エクサスケールコンピューティングのアプリケーションとアーキテクチャ── 「欲しかったのはこれじゃなーい!!」と叫ばないために

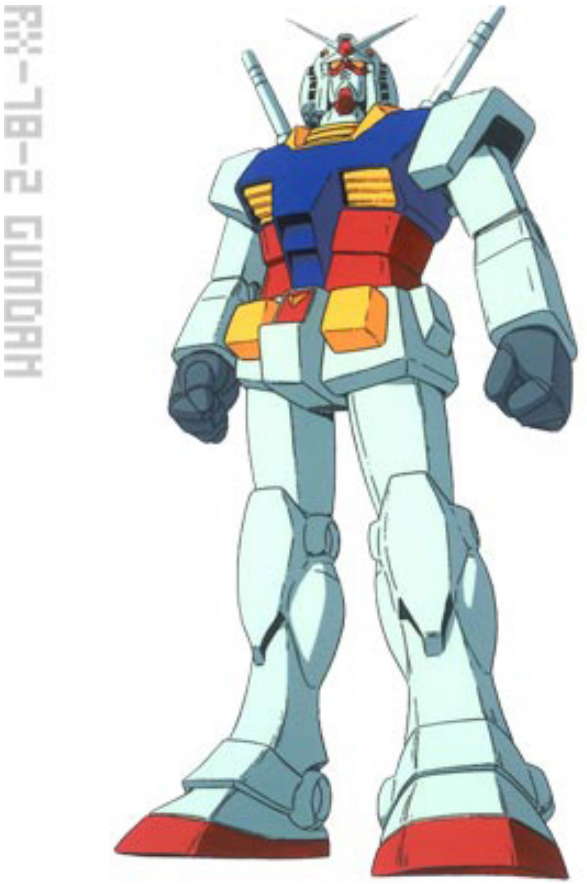
牧野淳一郎

東京工業大学理工学研究科

理学研究流動機構
(4/1 から 「地球生命研究所」?)

