

# 先進計算機構成論 08

---

東京大学大学院 情報理工学系研究科 創造情報学専攻

塩谷 亮太

shioya@ci.i.u-tokyo.ac.jp

# 質問や感想への回答

- row hammerやcold boot attackのようなものはどうやって発見されたのかが気になりました。メモリやCPUに詳しい人ならある程度思いつきやすいようなものなののでしょうか。
- メモリの読み出しについて、ビット単位での読み書きはあまり行われないイメージだったのですが、連続した領域にアクセスしたい場合にはcolumn selectorなどの拡張でなんとかしているのでしょうか？

# 質問や感想への回答

- 今やっているような話の復習ってどのようにしたら良いでしょうか？私は数理最適化が専門で、普段はハードウェアを意識することはほぼないのですが、全く知識をもたずにプログラムを書くのもどうなんだと思って受講しております。

# 質問や感想への回答

- ◇ "条件分岐を減らすと高速になるのテクに興味があります．例えば  $f(i) \mid i \in [0, 1000000)$  の最小値を求めたい場合，① の様に if を使うよりも，② の様に min 関数を使うほうが良かったりしますか？

- ◇ ①

- ◇ `int min_value = INF;`
- ◇ `for(int i = 0; i < 1000000; i++){`
- ◇  `int v = f(i);`
- ◇  `if(min_value > v){`
- ◇  `min_value = v;`
- ◇  `}`
- ◇ `}`

- ◇ ②

- ◇ `int min_value = INF;`
- ◇ `for(int i = 0; i < 1000000; i++){`
- ◇  `int v = f(i);`
- ◇  `min_value = std::min(min_value, v);`
- ◇ `}`

# 質問や感想への回答

- レジスタファイルというのはレジスタ値が登録された表という理解で合っていますか？
- Row hammerやCold Boot Attackなどの手法が大変興味深かったです。こういった話題が出てくる国際会議としては、主にどこをチェックしておくべきでしょうか？

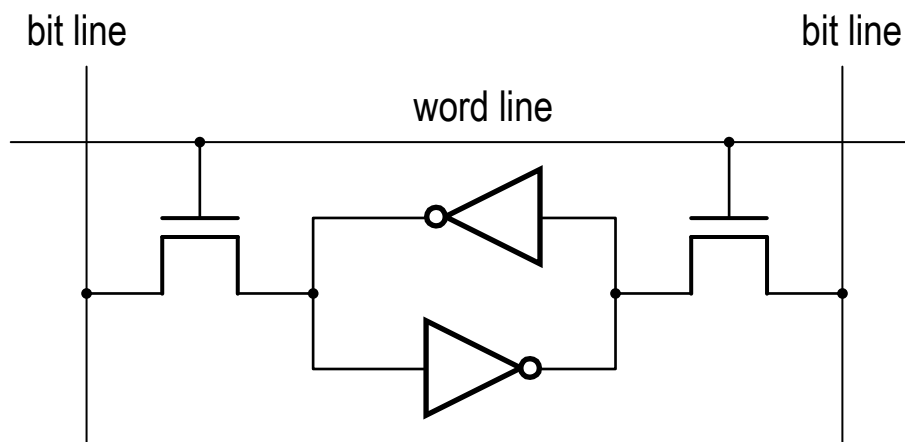
# 質問や感想への回答

- パーセプトロン素人なのでパーセプトロン予測器への学習方法がいまいちよく分かりませんでした。
- RASが出てきて、まだメモリの話に入ってもないのに何かしらをスタックするテーブルがめちゃくちゃあるな、と正直思いました。RASやBTBなどこんがらがってきたので、体系立てた図などあればうれしいです。

# 質問や感想への回答

- SRAMの書き込みのところなのですが、NOTゲートを2つ繋いでありbitlineも2本ありましたが片側1本でいいのでは？と思いました。それともNOTゲートのループで保持している内容を入れ替えるのにbitlineが2本必要なのでしょうか。

# SRAM の書き込み



## ■ 書き込み手順

1. ループ左右のどちらか 0 にしたい方のビットラインの電位を下げる
2. ワードラインをアサートして NMOS を ON に
3. インバータの状態を強制的にビットライン側から低電位にする

## ■ NMOS が低電位しか通せないなので、書き込みには 2 本いる



# 質問や感想への回答

- raw hammerやcold boot attackの話はB4のときに勉強しようとして一人ではよく分からなかったのですが、今回の話を聞いてなんとなく理解を深められた気がします。アタックのアプローチは少し異なると思いますが、良ければspectreの話も聞きたいです。
- 学部の際にmeltdownを触り程度で習ったので、その実際が気になります。

- Row Hammerで、何かキャッシュを回避して特定の行を過剰に集中してアクセスするとありましたが、どうやってキャッシュに乗らないようにしてアクセスさせるのでしょうか？

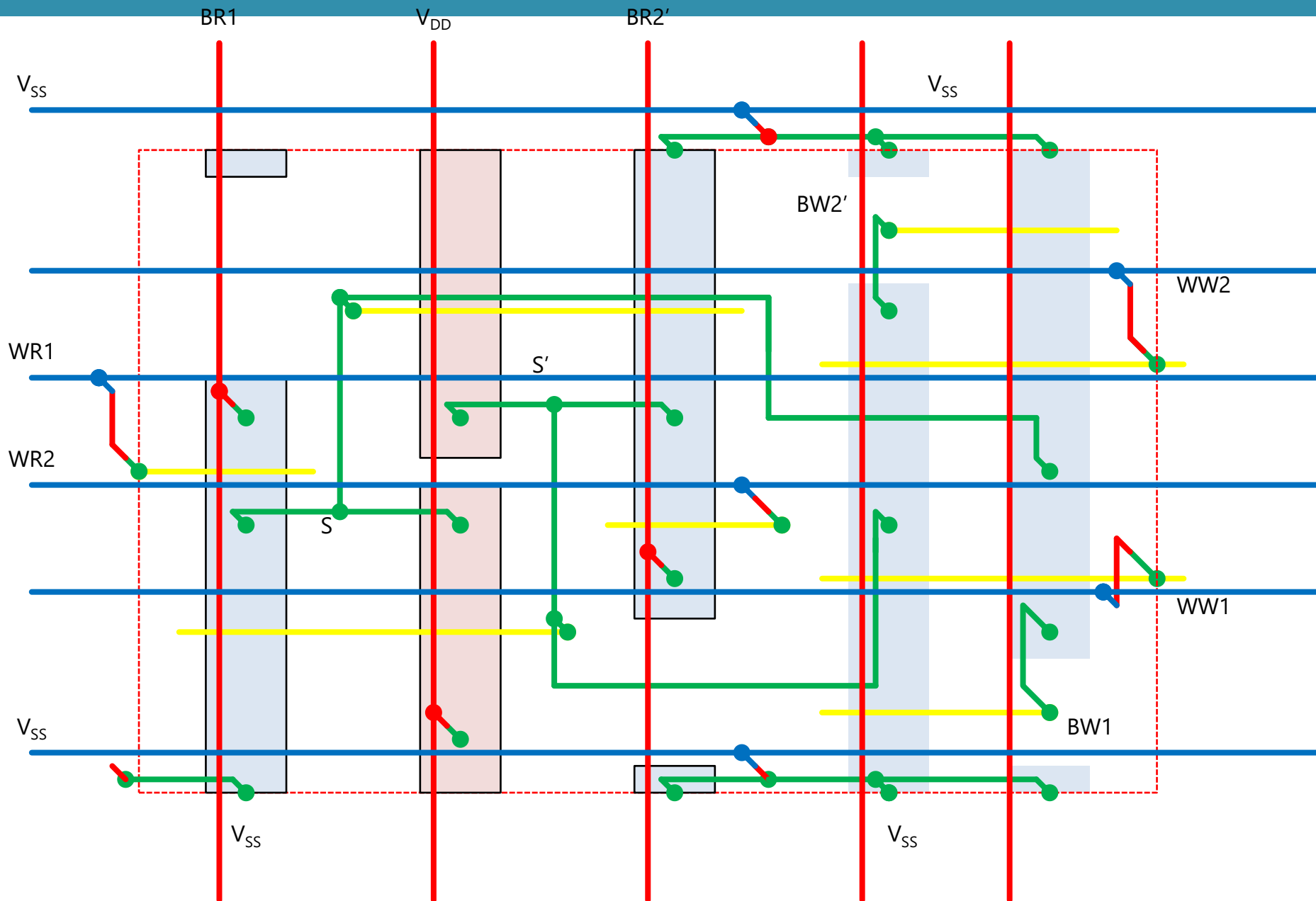
# 質問や感想への回答

- ジム・ケラーがIntelを退職するというニュースを見ました。彼は何者なんですか
- 私の知ってるおやつの人かな、と一瞬疑ったものの、知ってるおやつの人で安心しました。
- 授業でおやつの人が出てくるとは思いませんでした。自分はゲームのタイムアタック動画をよくみるのですが、最近はプレイヤーキャラに特定の動作をさせることで内部のメモリを直接書き換えてエンディング呼び出し命令を直接実行する、みたいな荒技をする人もいます。こういうのは専門家の人から見ても面白かったりするのでしょうか。

# 質問や感想への回答

- 基本的かもしれないが、回路中の「配線」というのは具体的にどういう線を使っているのかが気になった。やっぱり金なんだろうか？ 伸びるし。
- 配線が面積的に支配的になるという表現があったけど線が太いとかではなくて線間の距離がないと色々まずいということだよいのだろうか。
- あとそもそも回路内に線の交差がないか気になってしまった。交差なしで二次元的にめっちゃ頑張って作るのだろうか.....？"

# 3次元的な構造を考えながら，設計



# 質問や感想への回答

- メモリ以外にも物理的な性質を利用した攻撃で既知のものがありますか？
- コンピュータアーキテクチャ界隈のトップカンファレンスはISCAの他にどこがあるのでしょうか
  - ◇ ISCA, MICRO が一番上で、次点が HPCA, ASPLOS

# 質問や感想への回答

- p73で「大容量かつ高速なレジスタやキャッシュは作れない」理由としてアクセス時間を挙げていましたが、他の解説などではコスパが悪い(SRAMの製造コストがDRAMと比べて高い)ためと言われました。両方が理由として正しいのですが、講義で挙げた理由があまり言われないのが気になった。
- 今日の内容とは少しずれた話ですが、RAMを使ったSSDがフラッシュメモリを使ったものよりポピュラーでないのはどうしてですか？

# 質問や感想への回答

- ちなみに、DRAMのリフレッシュって、エネルギー的にはどれくらい大きいんでしょうか？リフレッシュをしなくていい場合を仮定してそれと比べてみると、割と大きかったりするんでしょうか？
- チップの面積を増やせばCPUからメモリへの同時アクセス数を増やせ、プロセッサの性能向上ができるとのことだったが現在それをしない要因は何なの気になった。



# 前回の内容

1. 分岐予測の続き
2. メモリ

# 今日の内容

1. 命令の並列実行の基本
2. データ依存
3. 静的命令スケジューリング
4. 動的命令スケジューリング
  1. 2020年は結局時間が足りなくて、動的命令スケジューリングの話はできなかった

