

# PHPの正規表現と 最長一致

hanawa (a.k.a. id:hnw)  
y at hnw dot jp

第29回PHP勉強会発表資料

# 第29回記念大会

- PHP勉強会おなじみのmagic number
- 次回記念大会は第0x29回（1年後）

# アンケート

- 正規表現に苦手意識がありますか？
- 最長一致って聞いたことありますか？

# クイズ

- 正規表現「 $(a+)([ab][ab])+$ 」を  $aaabbbb$  に対して適用した結果は？
  1.  $aaabb$
  2.  $aaabbbb$
  3. その他

# クイズの答え合わせ

- 正解：(1) `aaabb`  
(`preg_match`, `mb_ereg`)
- (2) `aaabbb` が正解のこともある(`ereg`)

→ 「最長一致」って何？

# 最長一致？

- 実は日本語の「最長一致」には2種類
  - greedy matching  
(PCRE, preg\_match/[mb\\_ereg](#))
  - longest matching (POSIX, ereg)
- greedyも「最長」と翻訳→[ほぼ誤訳](#)

# greedy matching

- greedy matching = 欲張りマッチ
  - 繰り返し表現を最大回数マッチさせる
  - 以降の表現を試して失敗したら、繰り返し回数を減らして以降の表現を試す (バックトラック)
  - それでもダメなら開始位置をずらす

# クイズの解説

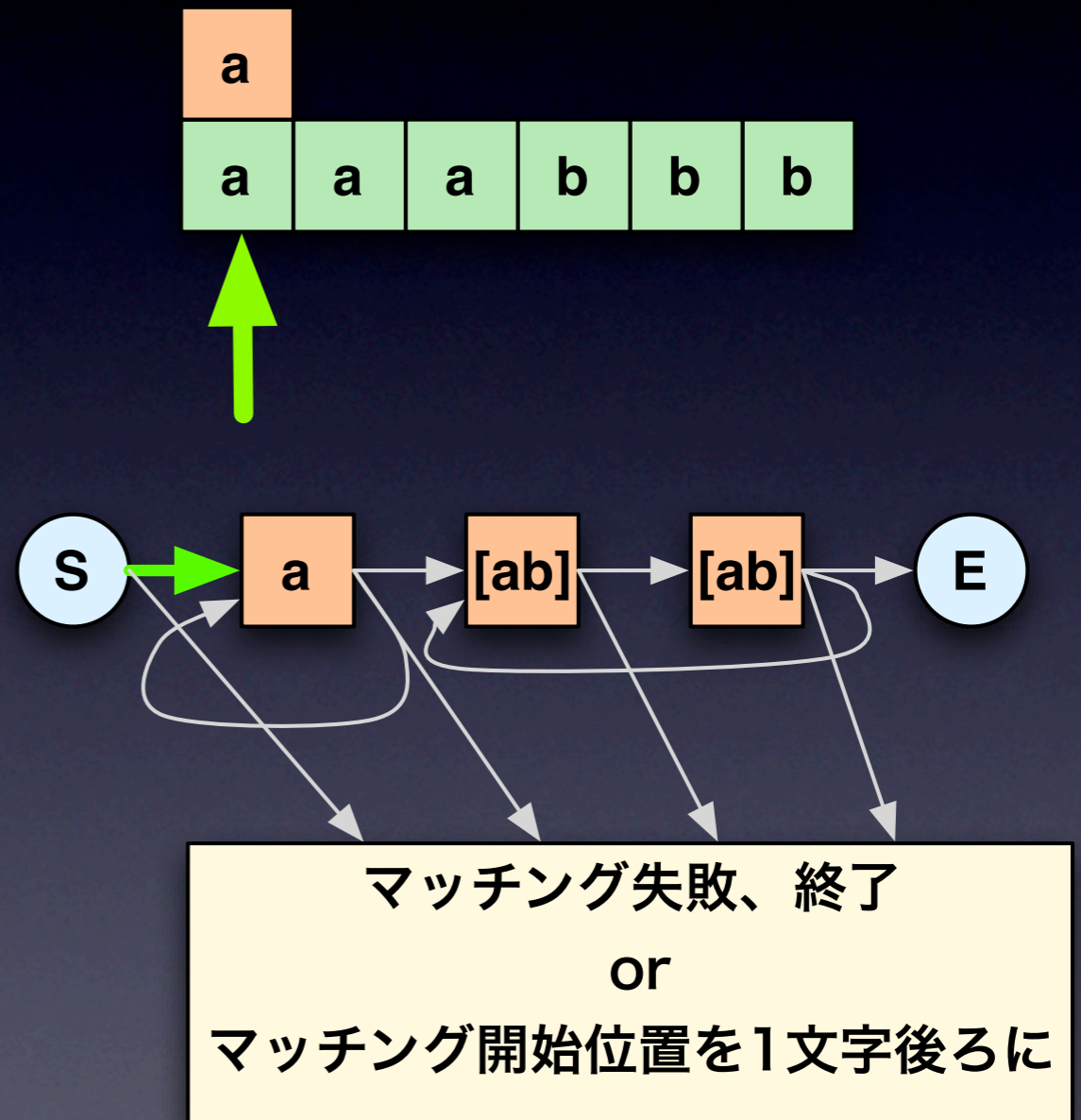
- 「 $(a+)(([ab][ab])^+)$ 」を  $aaabbb$  に対して適用
  - $a^+$ が可能な限り長くマッチ→ $aaa$
  - $bbb$ で $([ab][ab])^+$ を試す→ $bb$ 
    - 正規表現全体が最長とは限らない