理論懇シンポジウム 2012/12/23

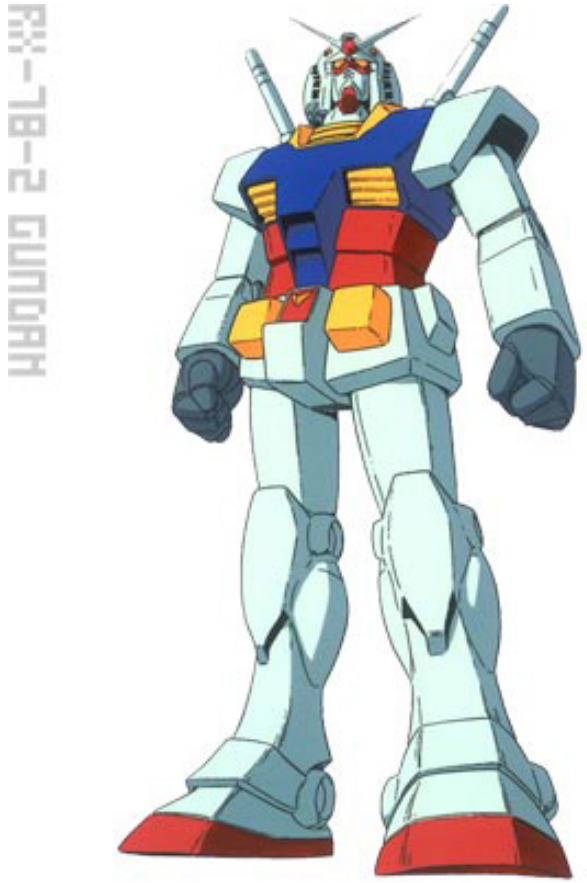
欲しかったのは



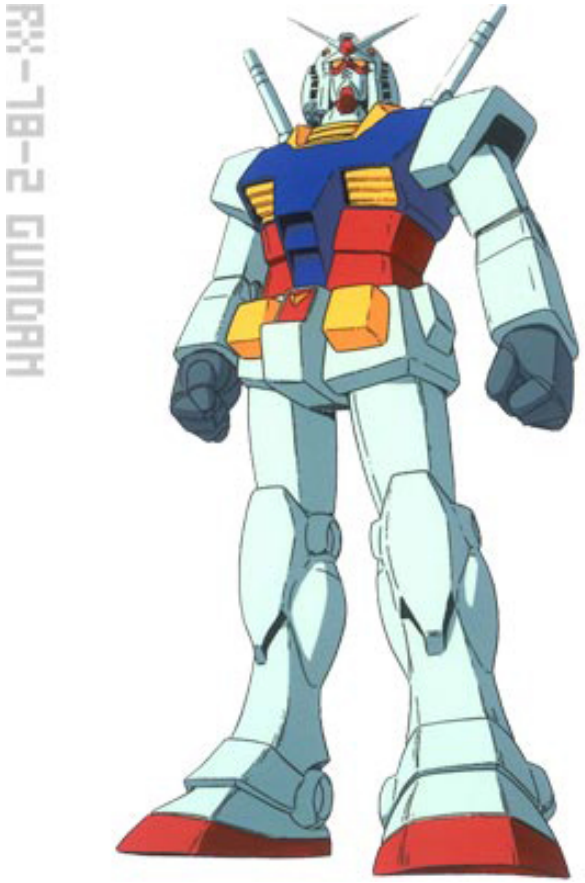
届いたのは

(http://www.zariganiworks.co.jp/korejanairobo/)

何か違う、、、



どうしてこうなった、、、



とならないために、、、

概要

- 技術トレンドとアプリケーション
- 「京」プロジェクトのデザインと概念設計
- 現在進行中のプロセス
  - 昨年度のアプリケーション部会等の復習
  - ４つのアーキテクチャイメージ
- おまけ: 今後の方向について

技術トレンドとアプリケーション

- この 40 年間、「大規模数値計算」をするのはどんどん大変になってきている。?
- それは何故か？
- どうするべきか？

というような観点から

何故大変になってきているか

- 計算機が速くなったので難しい物理をいれられるようになった
  — これはまあしょうがない
- 計算機を効率的に使うのが難しくなった (面倒になった/あるいは問題によっては不可能になった)
  — こっちは何故か？なんとかならないか？

HPC 技術の方向—ハードウェアの観点

- ここ 40 年間の計算機の能力の進歩の原動力: 半導体技術、特に LSI の微細化の進歩。元々のムーアの法則: 18ヶ月でトランジスタの数 2 倍
- 1970 まで: 計算機 1 台で、演算器 1 つ。演算器の高速化にトランジスタ使う (CDC7600, Cray-1)
- 1990: CMOS LSI だと 1 チップで高速な演算器を持つプロセッサ実現 (Intel 80860)。ここからマルチコアになりそうだが 10 年ならなかった。
- 2000: CMOS スケーリングの限界 (動作電圧が下がらない+クロックが上がらない):マルチコア・SIMD 化急速に進む。

この間ずっと: メモリバンド幅相対的には段々下がる:ピンリミット

つまり

2000 年以降急速に駄目になった。駄目さはどんどん上昇している。

- コア数・演算器数の増加、通信オーバーヘッドの相対的増加:
  大規模 (メッシュ数とか粒子数多い) 計算でないと性能でない
- キャッシュに入らないとすごく遅くなる
  大規模 (メッシュ数とか粒子数多い) 計算だと性能でない

あれれ？

実際には:演算に対してメモリアクセスが少なければいい。粒子法は結構なんとかできる。メッシュは演算量の多いスキームを頑張って最適化すれば、、、

これが欲しかったもの？

- 大規模で、1 ステップの計算が重いと性能がでることもある
- 小規模な系の長時間計算は？
- プログラムどやって書くの？N 君の (時間 $\times$ 興味) は有限

なんか考えとかないと次はもっと大変なことになっても知らないよ？

「京」プロジェクトのデザインと概念設計

- 大体の時系列
- ターゲットアプリケーション
- 「概念構築に関する共同研究」

大体の時系列

- 2004 年くらいから文科省の下の情報科学技術委員会、計算科学技術推進ワーキンググループで議論
- 2005 年の「中間報告」
  - 8 分野からの要求をまとめた (ことになっている)
  - ベクトル 2PF +スカラー 4PF +「特定処理計算加速機」20PF
- 2006 年度「次世代スーパーコ ンピュータ：概念構築に関する共同研究」実質的にはデザインコンペ。NFH の他、東大 (+天文台)、九大、筑波大等も参加