# 命令の並列実行

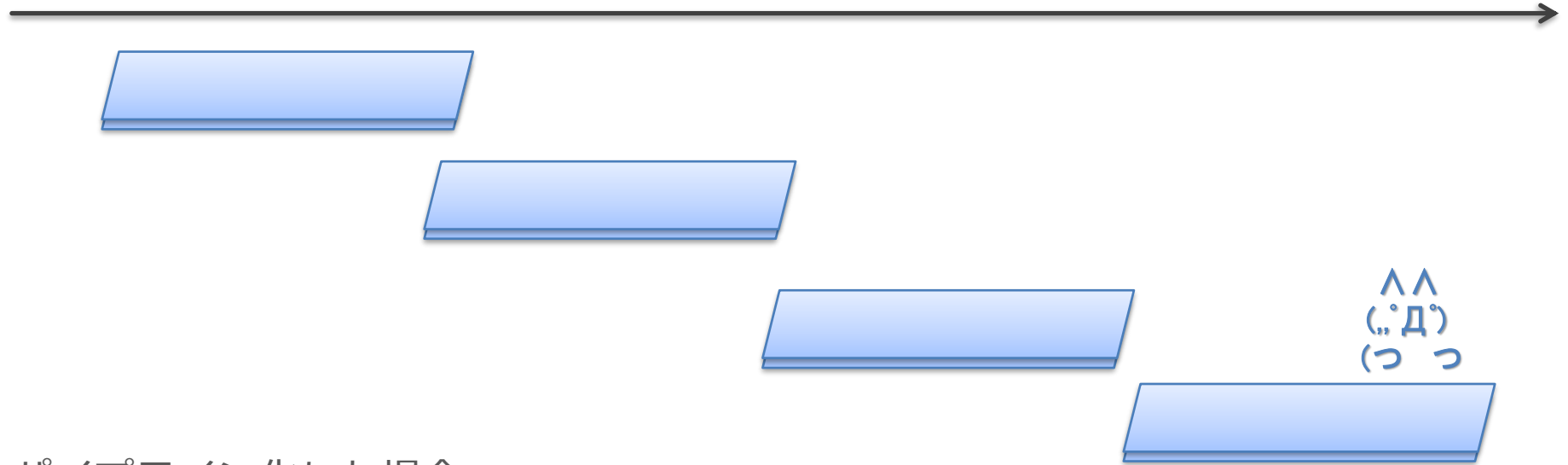
1. スカラ・プロセッサ
2. スーパスカラ・プロセッサ

# スカラ・プロセッサ

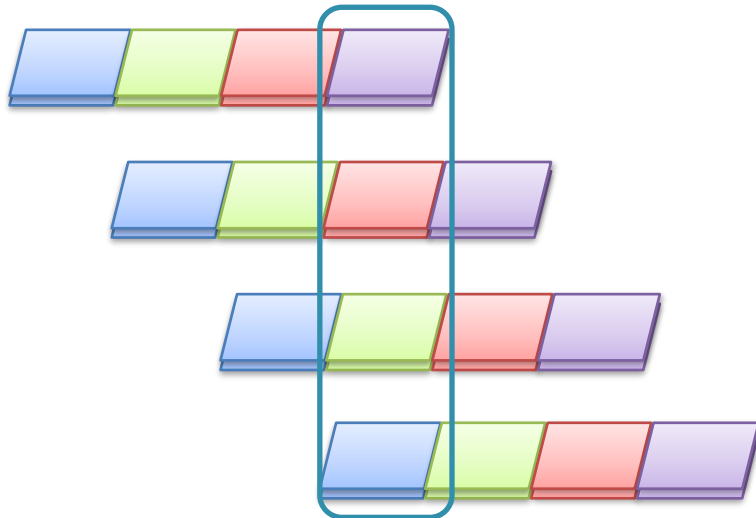
- 1 クロック・サイクルあたりに単一の命令を実行するプロセッサ
- パイプライン化：
  - ◇ 1 つの命令に関わる処理を分割して、毎サイクル並列に実行
  - ◇ パイプライン化されると「単一の命令を実行」になっていない？
    - 1 クロック・サイクルあたりでみると、単一の命令を処理

# パイプライン化

パイプライン化しない場合

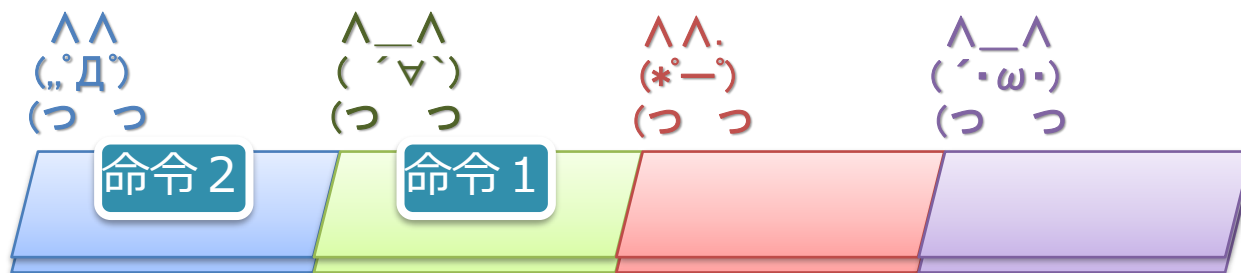


パイプライン化した場合



# パイプライン化による性能向上の限界

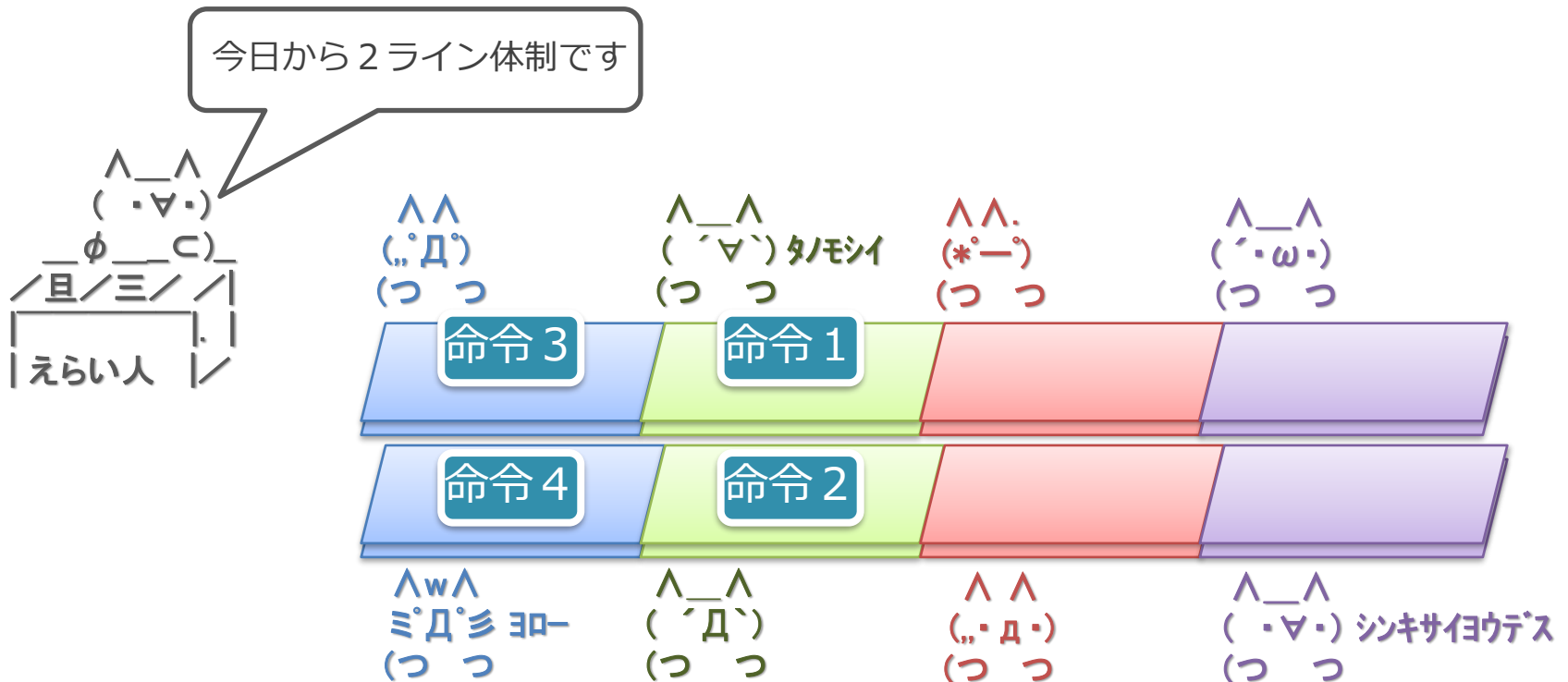
- パイプライン化による性能向上には限界がある
  - ◇ 回路的な理由による周波数向上の限界
    - D-FF の遅延
    - 電力と熱
  - ◇ アーキテクチャ的な理由による実効性能の限界
    - バックエッジによる実効性能の低下
    - (命令スケジュールを行うプロセッサ固有の性能低下もある)



# スーパースカラ・プロセッサ (Superscalar processor)

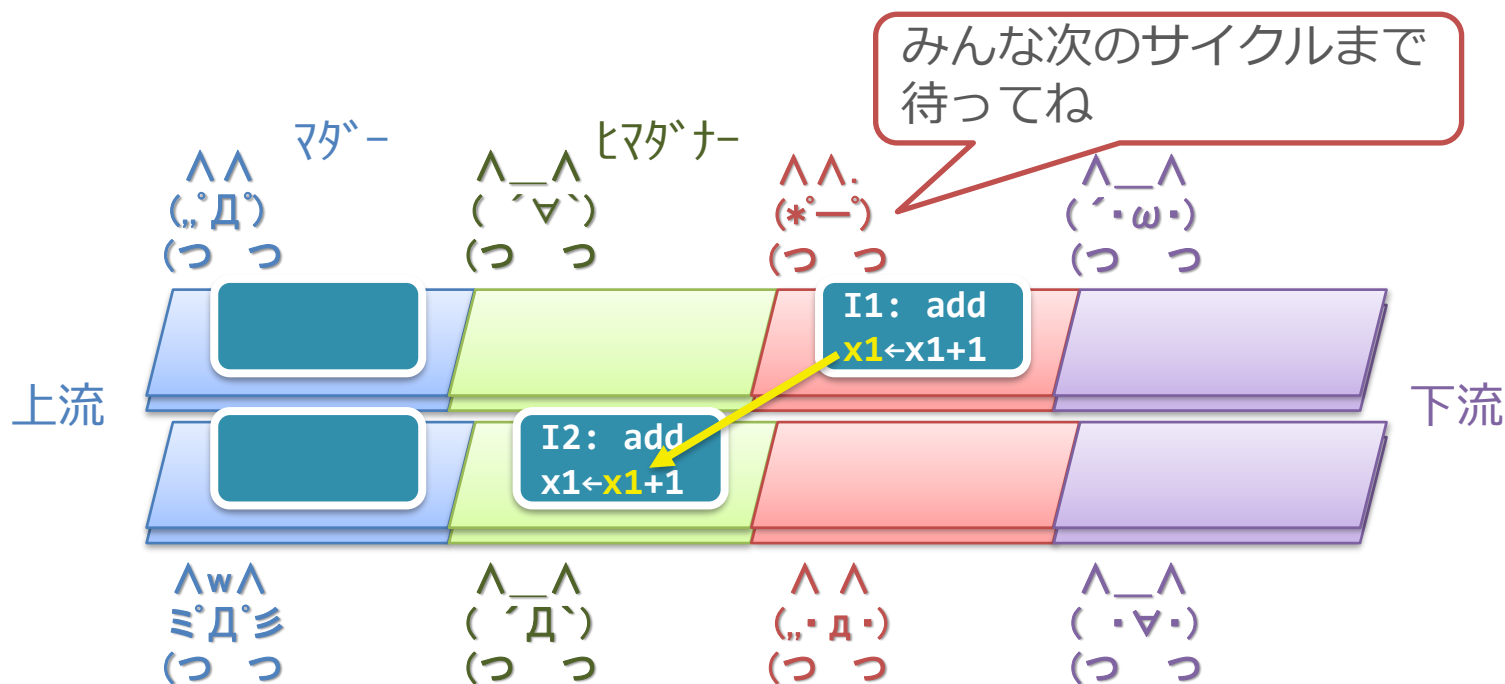
## ■ スーパースカラ・プロセッサ

- ◇ パイプラインや関連する演算器などを複数並べる
- ◇ 複数の命令を並行して処理して性能を向上



# 単純なスーパースカラ・プロセッサの動作

- 同時にフェッチしてきた命令間に依存がない場合は並列に実行
  - ◇ もし依存がある場合は、後続の命令全てを待たせて処理
    - パイプラインの上流側は全てストールさせる
  - ◇ プログラムの意味を保つため
    - 下の図だと I1 と I2 を並列に計算したらおかしくなる

