Principles of
Compiler Construction

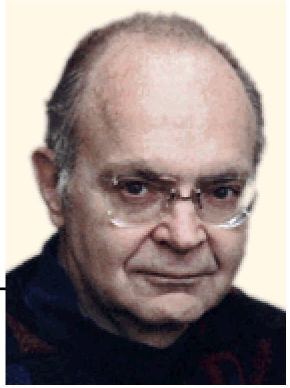
Step：
1文法预处理
2自底向上，增广文法 S'->S
 自自顶向下
3parsing table

Prof. Wen-jun LI
School of Computer Science and Engineering
lnslwj@mail.sysu.edu.cn

Lecture 7. LR Parsing

1. Introduction to LR Parsing
2. Motivation of LR Parsing
3. Simple LR Parsing
   - LR(0) and SLR(1)
4. More Powerful LR Parsing
   - LR(1) and LALR(1)
5. Ambiguity in LR Parsing
6. Error Recovery in LR Parsing

1. Introduction to LR Parsing



- Proposed by
  - D. Knuth (Stanford U.). **On the Translation of Languages from Left to Right**. Information and Control, 8(6), 1965, pp.607-639
    - Prof. 高德纳: The Art of Computer Programming(TAOCP), T_EX, Literate Programming, LR Parsing, Attribute Grammar, etc.
- Pros
  - Recognize almost all practical CFGs (more powerful than LL parsing).
  - High efficiency.
  - Detect errors as soon as possible.
- Cons
  - A large size of parsing table.
  - Hard to construct the parsing table manually.

2. Motivation of LR Parsing

- Critical: shift-reduce decision making
  - Maintain states to keep track of where we are in the parsing procedure.
- States are represented by items
  - $A \rightarrow X\ Y\ Z$ yields four items (LR(0) items)
    stack
    - $\boxed{A \rightarrow}\ \bullet \boxed{X\ Y\ Z}$ lookstack
    - $A \rightarrow X \bullet Y\ Z$
    - $A \rightarrow X\ Y \bullet Z$
    - $A \rightarrow X\ Y\ Z \bullet$
  - $A \rightarrow \varepsilon$ yields only one item: $A \rightarrow \bullet$

Items

- Different forms of items
  - A → X • a Z　　$S \to \cdot bBB$ ← 移进状态
    - A "shift" item　移进项目
    - Indicates the shift of symbol **a**.
  - A → X • B Z
    - A "reduce-expected" item. 待约项目
    - Indicates the expectation of reducing a B from the remaining input string.
  - A → X Y Z •
    - A "reduce" item　归约项目
    - Indicates a handle has appeared on top of stack.
  - S → X Y Z •
    - An "accept" item (special form of reduce items)
    - Indicates the state of accepting the input.

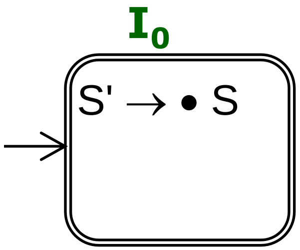
Augmented Grammar 拓展文法

- How and why of augmented grammars?
  - Add a new start symbol S'
    - S' $\rightarrow$ S
    - S $\rightarrow$ ...
  - Intent of augmenting a grammar
    - The start symbol will never appear in the body of any productions.
    - Then the accepting state (items) is unique: S' $\rightarrow$ S •

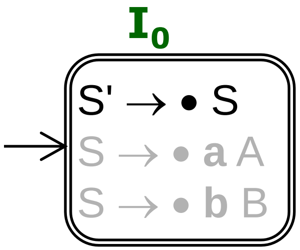
A Motivating Example

- Consider the following grammar

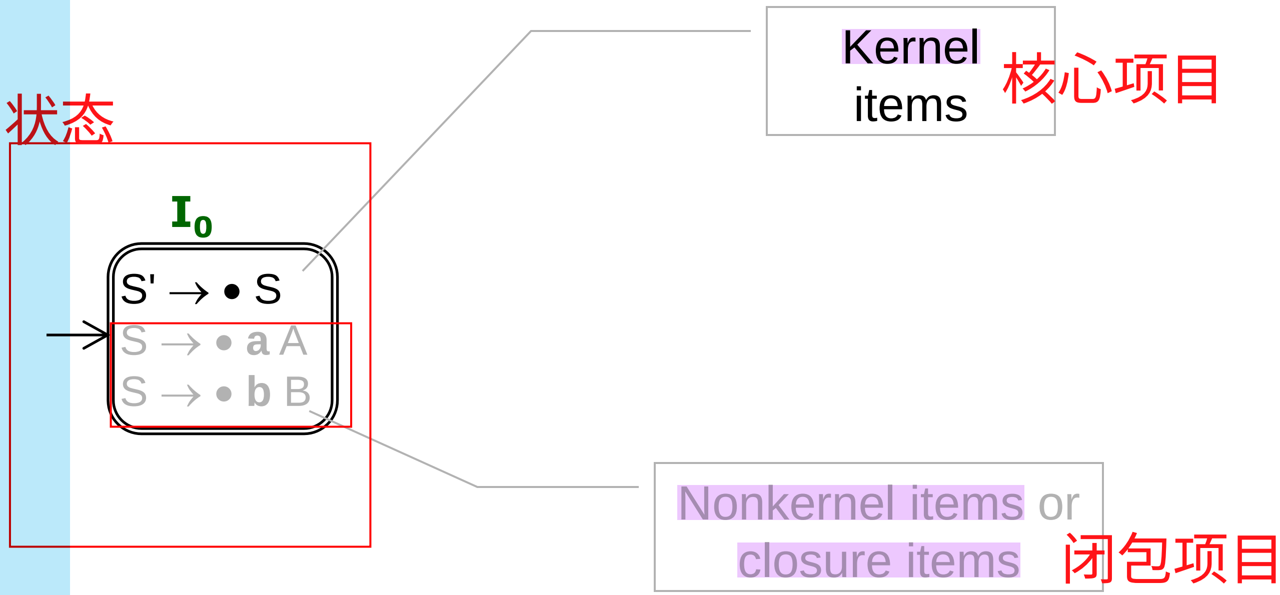
$$
\begin{aligned}
S' &\rightarrow S \\
S &\rightarrow \mathbf{a}\, A \mid \mathbf{b}\, B \\
\boxed{A} &\rightarrow \mathbf{c}\, A \mid \boxed{\mathbf{d}} \\
\boxed{B} &\rightarrow \mathbf{c}\, B \mid \boxed{\mathbf{d}}
\end{aligned}
$$