時系列続き

- 2005 年から 2006 年にかけて CSTP 評価委員会からボロクソに言われる。目標、アーキテクチャを文句がでないように色々変更。
- 2006 年初め (だったと思う) 開発実施本部設置、ヘッド:渡辺 (NEC OB)
- 2007 年夏: ベクトル (?PF) +スカラー (?PF) 合計 10PF 以上、と決定
- 2009 年春: ベクトル担当の N が撤退
- 2009 年 11 月 13 日 (金) 仕分け。「2 位じゃいけないんですか？」
- 2010 年 10 月。プロトタイプ機が Green500 4 位にランクイン
- 2011 年 6 月。8 割完成で Top500 1 位
- 2011 年 11 月。全ノード動作で Top500 1 位。

ターゲットアプリケーション

http://www.nsc.riken.jp/target-application/target-application.htm

21 個 (斎藤君のとか似鳥君のもある) 但し

実際に性能評価に使われたのは HPL, FFT と以下の 7 個

- SimFold, Modylas: 古典 MD
- GAMESS, RSDFT: 量子化学
- LANS, NICAM: 流体
- LQCD: 4 次元格子での CG 反復

この時点で、

- したい計算
- 高効率でできる計算
- 高効率でできるがオーバースペックな計算

の乖離

大雑把なアプリケーションの特性

- 古典MD、量子化学: メモリバンド幅もネットワークもあまりいらない
- 流体: メモリバンド欲しい。ネットワークはそこそこ
- QCD: メモリ量いらない。メモリバンド幅もネットワークも欲しい

これからわかること:

- 一台で全て満たそうとすると帯に短し襷に長しになる
- 低レイテンシ要求があるアプリケーションははいってない

「京」のデザインポイント

- CPU:128Gflops, 64GB/s, 16GB
- ICC: リンク 5GB/s x 2 x 10, CPU 20GB/s, レイテンシ:論文に書いてない
- 最近の計算機にしては B/F 重視
- 最近の計算機にしてはネットワーク重視
- CPU, ICC 別チップ

BG/Q との比較

	「京」	BG/Q	比率
ピーク速度 (GF)	128	204.8	–
メモリバンド幅 (GB/s)	64	42.6	2.4
ネットワーク (GB/s)	5	2	4
電力あたり性能 (GF/W)	0.85	2.1	(1/2.5)

演算性能あたりで、「京」は BG/Q の 2.4 倍のメモリバンド幅、 4 倍のネットワークバンド幅を持ち、 2.5 倍電気を食う。

どうしてこうなったか？

形式的な理由: 要求仕様にそう書いてあった

- HPL 10PF
- **電力** 30MW **以下**
- **アクセラレータ付けるならやはり** 10PF
- **この範囲でアプリケーションの性能を上げること**

つまり

- 「京」より演算性能を大幅に上げるのは、メモリバンド幅やネットワークバンド幅をあげなくてもいいなら電力・コストを大きくは増やさなくてもできたはず
- 逆に、 HPL 10PF だけならもっと安価にもできたかもしれない
- メモリバンド幅を増やすなら全く別の作りかたもあったかも
- レイテンシ小さくするのはまた別のアプローチがありえた

エクサはどこに向かっているのか？

- 昨年度の文部科学省のなんとか部会
- 4つのアーキテクチャイメージ
- 今年度の動き

昨年度の文部科学省のなんとか部会

- 去年の夏に突然つくれという話が発生した。
- アーキテクチャ・コンパイラ・システムソフトウェアのほうは SDHPC 検討グループ (大計のなにか) が横すべり
- アプリケーションは戦略分野等から人を集めて急拠立ち上げ。9-11 月に集中的にミーティング、検討。

以下昨年 11/15 の合同部会での牧野の報告から抜粋

予備検討の方針(1)

- アーキテクチャ部会での検討では、B/F、メモリ量、ネットワークバンド幅等について非常に狭い範囲しか想定していないように思われた
- 消費電力当り性能は、エクサスケール実現にとって大きな壁である。
- B/F はアプリケーションの効率に大きく影響する一方、消費電力当り性能にも大きな影響をもつ
- メモリ量、ネットワークバンド幅も、大きく変えれば電力に影響する

アプリケーション側で、 B/F、メモリ量、ネットワークバンド幅の必要量をだしておこう

予備検討の方針 (2)