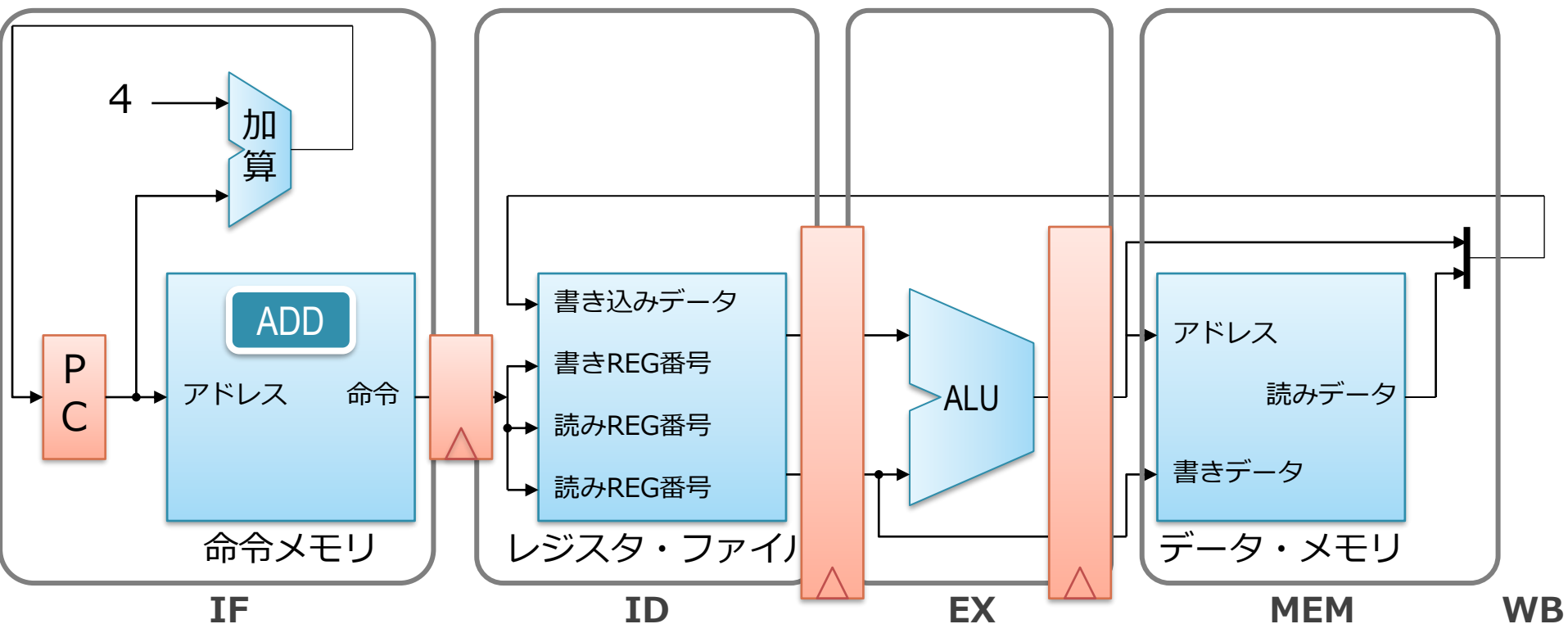




# パイプラインの復習

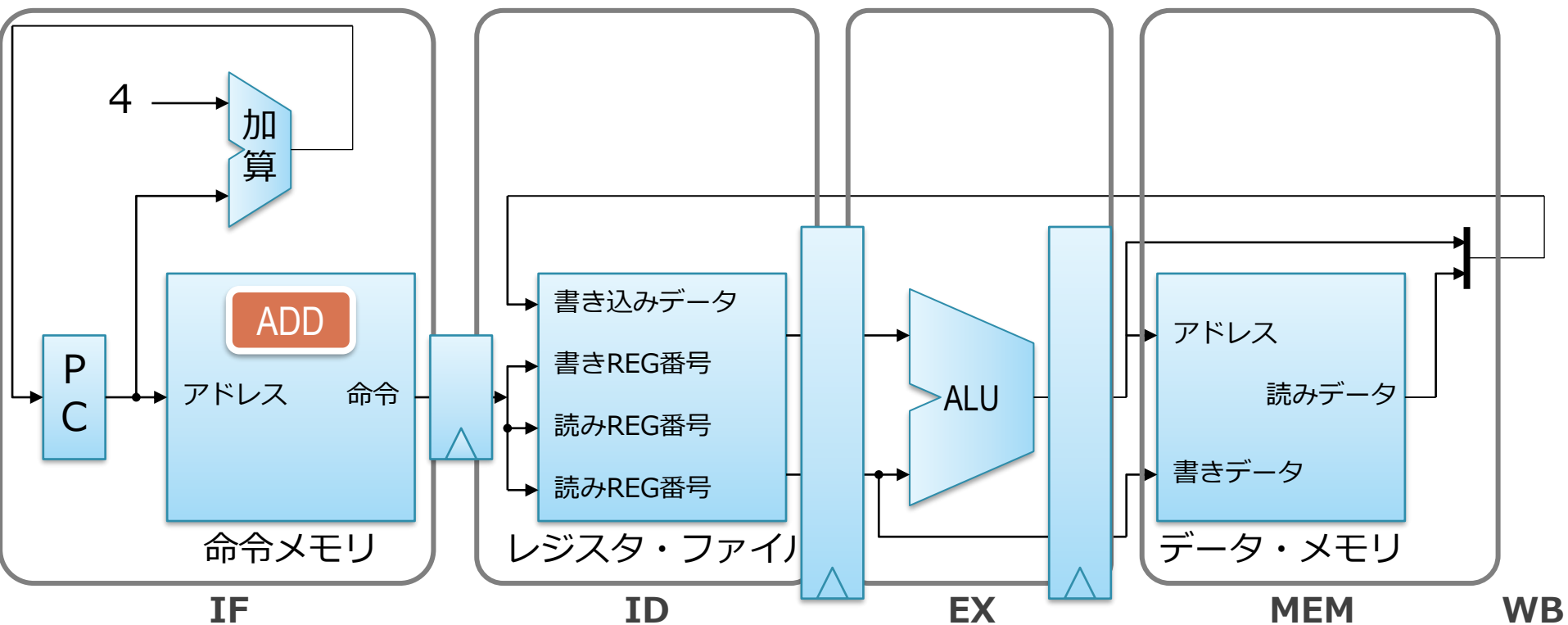
1. IF (**i**nstruction **f**etch)
2. ID (**i**nstruction **d**ecode)
3. EX (**e**xecution)
4. MEM (**m**emory)
5. WB (**w**rite **b**ack)

# パイプライン化されたプロセッサのブロック図



- ◇ 各ステージの間に, D-FF (オレンジの四角) を入れる
  - WB の書き込みについては, レジスタ・ファイル自体がクロックに同期して書き込みが行われるので D-FF は不要
- ◇ 各ステージの処理が早く終わっても, 次のクロックまでは D-FF で信号の伝搬は止まる