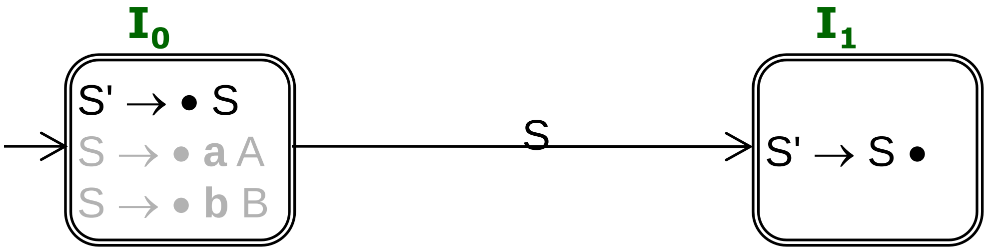
Initial state



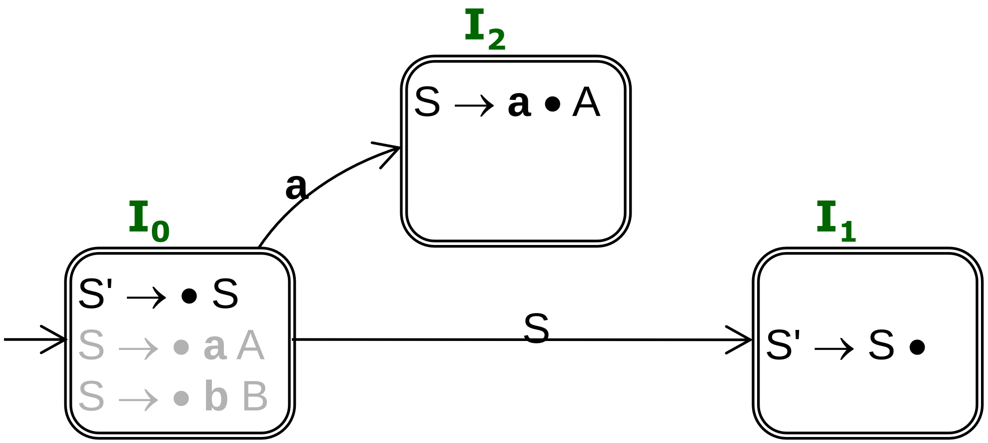
Equivalent closure



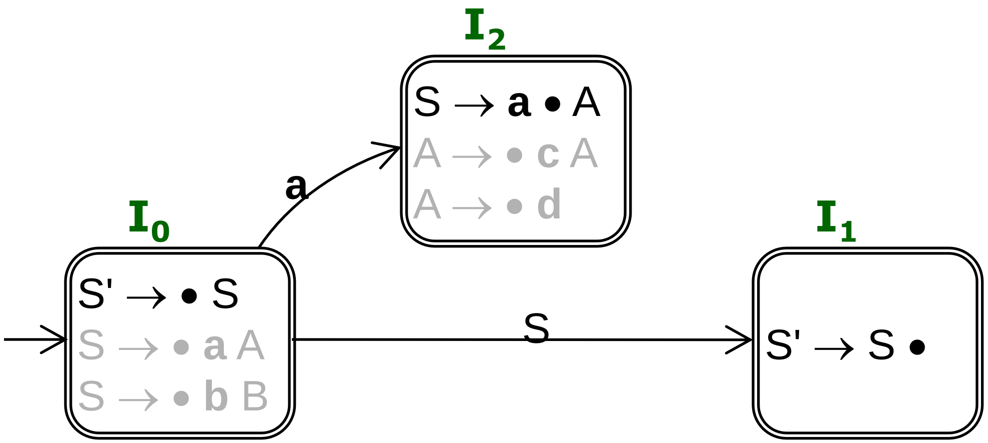
Equivalent closure



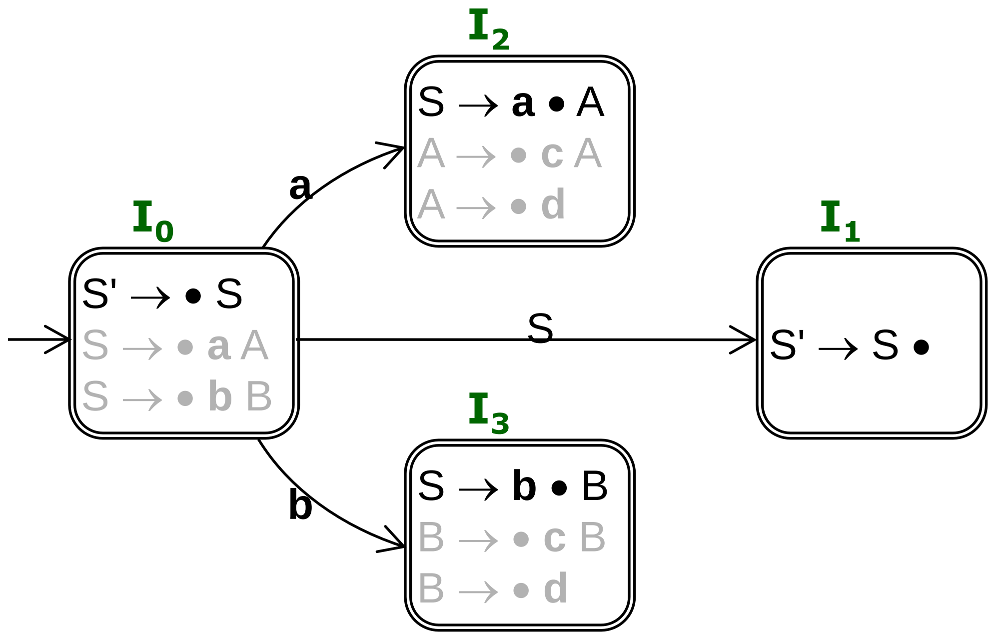
If the remaining string can be reduced to S (expected!)



