ分散**: どのようにデータ分散し通信を最小化できるように記述するのか
- ▶ **ヘテロジニアス**: 多数のメニーコアと少数のマルチコアの組合せ時において、計算対象をどのように分担させるのか、および、その記述性

プログラミングにおける技術要求

- ▶ 参考: [小柳ほか、2012] 小柳義夫、中村宏、佐藤三久、松岡聡: 計算科学 別冊: スーパーコンピュータ、岩波書店、2012年、pp.190-195
- ▶ **移植性**: ピュアMPIだけではない新しいプログラミングモデルを、いかに新しいアーキテクチャ上に、低いコストで移植できるか
- ▶ **再利用性**: 従来から使われているコード (レガシーコード) も利用できるような仕組み
- ▶ **省電力化**: 省電力化のためのPower knowを制御できる言語など
- ▶ **I/O**: 高性能なI/Oを考慮した言語や言語処理系

既存のプログラミング技術と問題点

- ▶ 参考:[小柳ほか、2012]小柳義夫、中村宏、佐藤三久、松岡聡:計算科学 別冊:スーパーコンピュータ、岩波書店、2012年、pp.190-195
- ▶ 従来成功してきた、逐次言語とMPIのSPMDモデル、およびピュアMPI実行の問題点
 1. 強スケーリング問題:超並列化による負荷バランスの維持をどうするか
 2. データ分割問題:通信レイテンシを隠ぺいするデータ分割をどうするか

既存のプログラミング技術と問題点

- ▶ 参考:[小柳ほか、2012]小柳義夫、中村宏、佐藤三久、松岡聡:計算科学 別冊:スーパーコンピュータ、岩波書店、2012年、pp.190-195
- 3. **システムバッファ問題**: 超並列化によるプロセス数の爆発的增加で、MPIのシステムバッファが膨大になり、起動できなくなる
- 4. **データ共有問題**: ノード内のデータの量は少なくなっても、スレッド並列化などの共有データに対する並列処理数は増加する。効率よく、スレッド並列化などの実行をするためには、どのように共有データを取り扱うか
 - ▶ 自動並列化コンパイラが取り扱う共有データ
 - ▶ OpenMP, OpenACCが取り扱う共有データ
 - ▶ CUDA, OpenCLによりデータ並列言語が取り扱う共有データ
 - ▶ その他 (Co-Array Fortran, Chapel, Unified Parallel C,) X10, XscalableMP

信頼性の問題

信頼性の問題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.30
- ▶ 故障があっても障害を引き起こさないこと(耐故障性)を上げる
- ▶ 耐故障性の処理完了は、平均故障間隔(MTBF)より小さい必要がある
- ▶ エクサスケールにおける故障
 - ▶ ハードウェア、ソフトウェアともに大規模化・複雑化するため、**難検知故障(Silent Error)**が増える。また、検知はできるが発生場所・要因の特定が困難な複合故障が増える。
 - ▶ 故障率が、構成要素の増加につれて増大する。
 - ▶ 平均故障間隔(MTBF: Mean Time Before Failure)、または平均中断間隔(MTTI: Mean Time To Interrupt)が短くなり、**平均故障間隔が35分 から39 分に。**
 - ▶ 1ソケット当たり年間0.1 回程度の故障が発生。**100 万ソケットのシステムを仮定すると、5 分程度の平均故障間隔。**

信頼性の問題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.32
- ▶ エクサスケールにおける耐故障技術
 - ▶ ECC(データに冗長な情報を追加)、ハートビート(一定間隔で故障しているか調査する)、異常状態検出(異常状態でないか検査する)、冗長実行(複数ハードウェアで実行)、再実行、チェックポイント・リスタート(正常状態を記憶しておき、故障時に読み込み再実行する)
 - ▶ 以下の要素技術が必要
 - ▶ 故障予測: 何処で、どのような故障が起きるか予測する
 - ▶ 故障解析: 故障による異常がどのように伝搬するか解析する
 - ▶ インターフェース: 検出した故障や異常を、他の部分と共有するための枠組み
 - ▶ 予防復旧: 故障に至る予兆を検知して予防的な再起動などを行う
 - ▶ 実装箇所
 - ▶ ハードウェア(ECCや冗長実行など)、ソフトウェア(プログラマが記載、プログラム言語が自動)