- アプリケーション、アルゴリズムにより B/F 等への要求は変わる
- 特に、同じアプリケーションでもアルゴリズムが変われば、またアルゴリズムが同じでも系のサイズ等だけによっても要求は変わる

といった問題があるので、

- 各分野に、重要なアプリケーション (計算法、系のサイズ等含めて) を選定してもらい、それぞれについて要求を見積もってもらう
- それらをいくつかのタイプに分類できるかどうか検討する

という方針を考えた。

アプリケーション

38 アプリケーション

分野	数
1	7
2	13
3	4
4	8
5	7

検討結果

(詳しい話は今日は省略)

- B/F とネットワークバンド幅は関係あり。
- B/F 要求高いがネットワークは弱くていいものはある。逆はない
- ランダムアクセス、非数値計算等、この軸ではよく表現できない要求もある
- 大雑把に数種類にタイプわけできそう

タイプわけの観点

観察:

- $\mathrm{B/F}$ は 0.1 以上の高いものと、桁で小さいものに分かれる
- メモリ要求にも非常に幅がある

注意事項:

- 分野によってはまだ十分な検討が進んでいない
- 分野によっては、そもそもアプリケーション・アルゴリズムの進化が速いために定量的な要求を明確にしにくいところもある

タイプわけの観点

「アプリケーションタイプ」でなくて「アーキテクチャタイプ」としてみた。

- そのほうが物理的制約をイメージしやすい
- アーキテクチャ側との議論もしやすい？

タイプわけ

以下の 4 タイプ

タイプ	B/F	メモリ量 (1TF)	消費電力 (1EF)	演算性能) (20MW)	バンド幅 (20MW
ベースライン	0.1	10-100GB	20MW	1EF	0.1EB/s
SoC	4	5-10MB	2-5MW	4-10EF	16-40EB/s
アクセラレータ	0.001	1-10GB	4-10MW	2-5EF	2-5PB/s
バンド幅重視	1	1TB	120MW	0.15EF	0.15EB/s

最終レポートでは

報告書での用語	アプリ部会のつけた名前	具体的イメージ
汎用	ベースライン	「京」の延長
容量・帯域	バンド幅重視	NEC SX-9 の延長
演算重視	アクセラレータ	GRAPE-DR,GPGPU
メモリ削減	SoC	…

注意

- 「汎用」は汎用ではない (「京」でうまく効率がでないアプリケーションはいくらでもある)
- 「容量・帯域」が本当にそのどっちかでも実現できる設計解があるかどうかは自明ではない
- 演算重視は要するに $\mathrm{B/F}$ 要求が低いものを対象にする
- 「メモリ削減」はオンチップメモリないし 3 次元実装でバンド幅を増やすのが本質。小容量になるのは結果。レイテンシもつめる。

汎用/ベースライン

- 「京」の延長
- B/F 0.1 $\sim$ 0.2?
- 富士通さん頑張って

演算重視/アクセラレータ

- メモリ帯域が少なくて良いとなったアプリケーションの大半が量子系 (密行列の直交化や対角化が計算量のほとんどを占める)
- 後は大規模な粒子系
- GPGPU 的なものでいいが、アクセラレータ側に外部メモリはあまりなくてもいい (そちらにメモリあるならホストはなんのため?という問題も)
- GRAPE-DR ベース？

メモリ削減/SoC

- 想定アプリケーション: 小サイズ MD、流体、QCD 等
- 大サイズ差分法を out-of-core でできるかどうかは要検討
- 外付けメモリがないか、極端に低バンド幅
- 軽量コアを非常に多数集積して、電力当り性能を上げる
- オンチップメモリに対しては 高 B/F
- メモリはチップ当り 1GB 以下程度？
- ネットワークは 100GB/s 程度？

容量・帯域/バンド幅重視

- SX-9 の延長
- NEC さん頑張って、、、

演算重視・メモリ削減の
もうちょっと具体的なイメージ

- 大雑把には: 70-80年代の大規模SIMDマシンを1チップ化。Goodyear MPP, CM, MasPar 等
- 例: CM-2。 2048 FPU, トータル 512MB メモリ
- 14nm だと 16384 FPU, 256-512MB メモリくらいが入るかも。16-32TF
- 1024-16384 チップで 32-520PF くらい。メモリは 512GB-8TB
- 小規模な計算がすごく速く走る。
- 大規模計算には工夫がいる？そもそも無理？