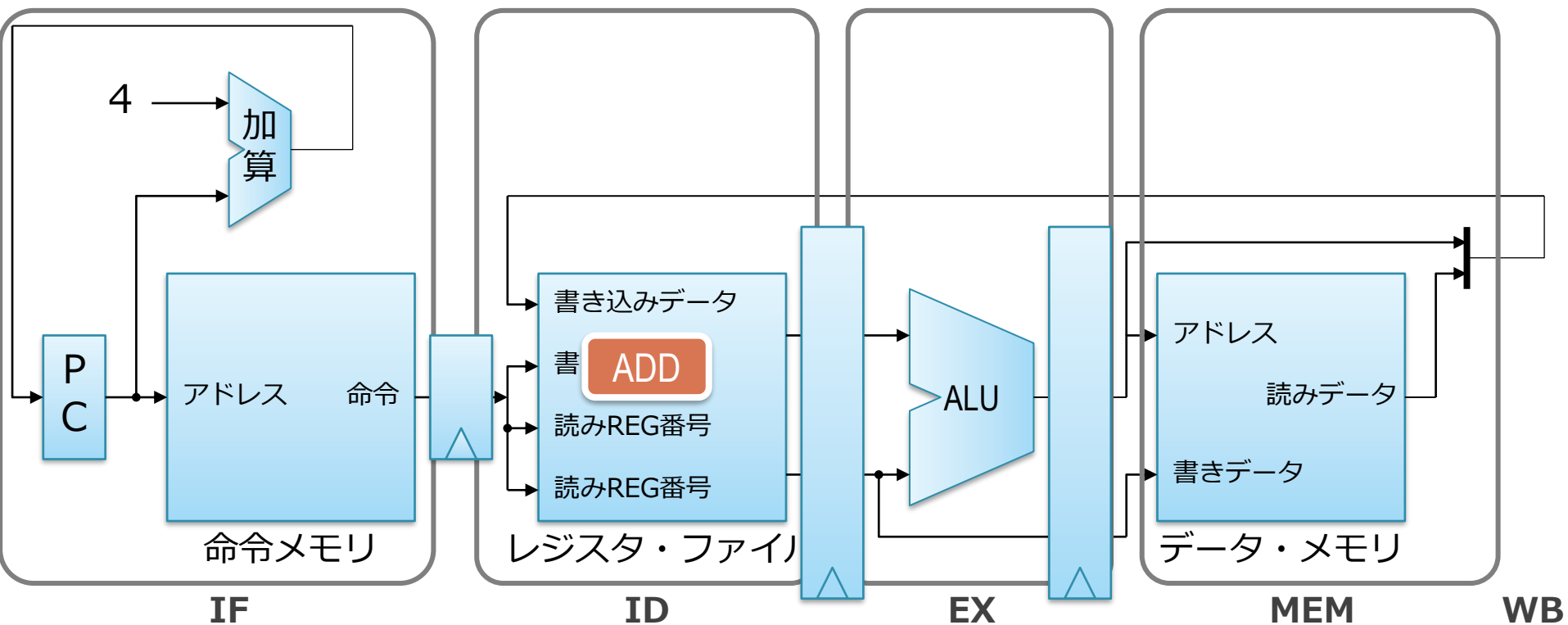
# フェッチ



## ■ IF (instruction fetch)

◇ 命令をメモリから取り出す（フェッチするという）

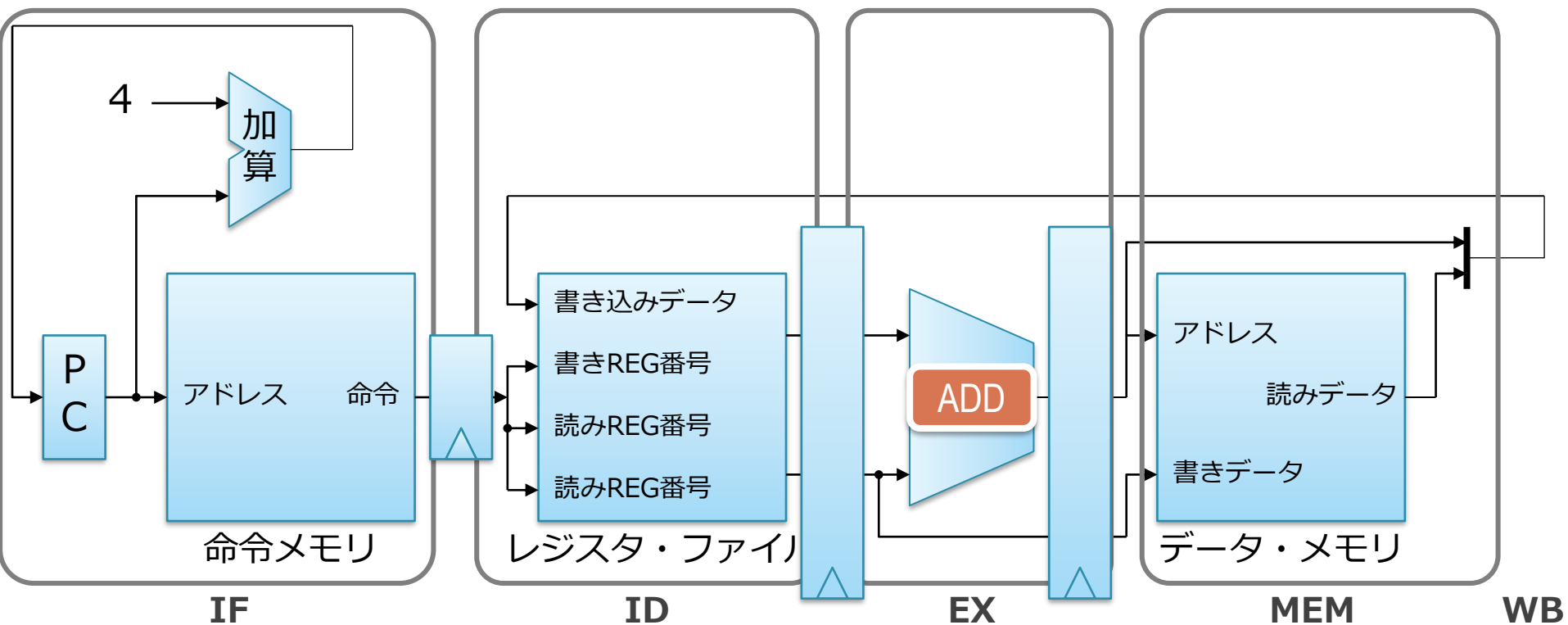
# デコード



## ■ ID (instruction **d**ecode)

- ◇ 取り出した命令の解析（デコードという）をする
- ◇ デコードしてレジスタ番号などを取り出し、レジスタを読み出す

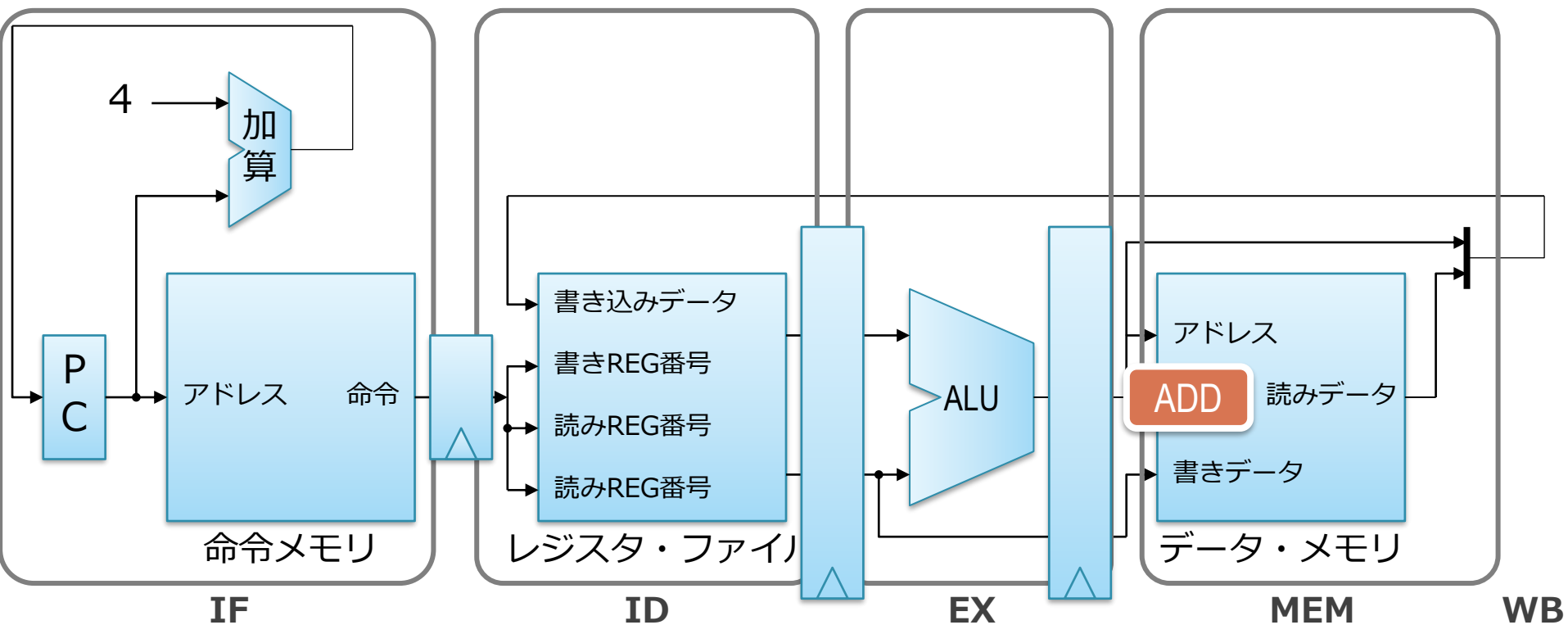
# 実行



## ■ EX (execution)

◇ 演算器で加減算や論理演算などを実行する

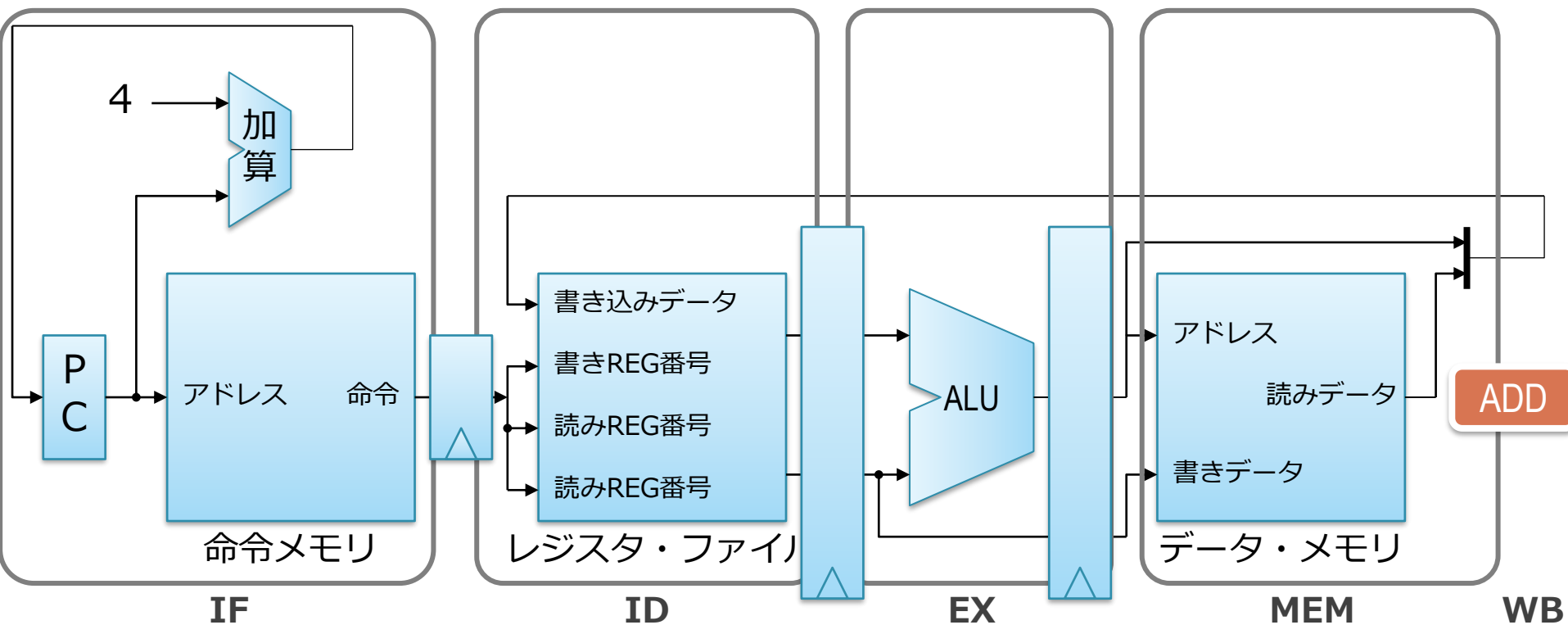
# メモリアクセス



■ MEM (**m**emory)

◇ データメモリにアクセス

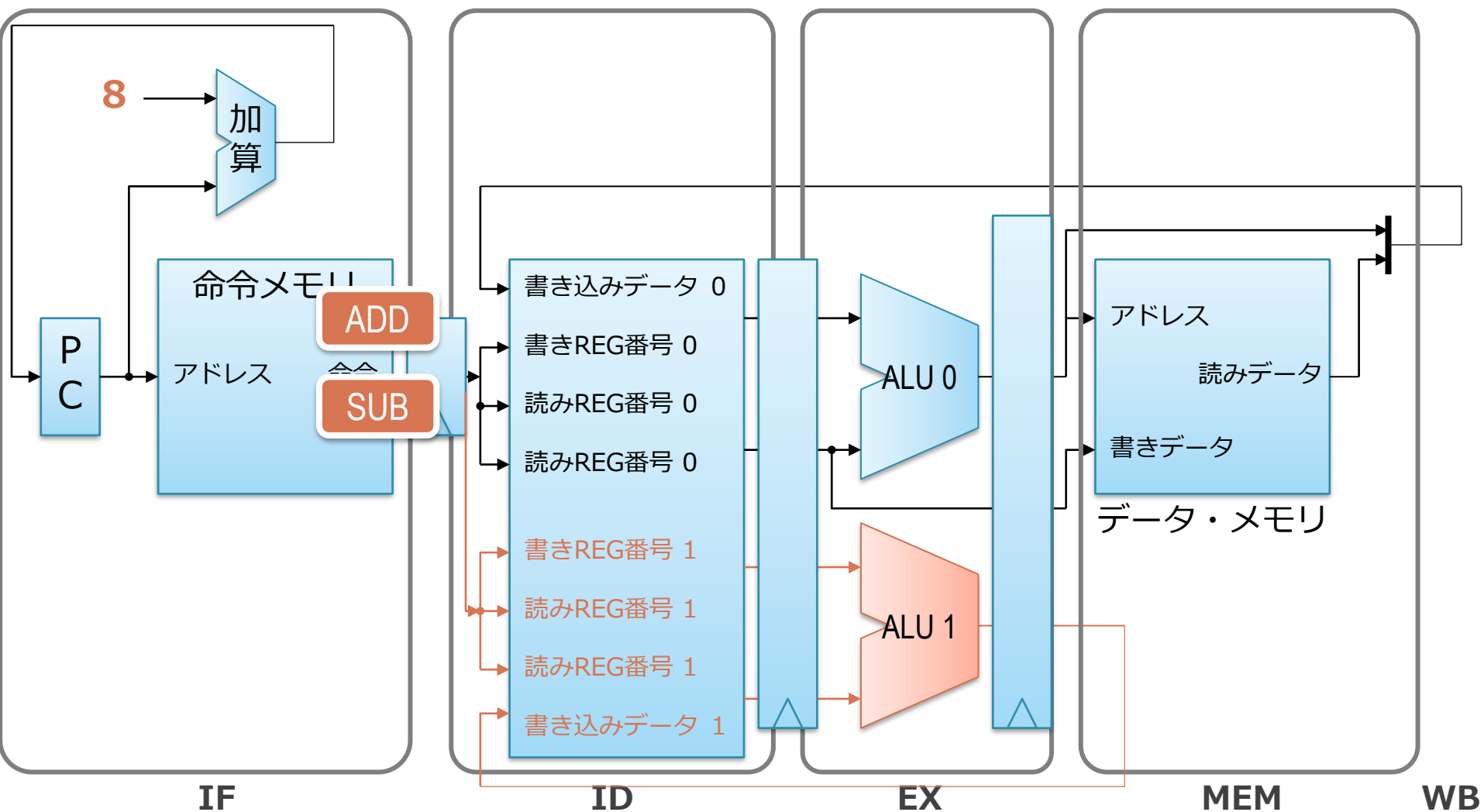
# 書き戻し



■ WB (**w**rite **b**ack)

◇ EX や MEM で得られた値をレジスタに書き戻す

# 単純な 2-way スーパースカラ・プロセッサの例

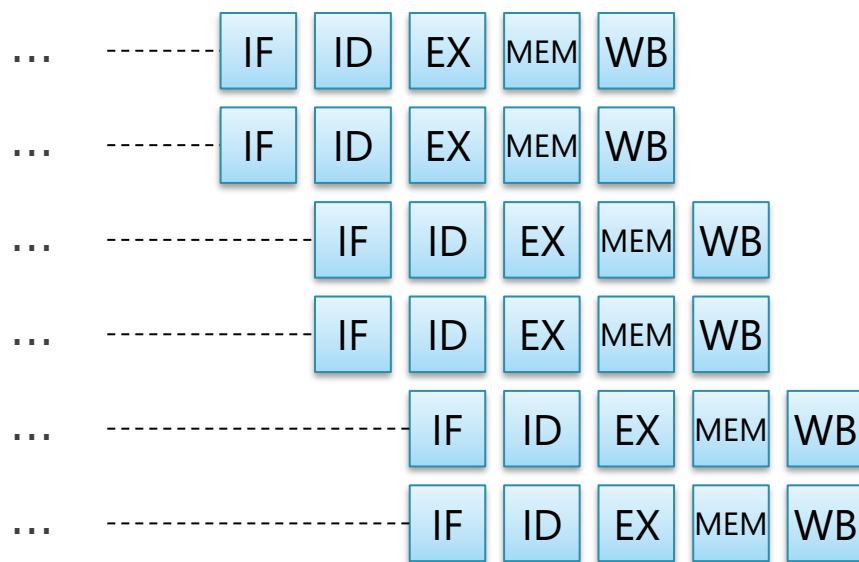


- ◇ フェッチ, レジスタ・アクセス, ALU を 2 命令分に拡張
- ◇ この例では, データ・メモリは 1 つのまま (並列実行に制限がある)