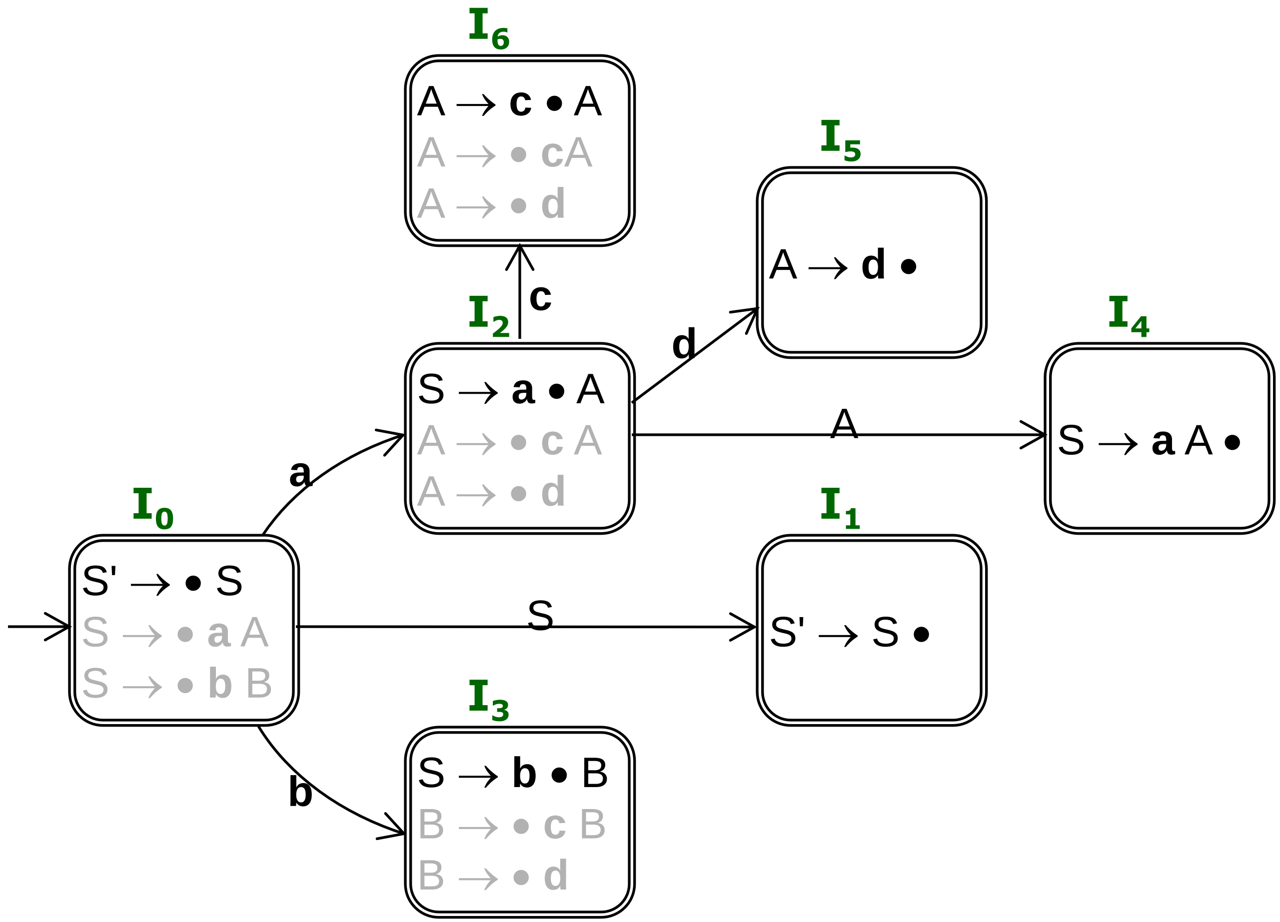
If the first symbol of remaining string is **a** (shift!)