# longest matching

- longest matching = 最長一致
  - 全体が最長 → **aaabbb**
  - 先に出現した表現が最長
  - 実装コストが（おそらく）高い
  - 検索コストも？

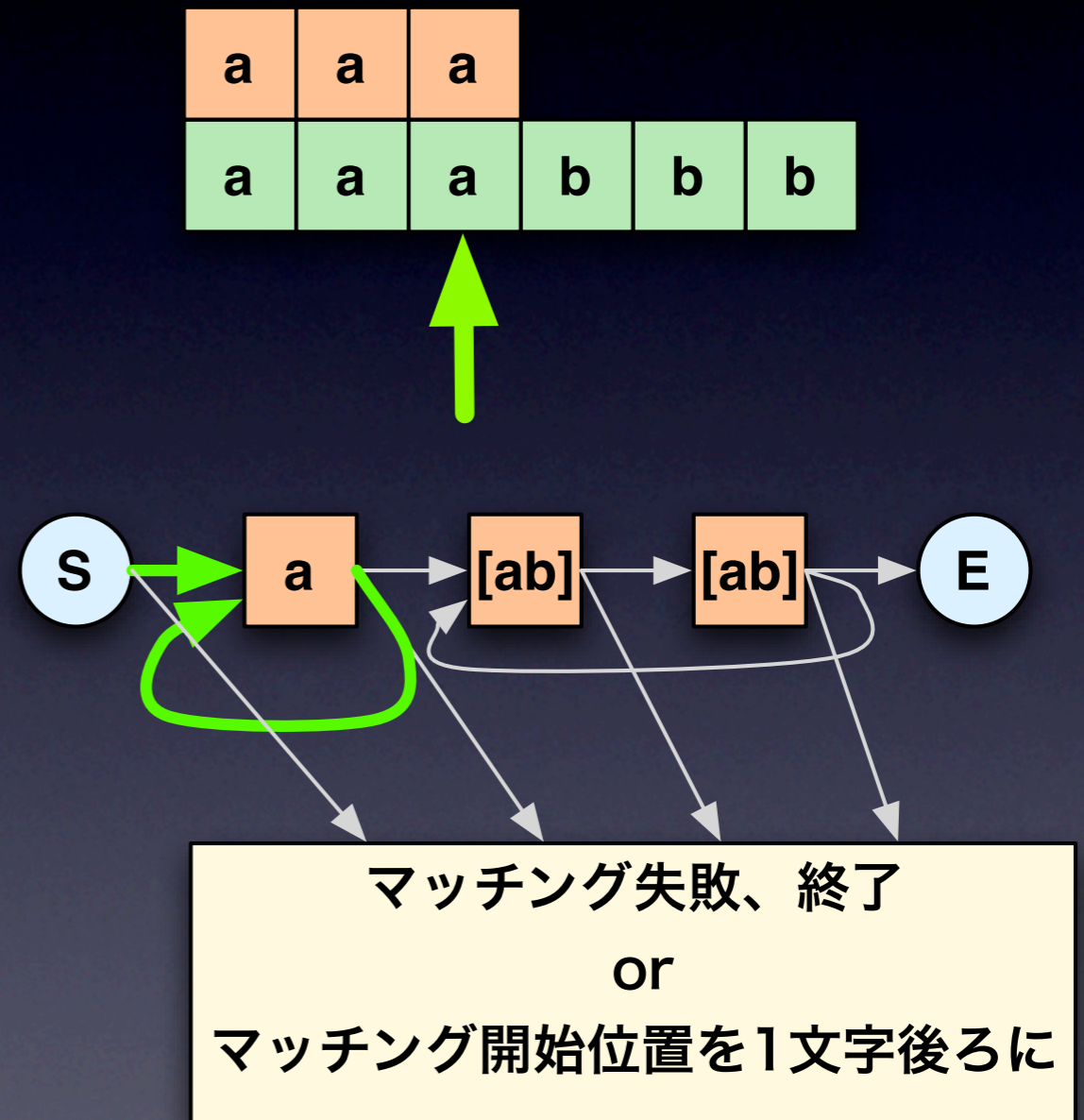
# greedy matching 図解

- 正規表現  
「 $a+([ab][ab])+$ 」
- 文字列 `aaabbb`



# greedy matching 図解

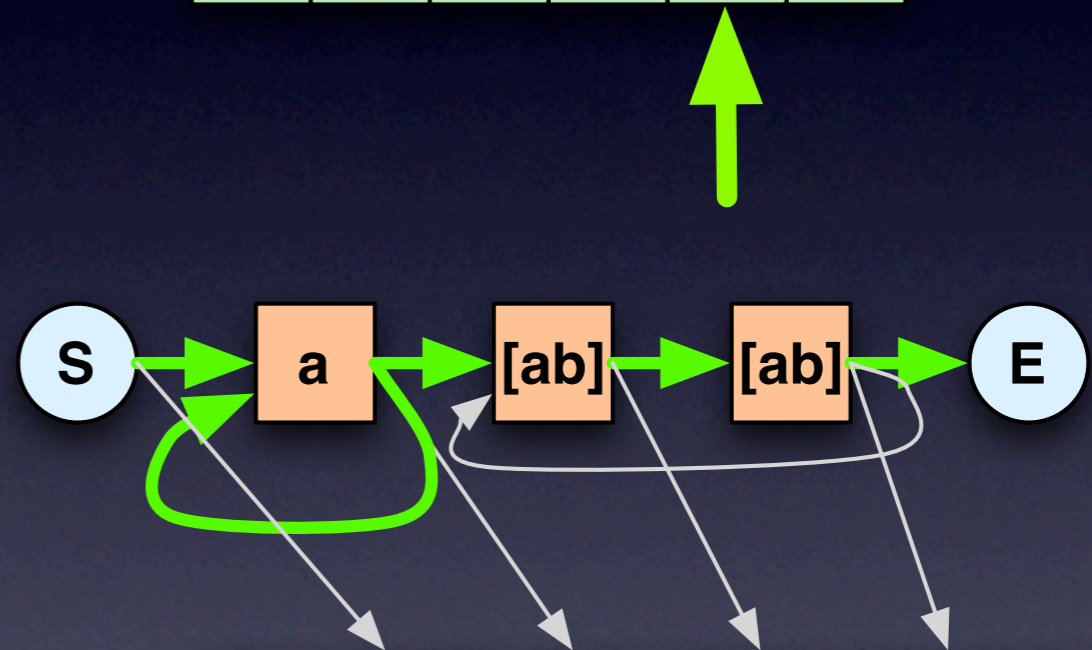
- 正規表現  
「 $a+([ab][ab])+$ 」
- 文字列 `aaabbb`
- 繰り返しを優先



# greedy matching 図解

- 正規表現  
「 $a+([ab][ab])+$ 」
- 文字列 `aaabbb`
- 繰り返しを優先

a	a	a	[ab]	[ab]	
a	a	a	b	b	b



マッチング失敗、終了  
or  
マッチング開始位置を1文字後ろに

# まとめ

- POSIX正規表現=longest matching
  - 性能↓の可能性：PCREを使おう
- PCRE=greedy matching
  - 「最長」は誤訳／繰り返しを欲張る

ご清聴

ありがとうございます

ございました