# 単純なスーパースカラによる性能向上

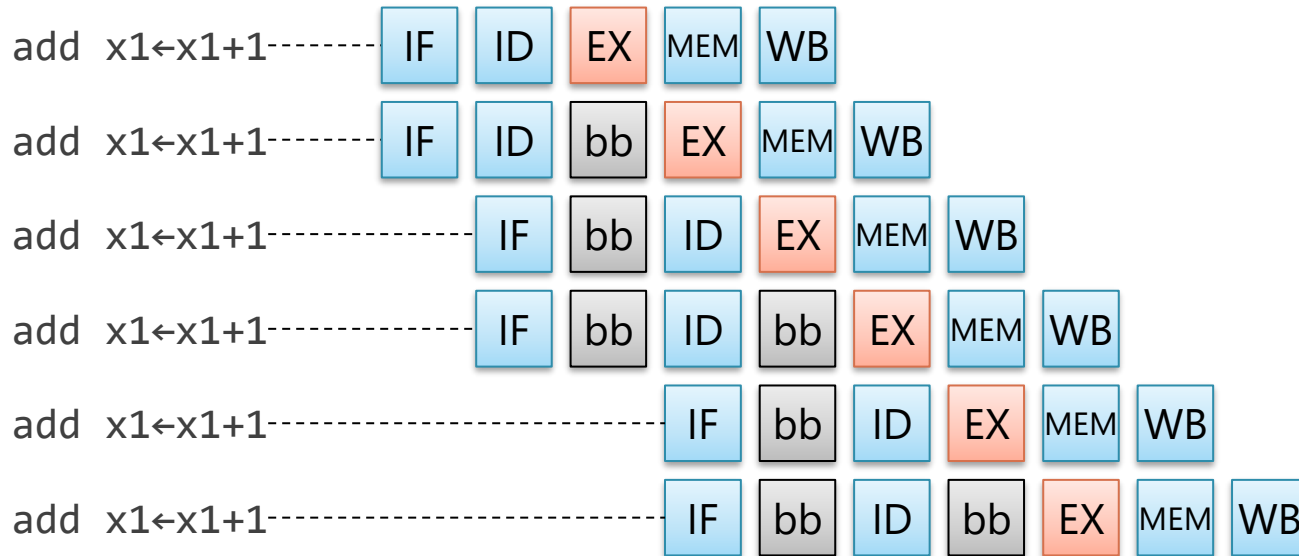
- 理想的には、並列に用意した資源の分だけ性能が向上
  - ◇ 2-way → 性能は2倍
  - ◇ 下の図は、理想的にパイプラインが回った場合



# スーパースカラによる並列実行の制約

- 実際はさまざまな制約があり，そんなに性能はあがらない
  - ◇ 2-way なら数割ぐらいの向上
- 典型的な制約の例：
  1. 同時にフェッチされた命令間に依存がある場合
  2. 構造ハザードが起きる場合
  3. 同時にフェッチされた命令内に分岐があり，他に飛ぶ場合

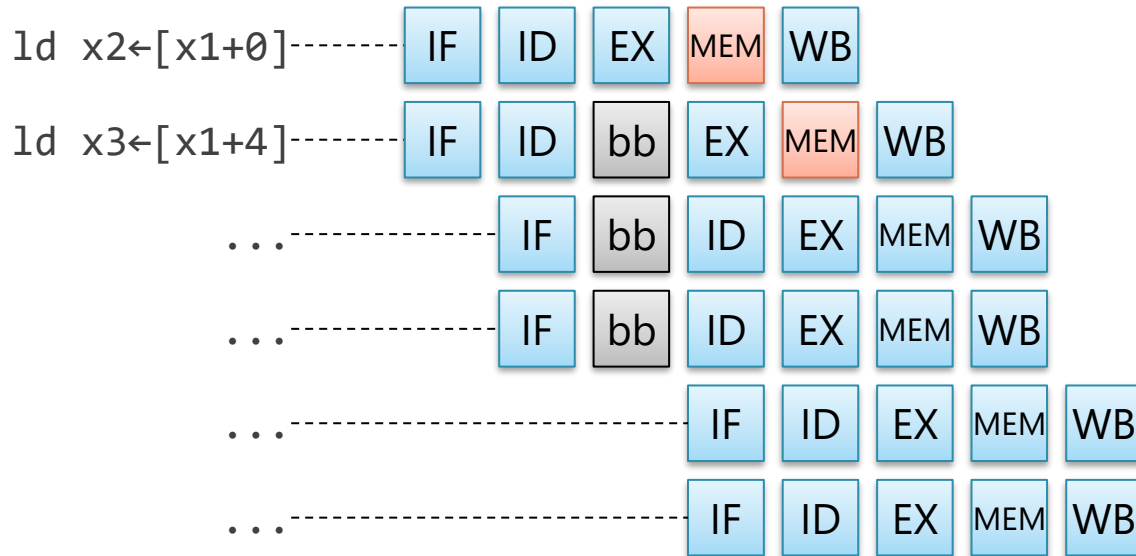
# 1. 同時にフェッチされた命令間に依存がある場合



- 最悪の場合：上記のように全ての命令間に連続に依存があるとき
  - ◇ 演算が逐次的に行われるようにバブルが入る
  - ◇ スカラ・プロセッサから全く性能があがらない

## 2. 構造ハザードが起きる場合

(便宜上メモリステージは EX で行うとしてます)



- 例：先ほどのブロック図のように，メモリは1つしかない場合
  - ◇ ロード命令は1サイクルに1つしか実行できない
  - ◇ 上記のように，ロードが連続するとバブルが入る
- 回路規模が大きい & 使用頻度が低い演算器はパイプライン間で共有されることが多い = 複数同時に来ると止まる
  - ◇ 乗算器，除算器，超越関数の演算器など

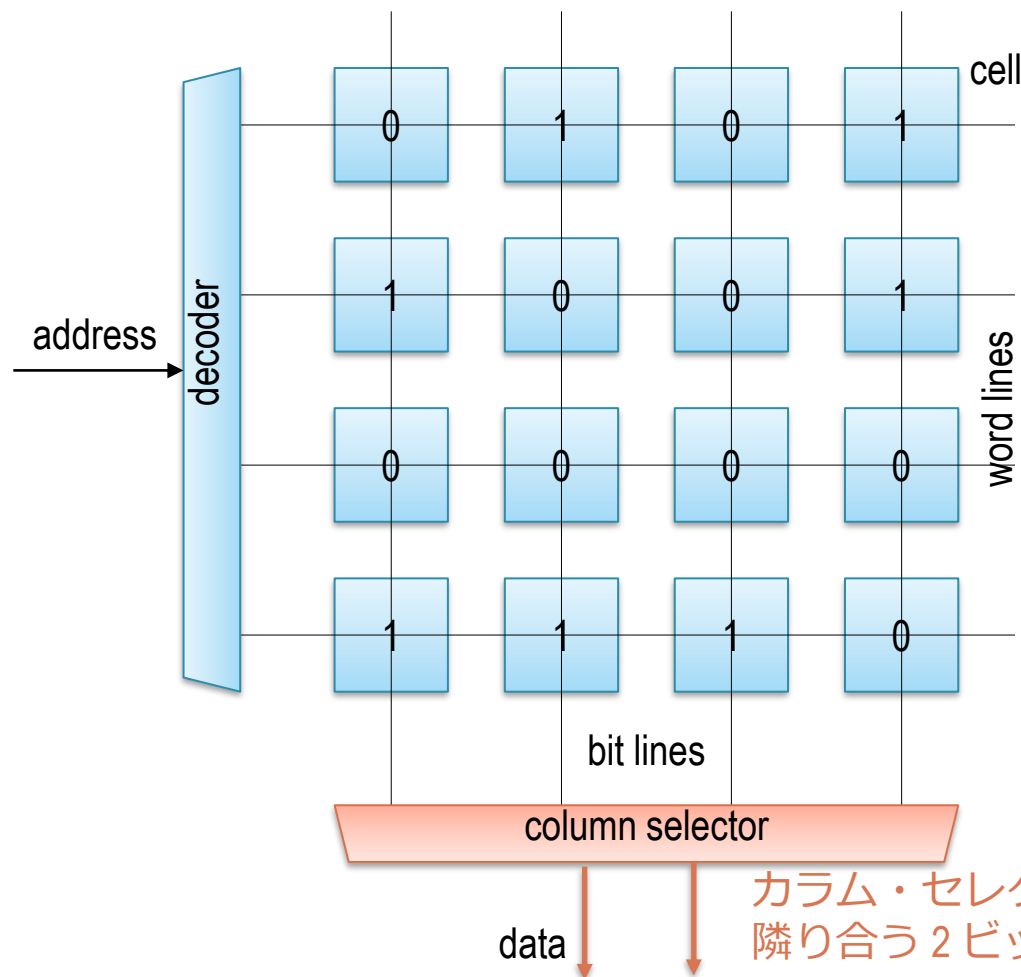
### 3. 同時にフェッチされた命令内に分岐があり、他に飛ぶ場合

- 1 命令目が分岐命令で成立する（と予測された）場合
  - ◇ 命令メモリが 1 ポートしかない場合，そこでフェッチが途切れる
- 例：下記のようなコードの場合

```
0x1000: bne ..., 0x1100
0x1004: ...
...
0x1100: add
```
- PC が今 0x1000 の場合，bne と add をまとめてフェッチしたい
  - ◇ そのためには，0x1000 と 0x1100 の 2 場所を読む必要がある
  - ◇ さらに，分岐予測では 2 個先までアドレスを予測する必要がある

# メモリでは連続箇所（連続した命令）を一気に読むのは一般に簡単

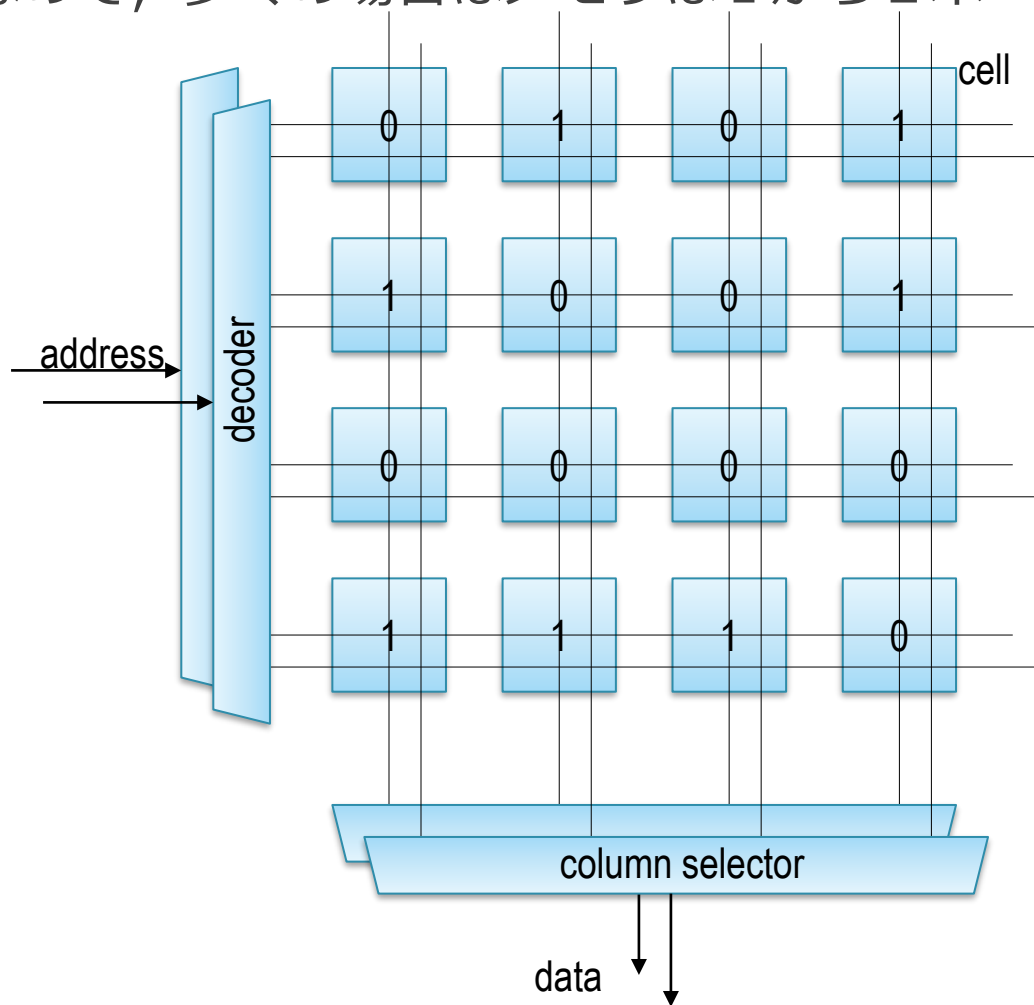
- ◇ もともと行単位で一気に読んでるため
  - カラムセクタでずらせばよい
- ◇ 他に、もっと大きな単位でマルチバンク化という方法が常に適用できる