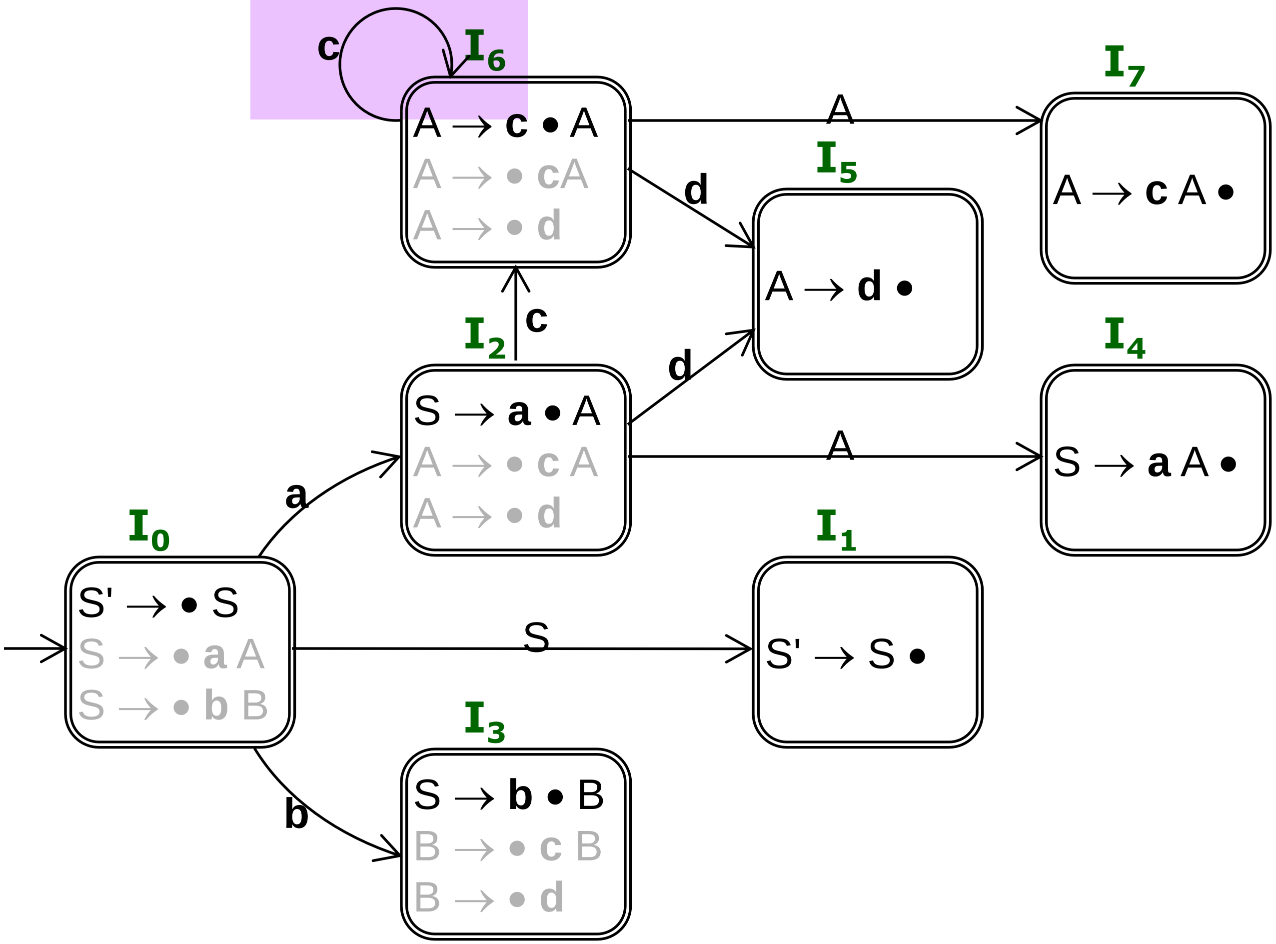
Equivalent closure in state I_2



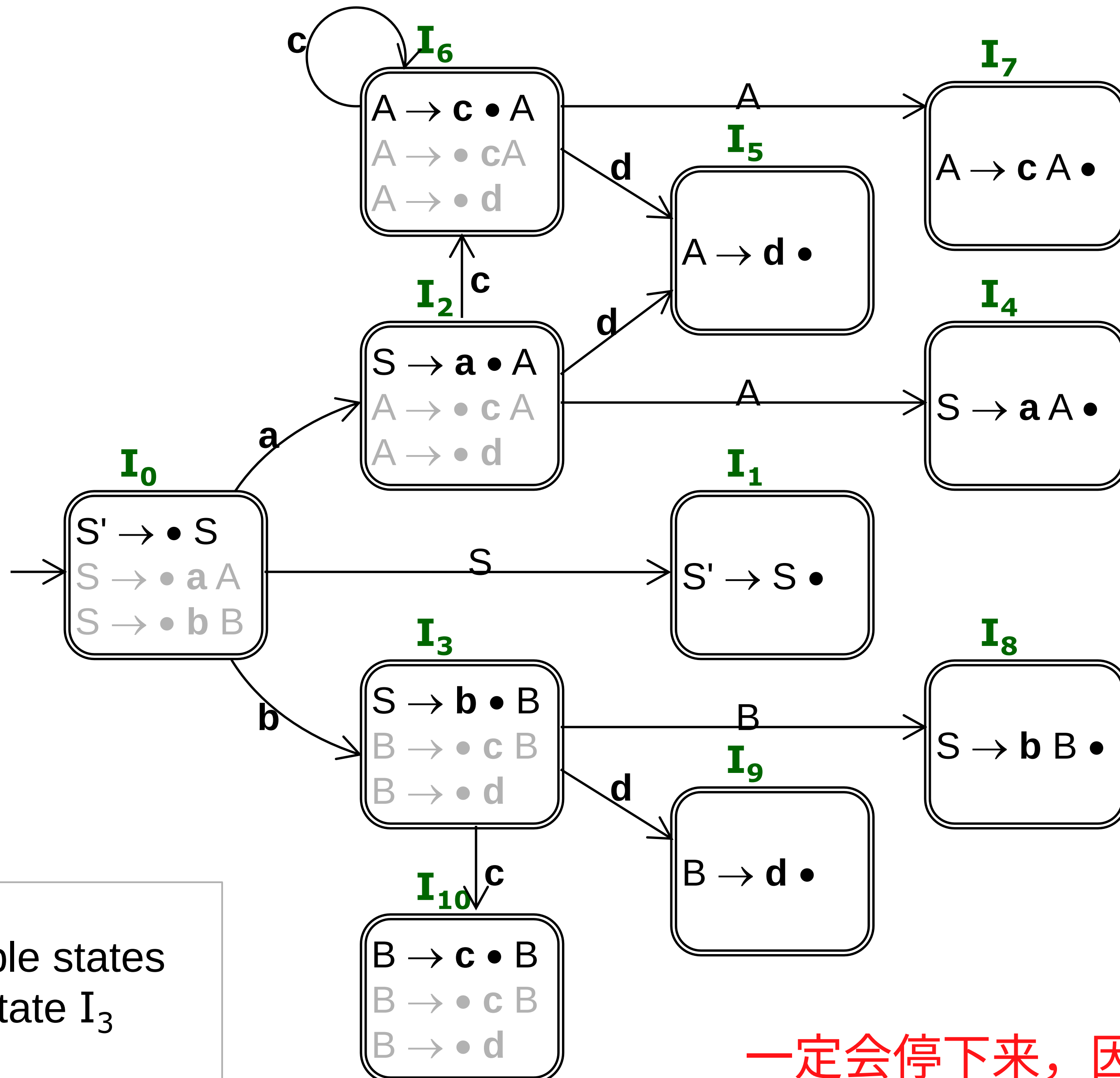
If the first symbol of remaining string is **b** (shift!)