信頼性の問題

- ▶ 参考: HPCI技術ロードマップ白書、2012年3月、pp.33
- ▶ エクサスケールにおける耐故障性の課題
 - ▶ 故障と対処モデルの解析
 - ▶ 故障分析・故障モデル・故障予測
 - ▶ Fault Resilience (多重または相補的に障害発生を防ぐ枠組み)
 - ▶ Fault resilience のためのフレームワーク
 - ▶ Fault resilienceを実現するシステムソフトウェア
 - ▶ アプリケーションレベル耐故障性実装支援: プログラム開発者がメモリ利用法やデータ更新のタイミングをディレクティブなどで指定し知識を与える。
 - ▶ ハードウェア・システムソフトウェアレベル耐故障
 - ▶ 新ハードウェアに対応した耐故障性技術
 - ▶ たとえば、GPUなどのアクセラレータにおけるチェック・ポインティング、不揮発メモリを利用した耐故障性の効率化、など

入出力の問題

入出力の問題

- ▶ 参考:[小柳ほか、2012]小柳義夫、中村宏、佐藤三久、松岡聡:計算科学別冊:スーパーコンピュータ、岩波書店、2012年、pp.196-198
- ▶ スパコンの用途として、演算のみだけではなく、データ処理を扱う分野(ビックデータ分野、データサイエンス分野)が注目されており、今後、スパコンの利用促進がされていくと期待される
- ▶ この時、演算時間よりも、ファイルI/Oなどの、入出力の時間が演算時間のほとんどとなる
 - ▶ 例)ゲノム分野:
ゲノムデータベースへのアクセス時間がほとんど

スパコンにおけるI/O能力

- ▶ スパコンのI/O能力は、ふつうの計算機をしのぐ性能をもっている
 - ▶ 最大で、数十GB/sec. ～数TB/sec.
 - ▶ ストレージサーバを並列にアクセスできるファイルシステム
 - ▶ PVFS、GPFS、Lustre、FEFS(富士通)、HSFS(日立) など
 - ▶ SSD(Solid State Drive)をノードに搭載し、一次記憶の時間を高速化する、など
- ▶ ただし、多数のファイルのアクセス(メタデータアクセス)、共有ディスクへのデータ書き込み、は遅いことがある
 - ▶ 共有ファイルへのデータ書き込みをせず、並列計算用のファイルシステムへデータをコピーしてから使う**ステージング**が一般的

I/Oにおける技術課題

▶ 以下の技術課題がある[小柳ほか、2012]

1. 容量・消費電力

- ハードディスクを構成する物理素子が電力を食うため、容量向上に制約がある

2. 転送速度

- 容量が確保できても、データ転送のハードウェア制約がある(参照:プロセッサの技術トレンド)
- 容量が2倍になっても読み出し速度は線形のため、速度は $\sqrt{2}$ 倍しかない。読み出し速度律速となる

3. 信頼性

- データ容量が増えると故障率も高まる。RAID5のような2重多重化では不十分で、最低でもRAID6のような3重多重化が望まれる。

4. ソフトウェア

- エクサバイト級のデータマネージは容易でない。
- OS、ミドルウェア、データベース、プログラム言語、I/Oライブラリ

現在の状況

日本における技術課題の解決状況

▶ HPCI技術ロードマップ白書、2012年3月、pp.15から転載

表 2-4 2018 年に向けた目標値と現在実現された数値の差分

課題	2018 年の目標値の例	2011 年までに国内で達成した成果
大規模並列	十万 CPU, 数百コア/CPU のシステム	88128CPU, 8 コア/CPU
ヘテロジニアスアーキテクチャ	特性の異なるコア（レイテンシコア・スループットコア）が混じり合った環境で高効率・高性能な計算を実現する	TSUBAME2(GPU)でノード内・ノード間の不均一性を有する大規模システムを構築
メモリ	B/F = 0.1 で高効率な計算を実現	メモリ階層を活用したライブラリ・フレームワークの開発
耐故障性	10 万プロセッサで MTBF=1 日以上	高信頼 CPU 約 10 万個を利用した Linpack で 28 時間動作させ 10PFLOPS を実現
電力	50GFLOPS/Watt の演算性能	アクセラレータ利用で約 1.6GF/Watt 大規模構成で 829MFlops/Watt (京)
生産性	上記環境での生産性向上	SCore, XcalableMP, HPC-MW で~1 万コア程度まで実績がある