カラム・セクタを、4ビットから隣り合う2ビットを選ぶように変更

# 任意の複数箇所を同時に読む = マルチポート・メモリが必要に

- ◇ 連続箇所を一気に読むのは簡単だが、独立した2カ所は大変
  - マルチポート・メモリはポート数の2乗で大きくなる
- ◇ なので、多くの場合はメモリは1から2ポート



### 3. 同時にフェッチされた命令内に分岐があり、他に飛ぶ場合

- 分岐をまたいでフェッチをしたい場合
  - ◇ 分岐の飛び元と飛び先の、2箇所をメモリから同時に読む必要がある
    - マルチポート・メモリが必要で回路の増大を招く
  - ◇ さらに、分岐予測では2個先までアドレスを予測する必要がある
    - 出来なくはないが、これもまた回路の増大を招く
- 実際には分岐にあたるとそこでフェッチを止めるのが普通



# 単純なスーパースカラによる並列実行のまとめ

- これまでに説明したような単純なスーパースカラではあまり大きな性能向上が期待できない
  - ◇ 2-way なら数割ぐらいの向上
- 同時実行幅を増やしていても、何かの制約ですぐ止まる
  - ◇  $n$  命令のうち 1 つでもひっかかってたらダメ

# 同時実行幅を増やしていても、何かの制約ですぐ止まる

## ■ どうするか？

### ◇ 1. 構造ハザード

→ ユニットを増やす

### ◇ 2. データ依存

→ **命令スケジューリング（後述）**

### ◇ 3. 分岐をまたぐ場合

→ 上に比べればあまり影響がないので放置

□ 分岐命令は4命令に1回ぐらいの出現なので、4並列ぐらいまでは顕在化しにくい

# 余談：「スーパスカラ・プロセッサ」という言葉

- 広義の「スーパスカラ・プロセッサ」
  - ◇ パイプラインや演算器を複数備え、複数命令を同時実行できるもの
- 単に「スーパスカラ・プロセッサ」と書いた場合、  
後述する「out-of-order 実行を行うスーパスカラ・プロセッサ」の意味でも使われることがある

# 今日の内容

1. 命令の並列実行
- 2. データ依存**
3. 静的命令スケジューリング
4. 動的命令スケジューリング

# 命令間の依存関係

## ■ 命令のスケジューリング

- ◇ プログラムの意味を変えずに、命令の実行順を並び変えること
- ◇ これによって並列に実行できる命令を増やす

## ■ プログラムの意味が変わらない = 依存関係をくずさない

- ◇ スケジューリングの背景として命令間の依存関係を整理しておく

# 命令間の依存関係

1. 制御依存
2. データ依存
  1. 真の依存
  2. 偽の依存

- 分岐とその後ろにある命令間の依存
  - ◇ 分岐命令の後ろにある命令は、分岐先がわかるまで実行不能
  - ◇ 分岐先が確定するまでどこを実行すれば良いか不明なため
- 分岐予測による投機実行により、効果的に解決できる

# データ依存

## 1. 真の依存

◇ フロー依存 : RAW (read after write)

## 2. 偽の依存


◇ 逆依存 : WAR (write after read)

◇ 出力依存 : WAW (write after write)



# 真の依存：フロー依存

## RAW (read after write)

- 文字通り, 同じレジスタを「書いた後に読む」際の依存
  - ◇ 「真の依存」, 「フロー依存」, 「RAW」は呼び方が違うだけでおなじものを指している
  - ◇ 一般に「データの依存関係」と言われたら思い浮かべるもの
- 真の依存の例：I1 が終わらないと I2 は実行できない  
I1: add **x1** ← x2 + 1  
  
I2: add x3 ← **x1** + 1

# 偽の依存 1 : 逆依存

## WAR (write after read)

- 同じレジスタを「読んだ後に書く」

- ◇ 真の依存（書いた後に読む）と方向が逆

- 逆依存の例：

I1: add x2 ← x1 + 1



I2: add x1 ← x3 + 1

- I1 と I2 の間には真の依存は存在しない

- ◇ 「←」の右の入力部分だけみると、順番を入れ替えても問題ない

- ◇ しかし、もしスケジューリングして I2 を先にやると x1 が破壊されてしまう

# 偽の依存 2 : 出力依存

## WAW (write after write)

- 同じレジスタを「書いた後に書く」

- 出力依存の例 :

I1: add **x1** ← x2 + 1



I2: add **x1** ← x3 + 1

- 逆依存と同様に, I1 と I2 の間には真の依存は存在しない
  - ◇ 「←」の右辺にある入力部分だけをみると, 順番を入れ替えても問題ない
  - ◇ しかし, スケジュールして I2 を先にやると I1 により x1 が破壊されてしまう

# 真の依存と偽の依存

- 真の依存は原理的に取り除きようがない
- 偽の依存は有限のレジスタを使い回すことによって発生する
  - ◇ いろいろ取り除きようがある

# 偽の依存の解消の例

- たとえばレジスタがたくさんあれば，事前に取り除ける

- ◇ 逆依存

I1: add x2←x1+1  
I2: add x1←x3+1



I1: add x2←x1+1  
I2: add x4←x3+1

- ◇ 出力依存

I1: add x1←x2+1  
I2: add x1←x3+1




I1: add x1←x2+1  
I2: add x4←x3+1

- 実際にはレジスタは無限に大きくできない

- ◇ 記憶回路の容量と速度はトレードオフがある

# 余談：値予測

- 真の依存を超えて実行を行う値予測（value prediction）という手法も研究されている
  - ◇ 依存元命令の演算結果自体を予測して，実行を先に進める
  - ◇ （この講義では基本的には真の依存は超えられないものとして話をします）
- 関数呼び出し時にメモリに退避させたレジスタの値を復帰させる場合などは結構精度高く予測できたりする
  - ◇ 一時期下火だったが，また研究されだした
- I2 は値予測によって予測された x1 の値を使って I1 より先に実行  


```
I1: add x1 ← x2 + 1
I2: add x3 ← x1 + 1
```

# 今日の内容

1. 命令の並列実行
2. データ依存
- 3. 静的命令スケジューリング**
4. 動的命令スケジューリング

# 静的命令スケジューリング

## ■ 静的命令スケジューリング

- ◇ コンパイラにより、並列実行できるように命令を並びかえる方法

## ■ 静的 vs. 動的

- ◇ 静的：

- 事前に並び替えておくので、CPU からみると実行順は変化しない

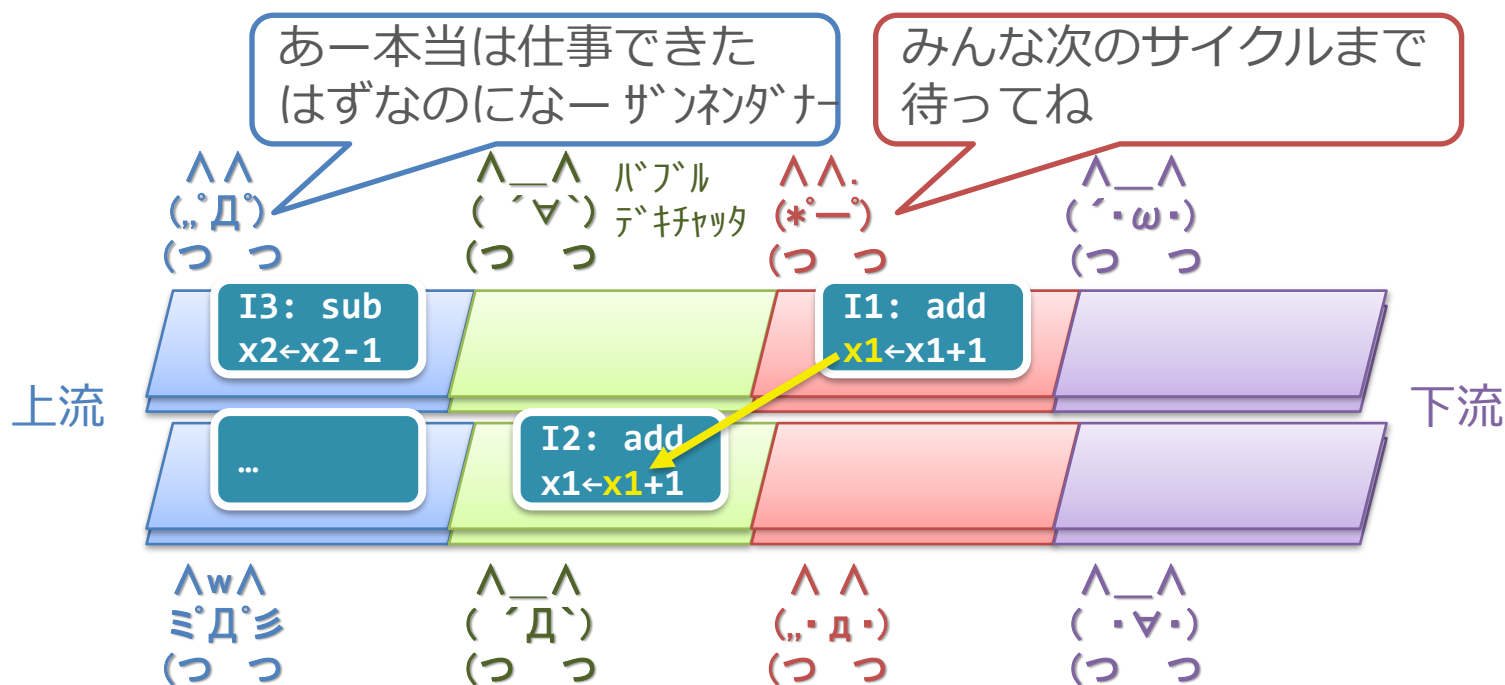
- ◇ 動的：

- CPU が実行時に並び替える



# 単純なスーパースカラでの実行の例

- 下記のコードでは I1 と I2 には真の依存があるが、I3 は無関係
  - ◇ しかし、上流が全部とまるので、I3 も実行できない
  - ◇ I1: add x1←x1+1
  - I2: add x1←x1+1
  - I3: sub x2←x2-1




# 静的スケジューリングによる解決

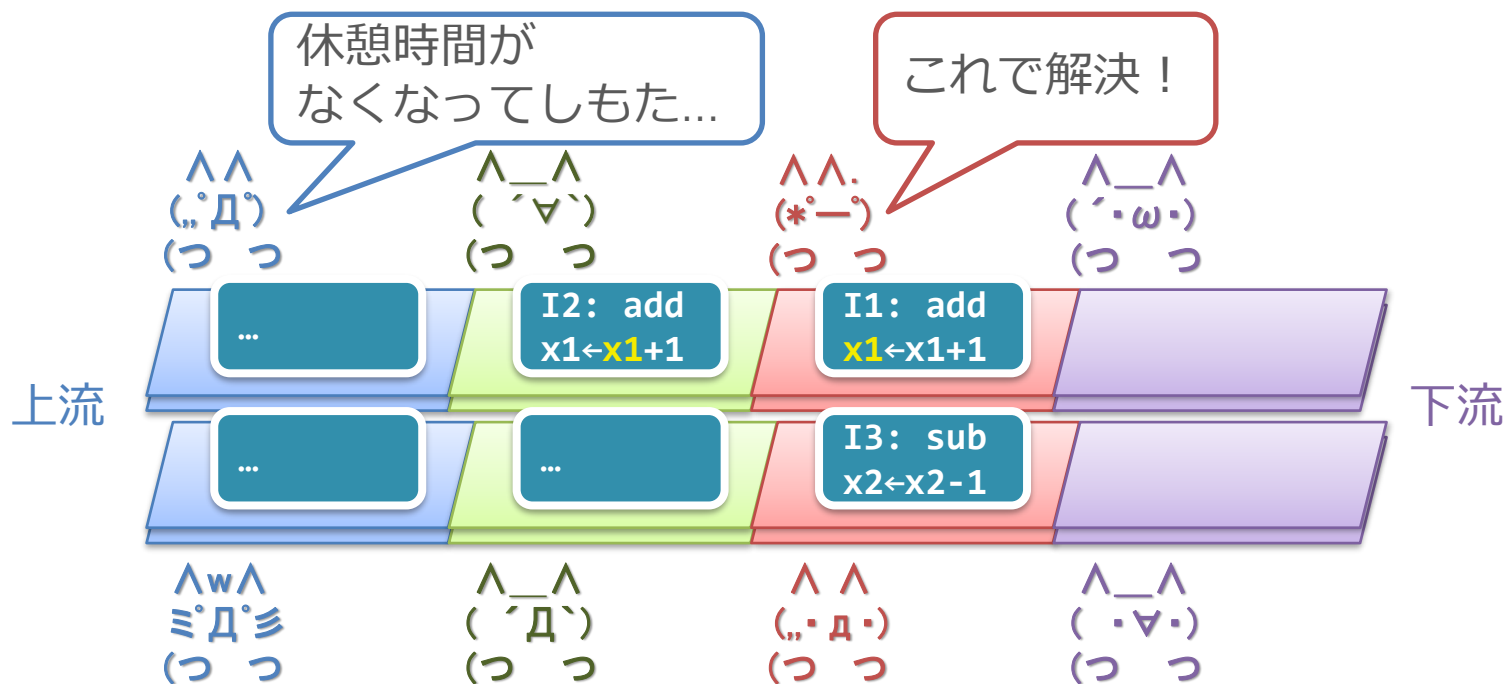
- I2 と I3 を入れ替えておけば、パイプラインはとまらない

◇ I1 と I3 が同時にパイプラインに投入される

◇ I1: add  $x1 \leftarrow x1 + 1$   
I2: add  $x1 \leftarrow x1 + 1$   
I3: sub  $x2 \leftarrow x2 - 1$