All possible states from state I_2

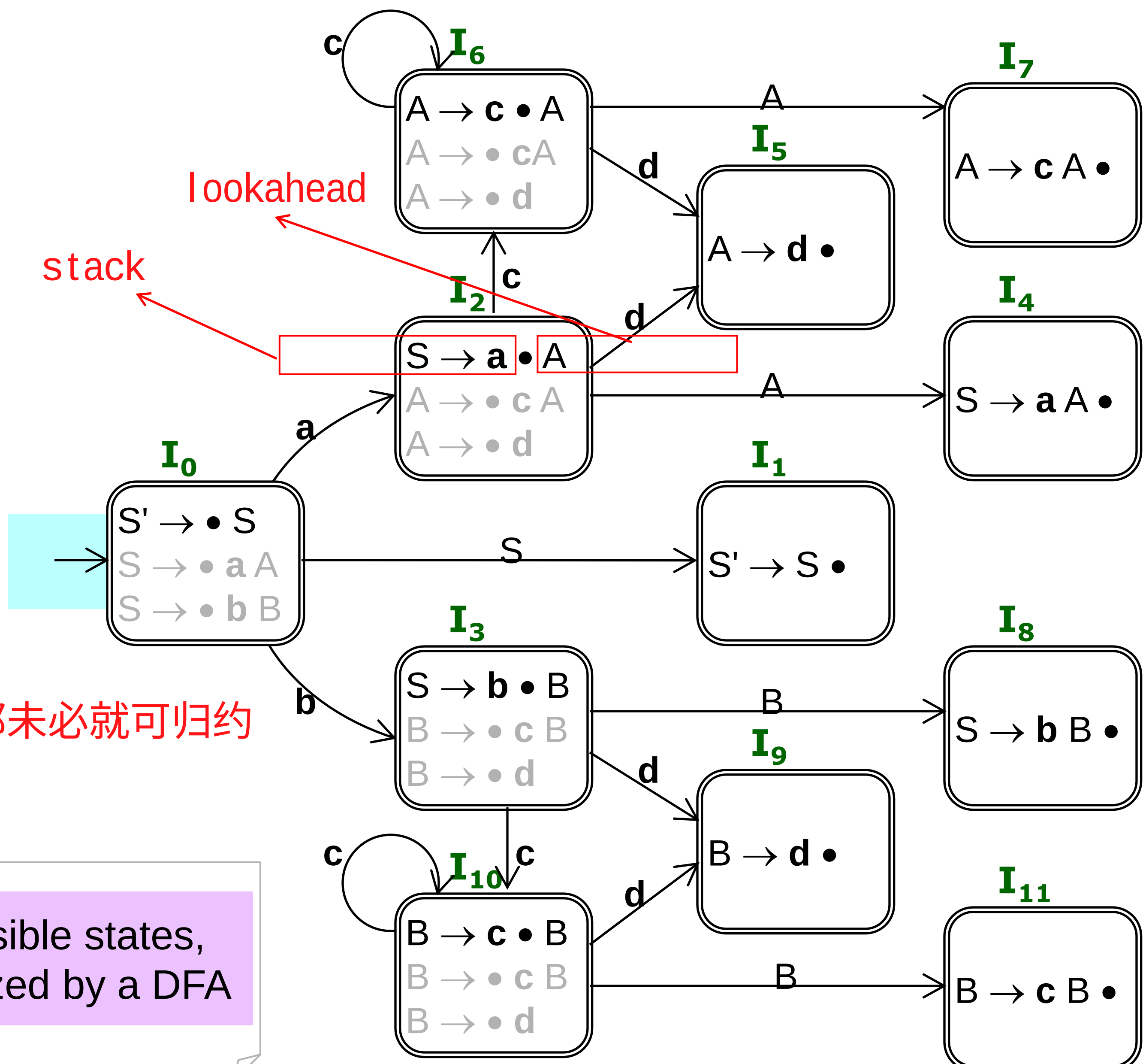


All possible states from state I_6



All possible states from state I_3

一定会停下来，因为是有穷的



形成了右部未必就可归约

All possible states, recognized by a DFA

所有状态都是终结状态

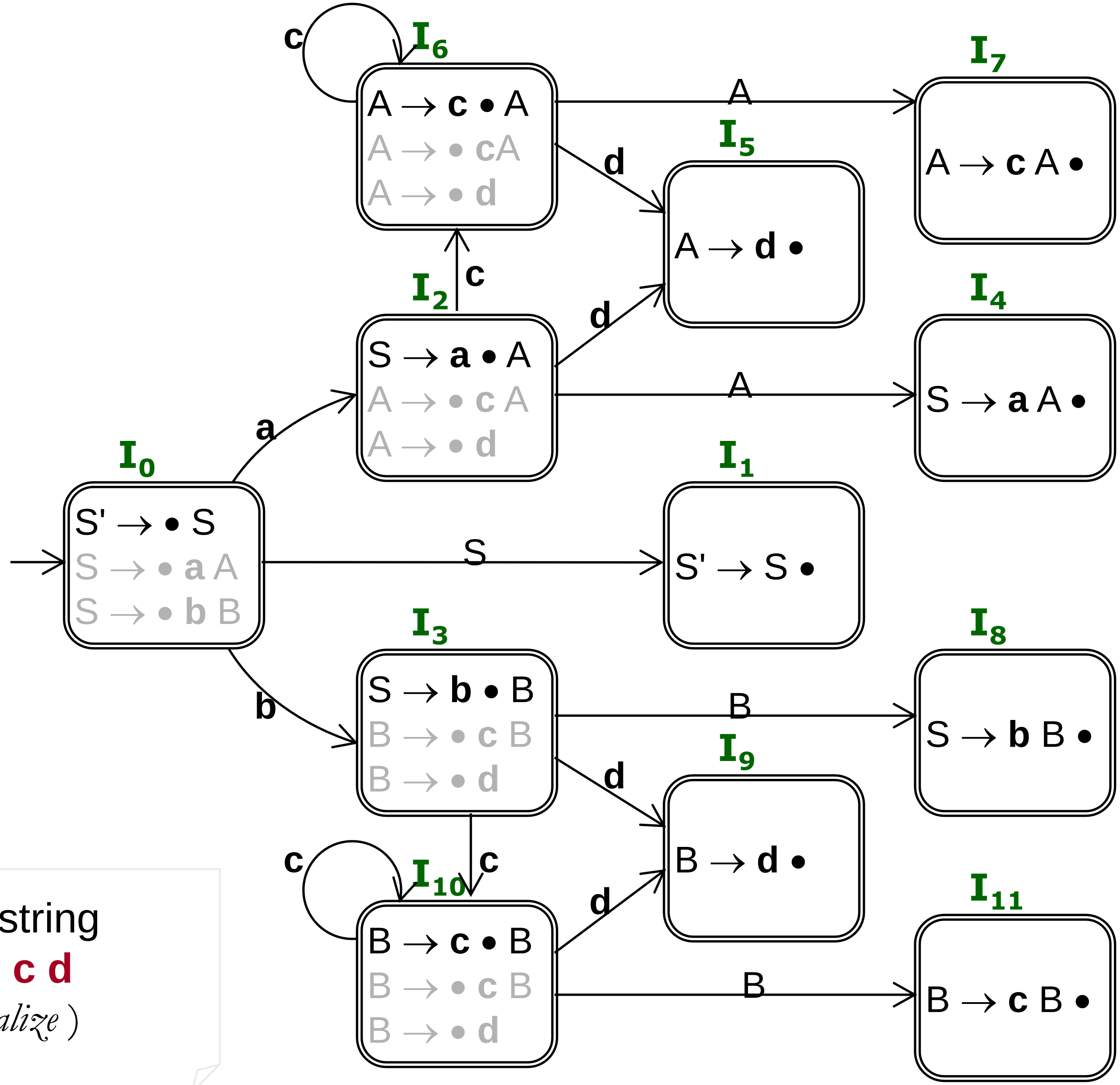
Working with the DFA

- Consider the following sentence:

  $\mathbf{a\ c\ c\ d}$

  We have the right-most derivation:

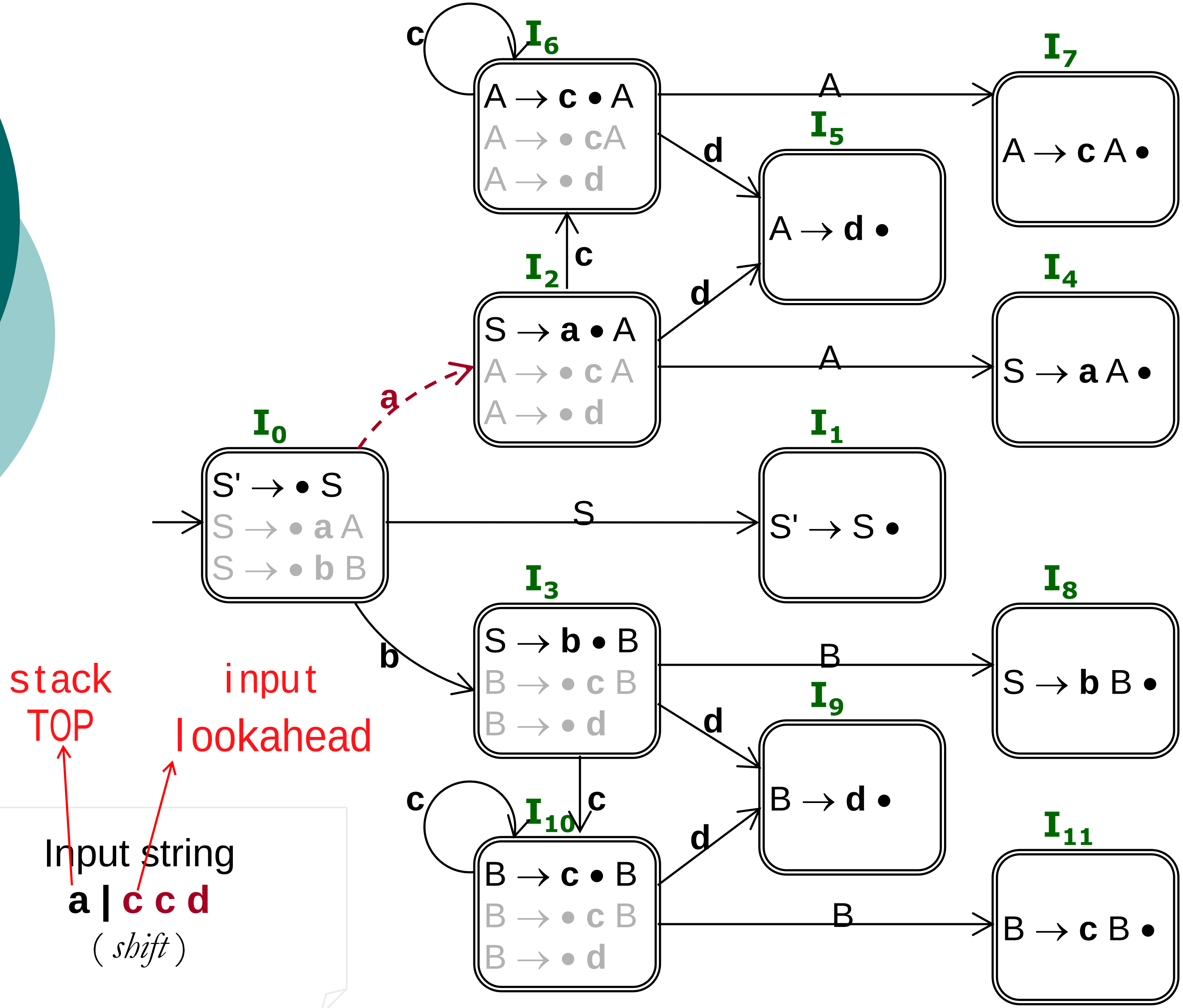
  $S' \Rightarrow S \Rightarrow \mathbf{a}\,A \Rightarrow \mathbf{a\ c}\,A \Rightarrow \mathbf{a\ c\ c}\,A \Rightarrow \mathbf{a\ c\ c\ d}$