今後の方向

- 今年度、来年度 feasibility study
- 筑波大+日立 (私もはいってる)「演算重視+メモリ削減」
- 東大+富士:「汎用」
- 東北大+NEC:「容量・帯域」
- その他、今年度なんとかワーキンググループが 3 個。関係は私にはよくわかってない
- その後？5 年くらいで開発？

最悪シナリオ

- 3 つの FS チームが「競争」に
- 同じようなデザインになる
- 1 つ作るだけになる

次悪シナリオ

- 「複数作る予算ないよね」
- 「汎用」じゃないと駄目だよね？
- 1 つ作るだけになる

前回のキーポイント: 概念構築のためのなんとか

- 概念構築といいつつすでに枠が決まっていた
  - HPL 10PF
  - 電力30MW以下
  - アクセラレータ付けるならやはり10PF
  - この範囲でアプリケーションの性能を上げること
- その結果、各社の提案が収斂した

以下、2012/6 に公開になった当時の会議資料から

(2012/6 公開)

秘 回収資料

資料2-2

次世代スーパーコンピュータの概念設計について 続き)

平成19年3月27日

理化学研究所
次世代スーパーコンピュータ開発実施本部

(2012/6 公開)

アーキテクチャ案の概要（汎用システム）（6月末時点）

- 基本要求仕様（性能評価の基準とするシステム構成）
  - 理論ピーク性能：10PFLOPS
  - 総メモリ容量：2.5ペタバイト

アーキテクチャ案	NEC	日立	富士通	筑波大学
コア数 （コア：1演算プロセッサ）	中並列 10万以下	高並列 10～50万	超並列 50～100万	
計算ノード数	～5万		10～15万	
高速演算機構	ベクトル		SIMD	
消費電力（本体のみ）	20-30 MW	10-20 MW	20-30 MW	10-20 MW
設置面積	3000 ㎡以上	1000-2000 ㎡	3000 ㎡以上	1000-2000 ㎡

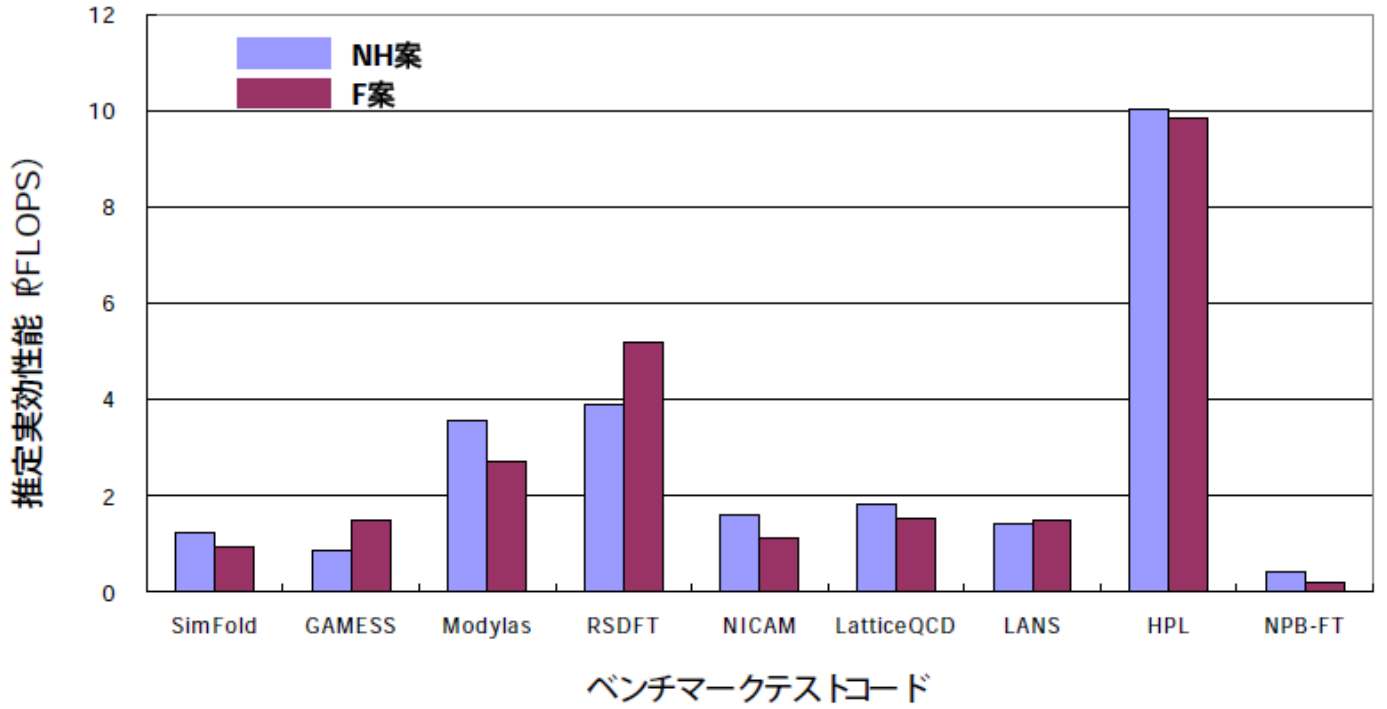
(2012/6 公開)