I1: add  $x1 \leftarrow x1 + 1$   
I3: sub  $x2 \leftarrow x2 - 1$   
I2: add  $x1 \leftarrow x1 + 1$



# VLIW : Very Long Instruction Word

- 静的スケジューリングを前提とした CPU のアーキテクチャ
- 通常の命令相当の操作を複数まとめたものを 1 つの命令とする
  - ◇ 命令セットの仕様として, 1 つの VLIW 命令内では依存関係を持たないようにする
  - ◇ = フェッチした後は必ずそれらは並列実行できる

◇ I1: add x1←x1+1  
I2: add x2←x2+1  
I3: sub x3←x3-1  
通常の命令セット

I1: 

add x1←x1+1  
add x2←x2+1  
sub x3←x3-1

VLIW ではこれで 1 命令

# VLIW の利点と問題点

## ■ 利点：

- ◇ ハードウェアがすごく簡単

- スーパースカラでは依存があったら止める機構があった

- VLIW では仕様として命令内に依存は発生しないので，不要

## ■ 問題点：

1. 性能向上に限界がある

2. 互換性がとりにくい

# 1. 性能がいまいち出ない

- VLIW は静的スケジューリングに全面的に頼っている
  - ◇ しかし、それで出来る並び替えは結構自由度が低い

# 静的スケジューリングが難しい例 1

- 例1：分岐を乗り越えた並び替えは難しい  
(不可能とはいっていない)

- ◇ 3行目のメモリ・アクセスを1行目の位置まで引き上げることは困難
- ◇ うかつにやると例外が起きて落ちる

```
1: i = i + 1
2: if (flag)
3:     a = *ptr; // flag が false の時は ptr は NULL
```

# 静的スケジューリングが難しい例 2

- 例 2 : ポインタ参照の順番を入れ替えるのは難しい  
(不可能とはいっていない)

- ◇ 2 行目と 3 行目のメモリ・アクセスを入れ替えることは困難
- ◇ うかつにやると意味が変わる

```
1: func(STRUCT* s, int* a){  
2:     s->a = 1;  
3:     int b = *a;  // a は s->a を指してる可能性がある
```

# 余談：C 言語などでのポインタ経由アクセス

- 以下では, `a` と `c` のために2回分のロード命令が生成される
  - ◇ 間にグローバル変数へのアクセスが入ると, 一回 `*ptr` をロードしてレジスタに置いた値が使い回せない
  - ◇ オブジェクトへのメンバへのアクセスでも同じことがおきる
  - ◇ ローカル変数に1回コピーしてからアクセスしたほうが速い
- ```
int g = 0;
func(int* ptr){
    int a = (*ptr) + 1;
    int b = (*ptr) + 2; // 最適化されて上のロード結果を使用
    g = 1; // ptr が g を指している可能性がある
    int c = (*ptr) - 1;
```



# 互換性がとりにくい

- 静的に CPU の挙動を仮定して命令をスケジュールする
  - ◇ = その仮定をくずれるとまずい
- 要因：
  1. 並列実行幅が固定されている
  2. 実行タイミングを仮定してスケジュールされている

# 1. 並列実行幅が固定されている

- 仕様として「N 命令相当を 1 つの VLIW 命令 とする」としている
  - ◇ 性能を上げるために N を後から増やそうと思っても増やせない
  - ◇ 既存のコードが動かなくなってしまう

- たとえば N を 2 から 4 にすると互換性がとれない

◇ I1: `add x1←x1+1`  
`add x2←x2+1`

I2: `sub x1←x1-1`  
`sub x2←x2-1`

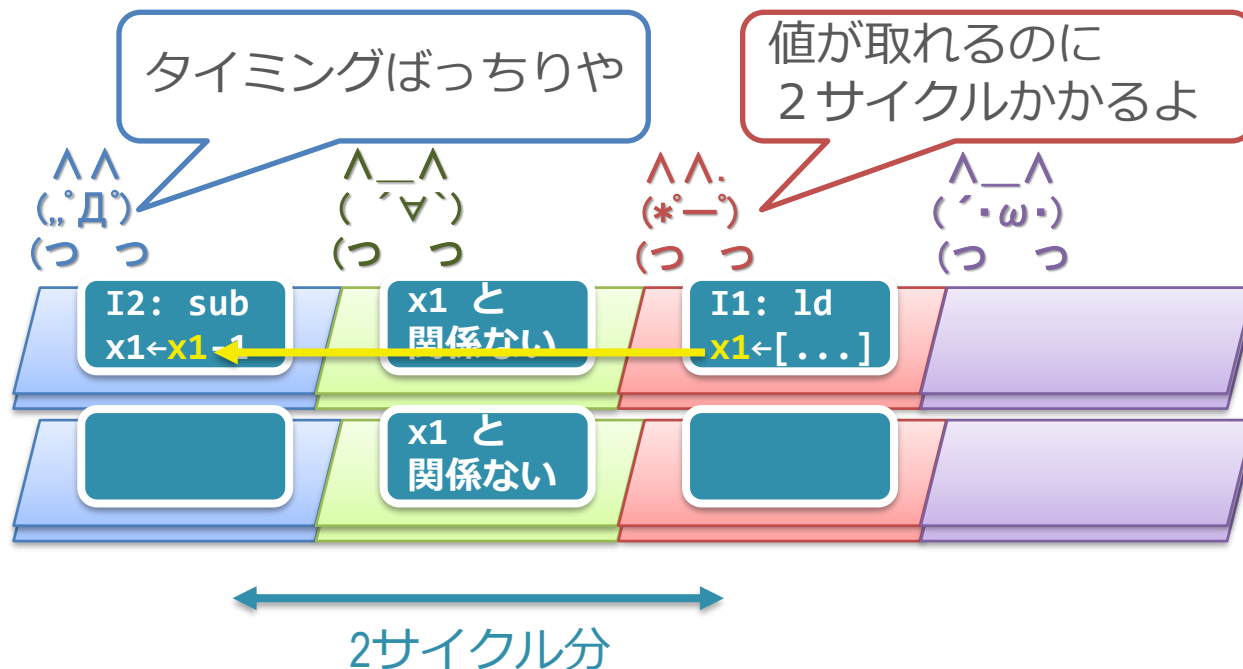
ある VLIW バージョン 1

I1: `add x1←x1+1`  
`add x2←x2+1`  
`sub x1←x1-1`  
`sub x2←x2-1`

ある VLIW バージョン 2  
そのまま実行すると仕様違反

## 2.実行タイミングを仮定してスケジュールされている

- 複数サイクルかかる命令は、それに合わせてスケジュールされる
  - ◇ 「M サイクル後に結果が使用できる」前提でパイプラインが止まらないように事前に命令が並べてある
- 以下では、I1: ld の値が使えるタイミングで I2 が実行できるよう両者を離してある

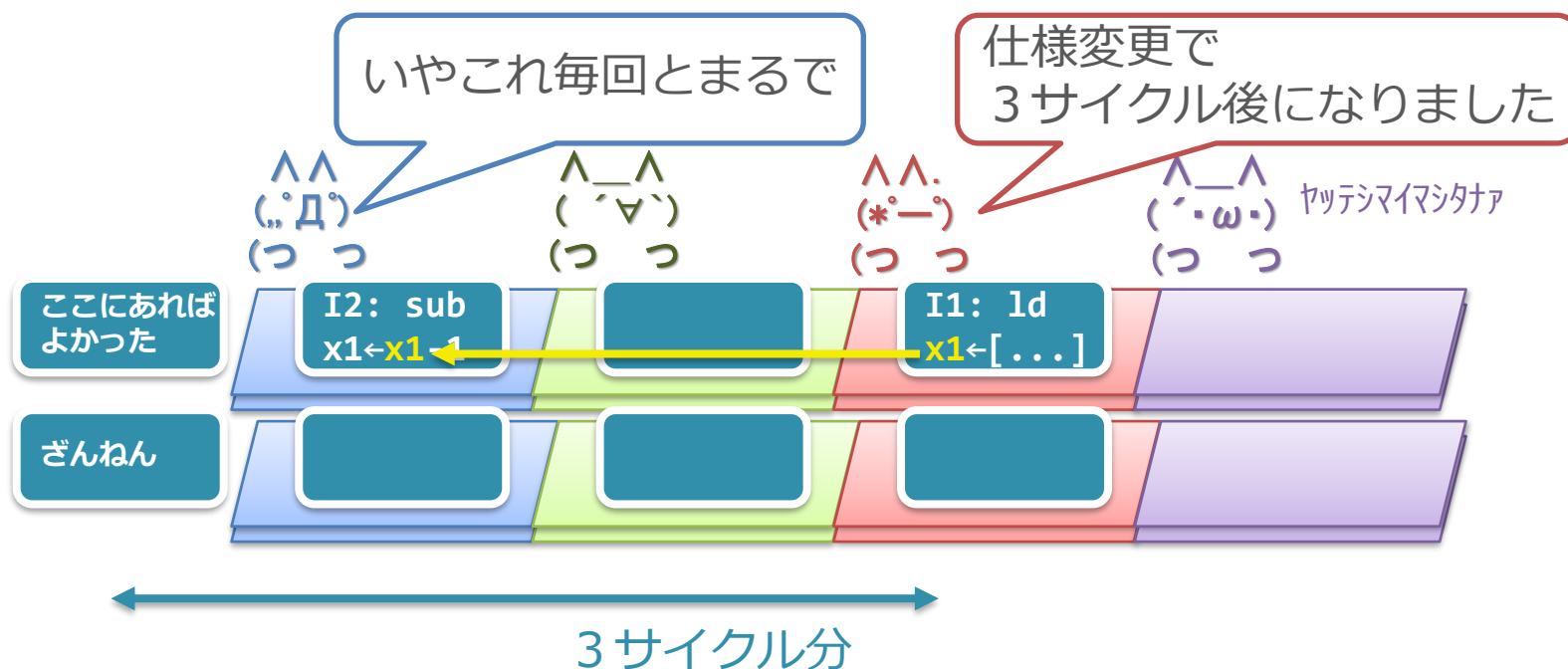


## 2. 実行タイミングを仮定してスケジュールされている

- M を変化させにくい
  1. 短くなる = 既存のコードは恩恵を受けられない
  2. 長くなる = 毎回バブルが発生して性能ががたおちする
    - 想定したタイミングに入力が揃わない
- たとえば、次世代の以下のような CPU 作る場合を考える：
  1. 乗算器を改良してレイテンシが短くなった
  2. キャッシュを倍にしてヒット率をあげたがレイテンシが多少伸びた

# キャッシュのレイテンシが伸びた場合

- Id のレイテンシが 2 から 3 に変更となった場合
  - ◇ 2 サイクルにジャストで合わせて I2 をスケジューリングしておくと、毎回バブルが発生することに