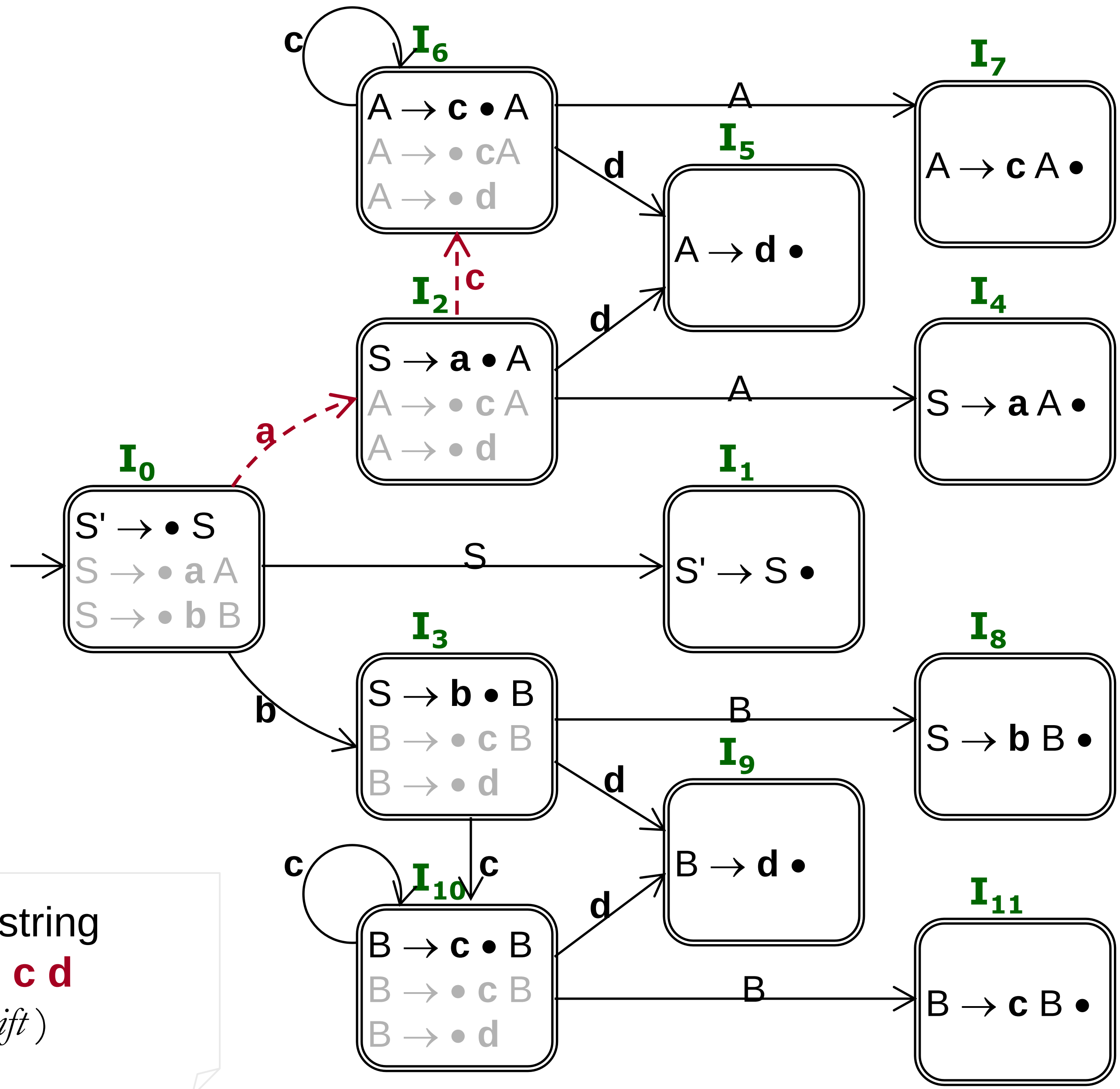


Input string

| a c c d

(*initialize*)