提案システムの演算部性能の比較

			NH案	F案
演算コア	動作周波数（GHz）		2	
	演算性能（GFLOPS）		64	16
	演算加速機構 （演算器数）		ベクトル型 （16: 2FMA x 8VPP）	SIMD型 （4FMA）
	レジスタファイル		ベクトルレジスタ 256要素×64本	スカラレジスタ 128本
CPUチップ （計算 ノード）	演算性能（GFLOPS）		256	128
	演算コア数		4	8
	メモリバンド幅 （Byte/Flop）		0.5	
	L2 キャッシュ	容量（MB）	8	6
		Byte/Flop	4	2
		特殊機構	選択的登録機構	ライン・ロック機構

(2012/6 公開)

ベンチマーク・テストによる性能予測 (詳細9本)



- ターゲット・アプリケーションから7本のベンチマーク・テスト, 及びHPL, NPB-FTについて, 実効性能を推定.
- いずれのベンチマーク・テストもほぼ同等の性能.

NH/F の比較

- ベクトル・スカラーで違うはずだったが、アーキテクチャパラメータはほとんど同じものに収斂
- さらに消費電力等も収斂
- さらにベンチマーク性能も収斂

おまけ: アクセラレータについては？

アーキテクチャ案の概要（アクセラレータ）（6月末時点）

- 基本要求仕様
  - アクセラレータ部の理論ピーク性能：10PFLOPS
  - 汎用サーバ（ホスト）のI/Oインターフェースに接続
  - ホストより指定された演算処理をアクセラレータで実行し，結果をホストに格納

アーキテクチャ案		国立天文台	東京大学
アクセラレータ	アーキテクチャ	SIMD型 プロセッサアレイ	
	プロセッサチップ数	約 15,000	約 20,000
	ボード数	4,000	2,500
ホストサーバ数		2,000	2,500
アクセラレータ部 消費電力		-10 MW ※平成24年6月公開時の注意書き 消費電力については、10MW以下、10-20MW、20-30MWの範囲でまとめたもの。提案は、ホスト部を除いて、1.7MWであった。	-10 MW ※平成24年6月公開時の注意書き 消費電力については、10MW以下、10-20MW、20-30MWの範囲でまとめたもの。提案は、ホスト部を除いて、0.68MW（案1）、0.88MW（案2）であった。

(2012/6 公開)

概念設計評価時の判断

- アクセラレータ案の説明資料には消費電力「-10MW」という謎の表現が。議事録には「10MW 以下くらい」とある。
- 実際の提案の数字は公開時注意書きにあるように 0.88-1.7MW

アクセラレータを採用しなかった理由

公式の説明

2者のシステム構成により、目標性能達成の見込みが確認できたため、アクセラレータの採用は考慮しない

- 目標 = LINPACK 10PF + HPCC 4 種1位
- アプリケーションのことはすっかり忘れられた、、、

まとめ

- エクサスケールにいくにあたっての技術的問題及び社会的・政治的状況を概観した
- プロジェクトは (実現可能なら) 設定した目標通りのものが実現される
- 国家プロジェクトだと何故か、Top500 1 位とか HPCC 1 位とかが目標になったりする。
- こっちでやりたい (かつ、できる) ことからあまりずれないように注文はつけていかないといけなさそう。