# 実行タイミングを仮定してスケジュールされている ことの他の問題

- そもそも実行時にレイテンシが動的に変化する場合是对応困難
  - ◇ キャッシュのヒットとミスが場合によってかわるようなロード
- コンパイラではあらかじめヒットかミスを仮定してスケジュール
  - ◇ プロファイラで事前に特性をとって、それに基づくことである程度緩和はできる

# VLIW の例 : Intel Itanium

- インテルと HP で作った VLIW プロセッサ
  - ◇ 2000 年代前半ぐらいまでは x86 からこれに移行しようとしていた
  - ◇ EPIC アーキテクチャと言われる命令セットを持つ
- これまで述べたような VLIW の問題を緩和するような機構を色々投入
  - ◇ 命令セットの互換性をとりながら同時実行幅を増やす
  - ◇ 分岐を跨いだロードの移動をハードで支援

# Intel Itanium の性能

- しかし, x86 よりも全然性能がでなかった
  - 1. 静的スケジューリングの限界
  - 2. レイテンシを仮定したコード
  - 3. クロックが上げられなかった
    - 1. 2. に関連して, キャッシュ・アクセスのステージ数を増やしてクロックを上げることができない
    - 2. VLIW の問題緩和の機構のせいで返って複雑化



# Intel Itanium の末路

1. 当時 32 ビットから 64 ビットへの移行の要求が高まっていた
  - ◇ 主にメモリ使用量を増やすため
    - 32 ビットのアドレスで表せるのは 4GB まで
  - ◇ Itanium はこのための 64 ビット CPU でもあった
2. インテルは互換 CPU の製造開発を許したくなかった
  - ◇ しかし既に与えたライセンスは取り消せない
  - ◇ 64 ビット世代で内容を刷新して今度は独占を目指した
3. AMD が独自に x86-64 を策定
  - ◇ Itanium がさっぱり性能でないので、MS が見切りをつけて Windows の x86-64 対応を開始
4. 後追いでインテルも x86-64 の CPU を開発
  - ◇ Itanium は一応製造されているが、2021 年に最終出荷で終了

# VLIW は全くダメなのか？

- 以下のような場面であれば有用
  - ◇ 絶対性能よりも、ハードが小さいこと（電力）の要求が高い
  - ◇ 動作させるソフトウェアが限られている
    - 互換性が問題になりにくい
- 典型的には、組み込み CPU が該当
  - ◇ 簡単なハードでそこそこの性能がほしい時に有用
- CPU を作る学生実験で性能出したい場合なんかでも有望
  - ◇ 実装が簡単 & 少数の課題となるプログラムさえ速ければよい
  - ◇ 人力静的スケジュールで最適化する
    - いろんな仮定をぶちやぶれる

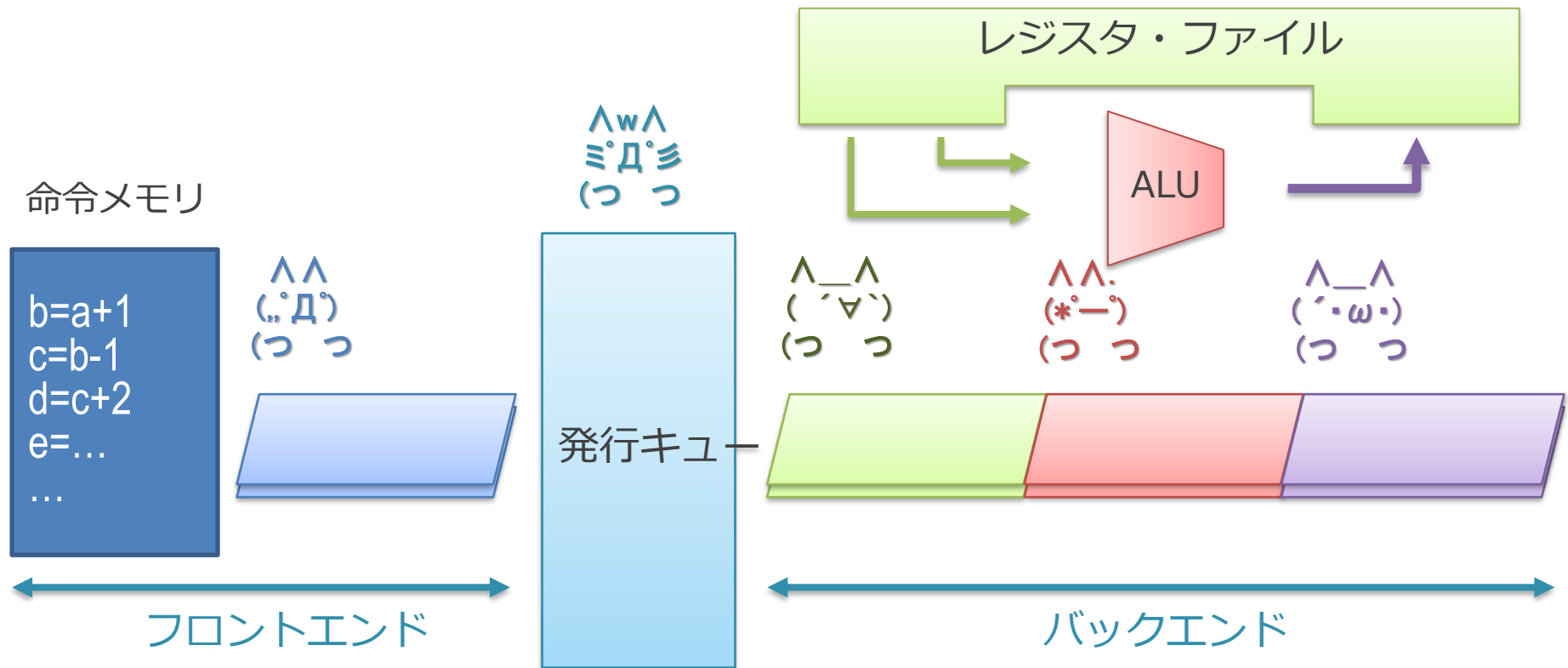
# 今日の内容

1. 命令の並列実行
2. データ依存
3. 静的命令スケジューリングと VLIW
4. 動的命令スケジューリング (のさわり)

# 動的命令スケジューリング

- CPU により, うまく並列実行できるように命令を並びかえる方法
  - ◇ 静的 : 事前に並び替えておくので, CPU からみると変化しない
  - ◇ 動的 : CPU が実行時に並び替える
- スカラ/スーパスカラとは直行した概念
  - ◇ ...ではあるが, 普通は動的スケジューリングを行う CPU はスーパスカラ
  - ◇ スカラで動的スケジューリングをやってもあまり意味がないから
- 現在主流の CPU は, 基本的にみなこのタイプ

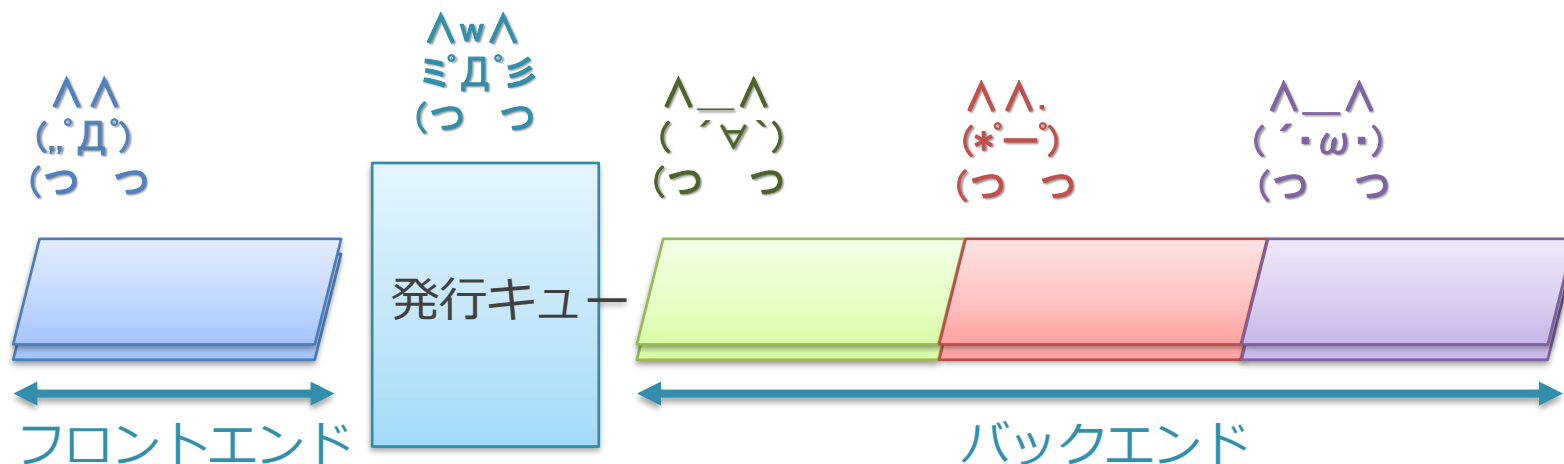
# 動的命令スケジューリングを行う CPU の構造



## ■ 発行キューによって前後に分離された構造を持つ

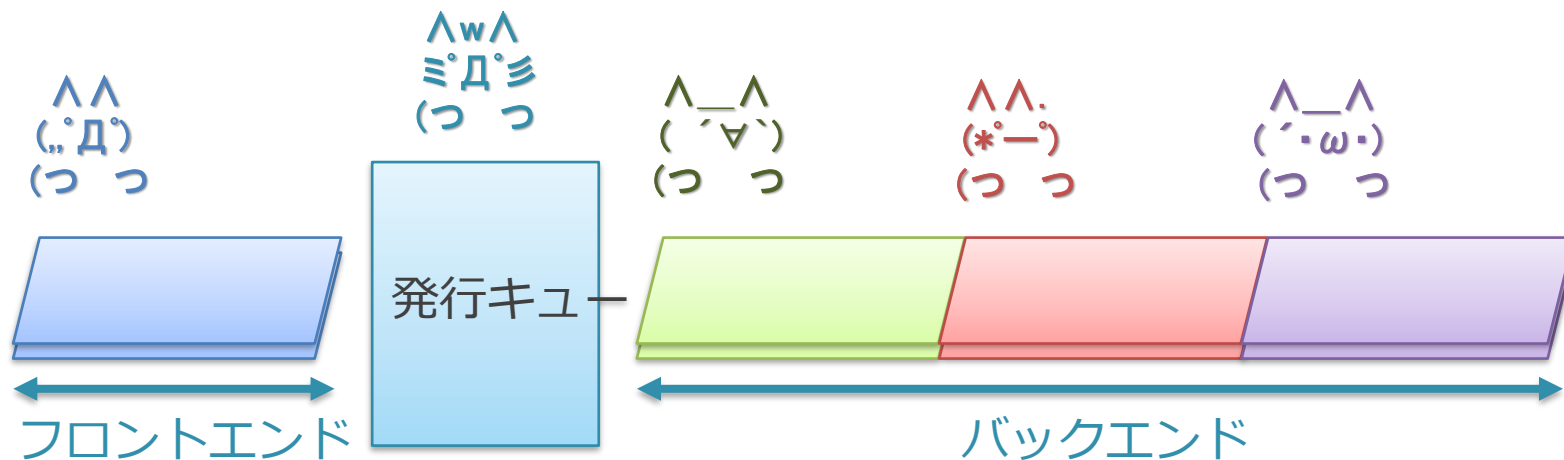
1. フロントエンド : 命令を供給
2. 発行キュー : 命令の待ち合わせ
3. バックエンド : 命令を実行

# 大ざっぱな動作



1. フロントエンドで命令を順にフェッチ
2. 発行キューに投入
3. 実行可能なものから順にバックエンドに命令を送信
4. レジスタを読んで演算器で実行し書き戻す

# 言葉の定義 1



- ◇ ディスパッチ：フロントエンドから発行キューに命令をいれること
- ◇ 発行 (issue)：発行キューからバックエンドに命令を送ること
- ◇ 完了 (complete)：バックエンドで命令の処理が終わること
  
- ◇ 微妙にこのあたりの用語は文献ごとに統一されていないので注意
  - インテルは昔からかたくなにディスパッチと発行を逆の意味で使う





# 今日の内容

1. 命令の並列実行
2. データ依存
3. 静的命令スケジューリングと VLIW
4. 動的命令スケジューリング（のさわり）

# 出欠と感想

- 本日の講義でよくわかったところ, わからなかったところ, 質問, 感想などを書いてください (なんか一言書いてね)
  - ◇ LMS の出席を設定するので, そこにお願いします
  - ◇ パスワード : super
- 意見や内容へのリクエストもあったら書いてください