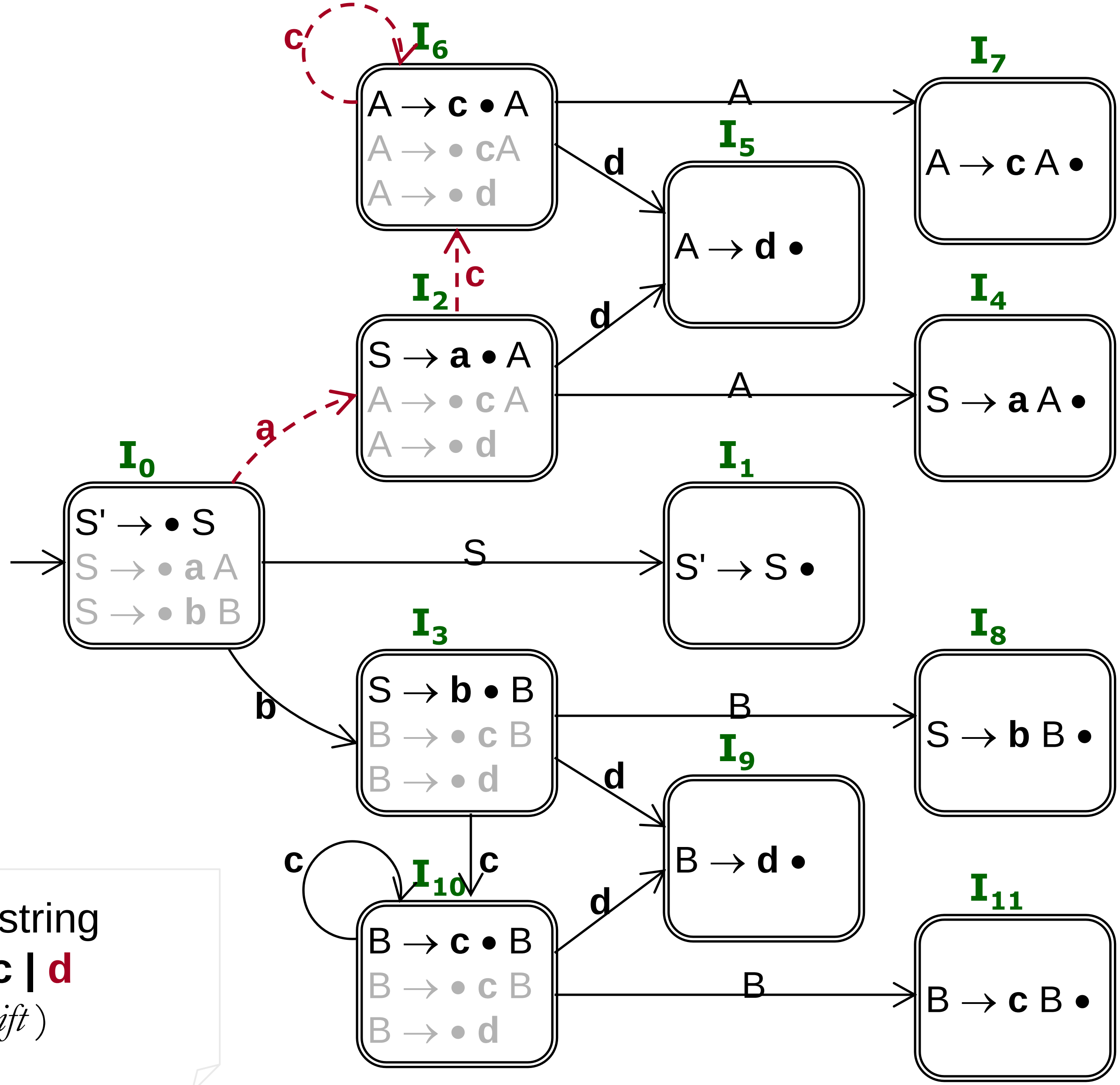

I0
S' → • S
S → • a A
S → • b B
I1
S' → S •
I2
S → a • A
A → • c A
A → • d
I3
S → b • B
B → • c B
B → • d
I4
S → a A •
I5
A → d •
I6
A → c • A
A → • cA
A → • d
I7
A → c A •
I8
S → b B •
I9
B → d •
I10
B → c • B
B → • c B
B → • d
I11
B → c B •
a
b
c
d
S
A
B
stack
TOP
input
lookahead
Input string
a | c c d
(shift)



Input string

a c | c d

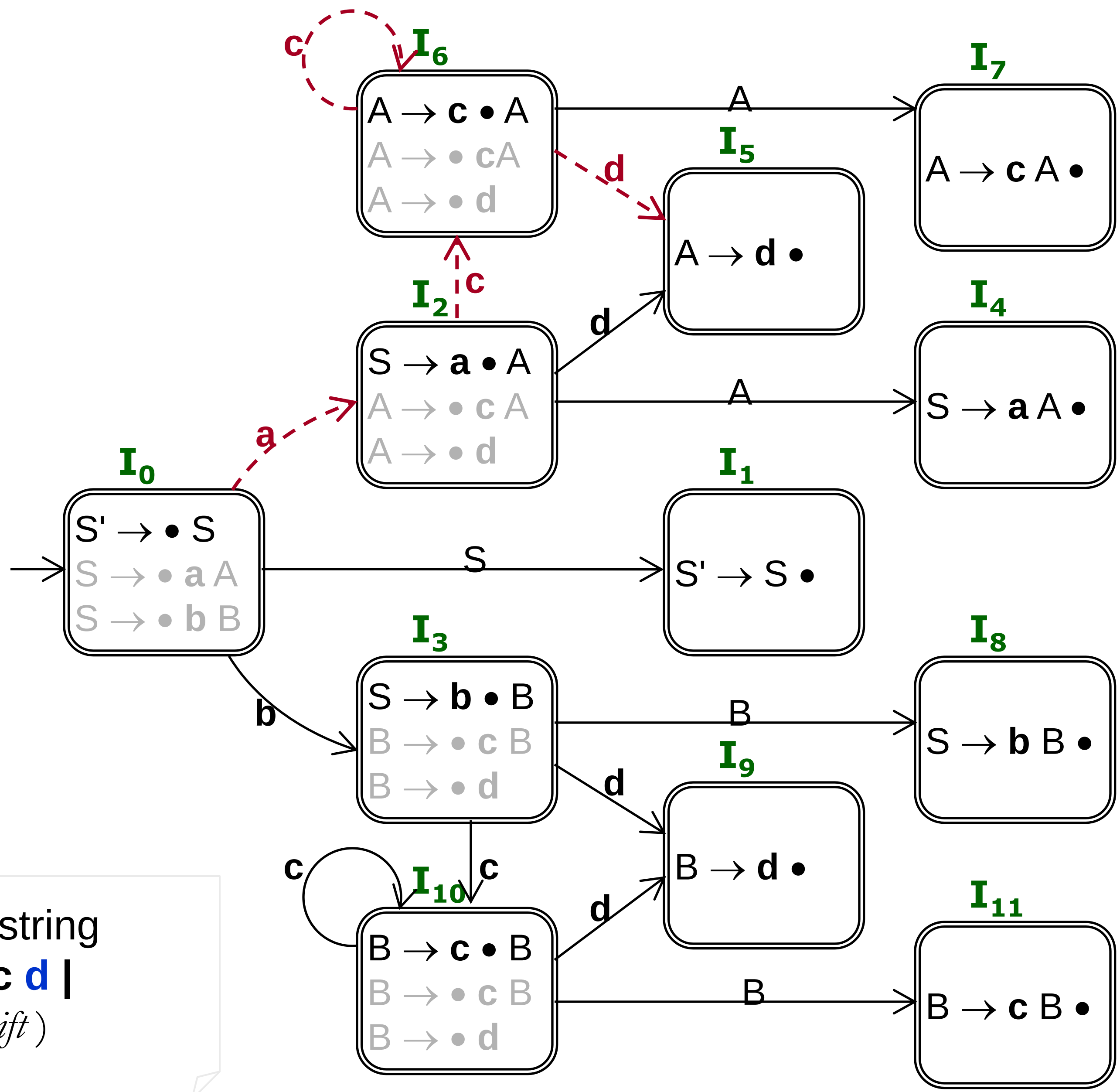
(*shift*)



Input string

a c c | d

(*shift*)