次世代システム調査研究と コ・デザイン

HPCIと 将来のHPCIシステムのあり方の調査研究

- ▶ HPCI (High Performance Computing Infrastructure)とは
 - ▶ https://www.hpci-office.jp/pages/concept?parent_folder=21から抜粋
 - ▶ 科学技術の分野における国際競争が激化する中で、日本が科学技術で世界をリードするため、創造的・先端的な研究開発の推進が重要
 - ▶ 計算科学に基づく数値シミュレーションは実験や観測が困難な現象の解明や技術開発における精度の向上等に大きく寄与しており、今後、情報科学技術の分野のみならず多様な物質・材料の構造・物性及び機能の解析や、遺伝子レベル及び人体全体の解析等を通じて、ナノテクノロジー・材料、ライフサイエンス、ものづくり、環境、防災、航空・宇宙及び原子力等、広範な分野において、計算科学技術を通じて実施された研究開発が、国際競争力の飛躍的な向上につながる研究成果を世界に先んじて創出していくことを期待
 - ▶ このため、「京」を中核として他の全国の主要なスーパーコンピュータを高速ネットワークでつなぎ、革新的ハイパフォーマンス・コンピューティング・インフラ(HPCI)を構築する

HPCIと 将来のHPCIシステムのあり方の調査研究

▶ 将来のHPCIシステムのあり方の調査研究とは

- ▶ http://www.mext.go.jp/b_menu/shingi/chousa/shinkou/020/shiryo/attach/1319672.htm から抜粋
- ▶ HPC技術等のHPCIシステムの高度化に必要な技術的知見を獲得することを目的とした調査研究を平成24年度、平成25年度に実施
- ▶ 10年後の我が国の社会的・科学的課題の解決という視点から複数のシステムを厳選
- ▶ システムの選定にあたっては、アプリケーション&アーキテクチャ・コンパイラ・システムソフトウェア合同作業部会においてとりまとめられた、「今後のHPCI技術開発に関する報告書」を踏まえて行う
- ▶ 各システムについて、ハードウェアの技術動向調査、システム設計研究・システムソフトウェアの検討等を行い、5～10年後の我が国のHPCIシステムに必要な技術的知見を獲得
- ▶ 事業終了年度に各システムに関し評価を行い、その結果を踏まえ、今後のHPCI構成システム等のあり方に反映

HPCIと 将来のHPCIシステムのあり方の調査研究

- ▶ 将来のHPCIシステムのあり方の調査研究、採択機関
 - ▶ http://www.mext.go.jp/b_menu/shingi/chousa/shinkou/020/shiryo/__icsFiles/afieldfile/2012/08/07/1323974_1_2.pdf から抜粋
 - ▶ 東京大学、筑波大学、東北大学の各々を中心とする3チーム
 - ▶ アプリケーション分野からみた将来のHPCIシステムのあり方の調査研究〔事業代表者:理研 富田浩文〕
 - ▶ レイテンシコアの高度化・高効率化による将来のHPCIシステムに関する調査研究〔事業代表者:東大 石川裕〕
 - ▶ 演算加速機構を持つ将来のHPCIシステムのあり方の調査研究〔事業代表者:筑波大 佐藤三久〕
 - ▶ 高メモリバンド幅アプリケーションに適した将来のHPCIシステムのあり方の調査研究〔事業代表者:東北大 小林広明〕

実施内容について

- ▶ 公開資料は、理研FSで提供
 - ▶ <http://hpci-aplfs.aics.riken.jp/document.html>
- ▶ ここでは、東大FSの内容について紹介する

参考文献

- ▶ HPCI技術ロードマップ白書
 - ▶ <http://www.open-supercomputer.org/workshop/sdhpc/>
- ▶ 小柳義夫、中村宏、佐藤三久、松岡聡：計算科学 別冊：スーパーコンピュータ、岩波書店、ISBN978-4-00-011307-6、2012年

レポート課題

- I. [L10] エクサスケールコンピューティングを実現するための技術的な障壁について、現在の技術のレベルと、要求されるレベルを調べ、その要点をまとめよ。

また調べた事項から予想される、将来の研究課題について、自分の意見を主張したうえで、考察せよ。