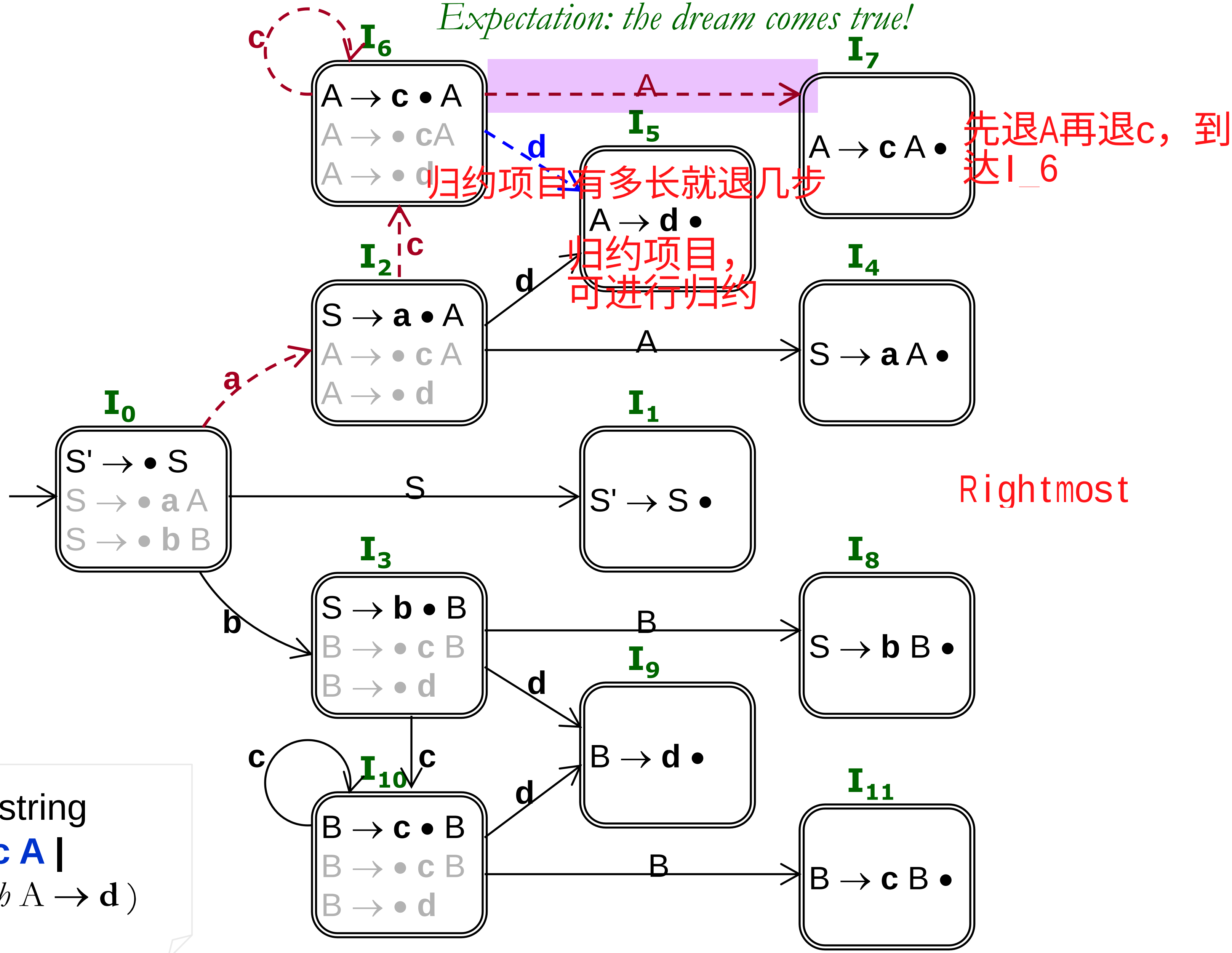


Input string

a c c **d** **|**

(*shift*)

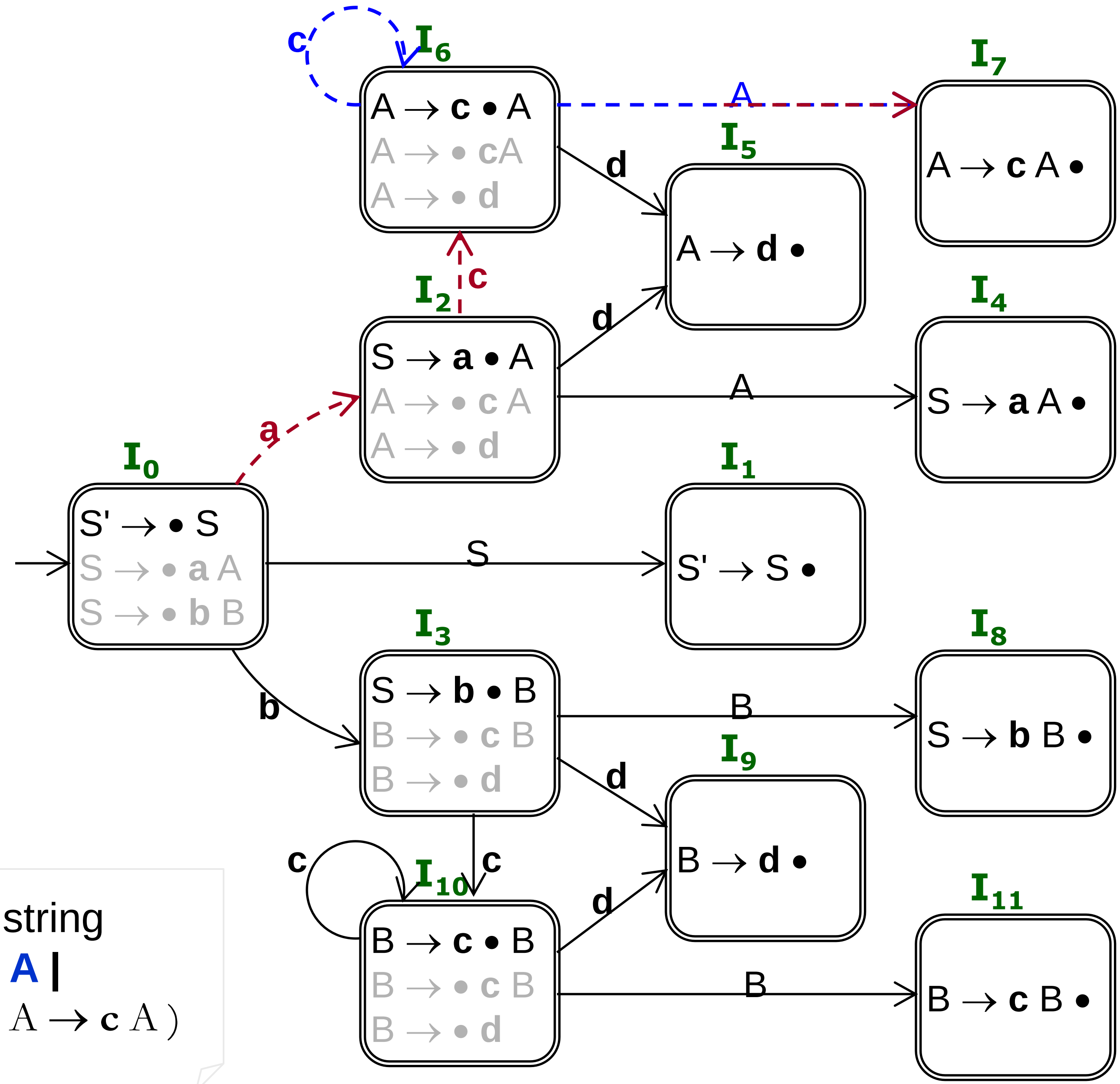
Expectation: the dream comes true!



Input string

a c c A |

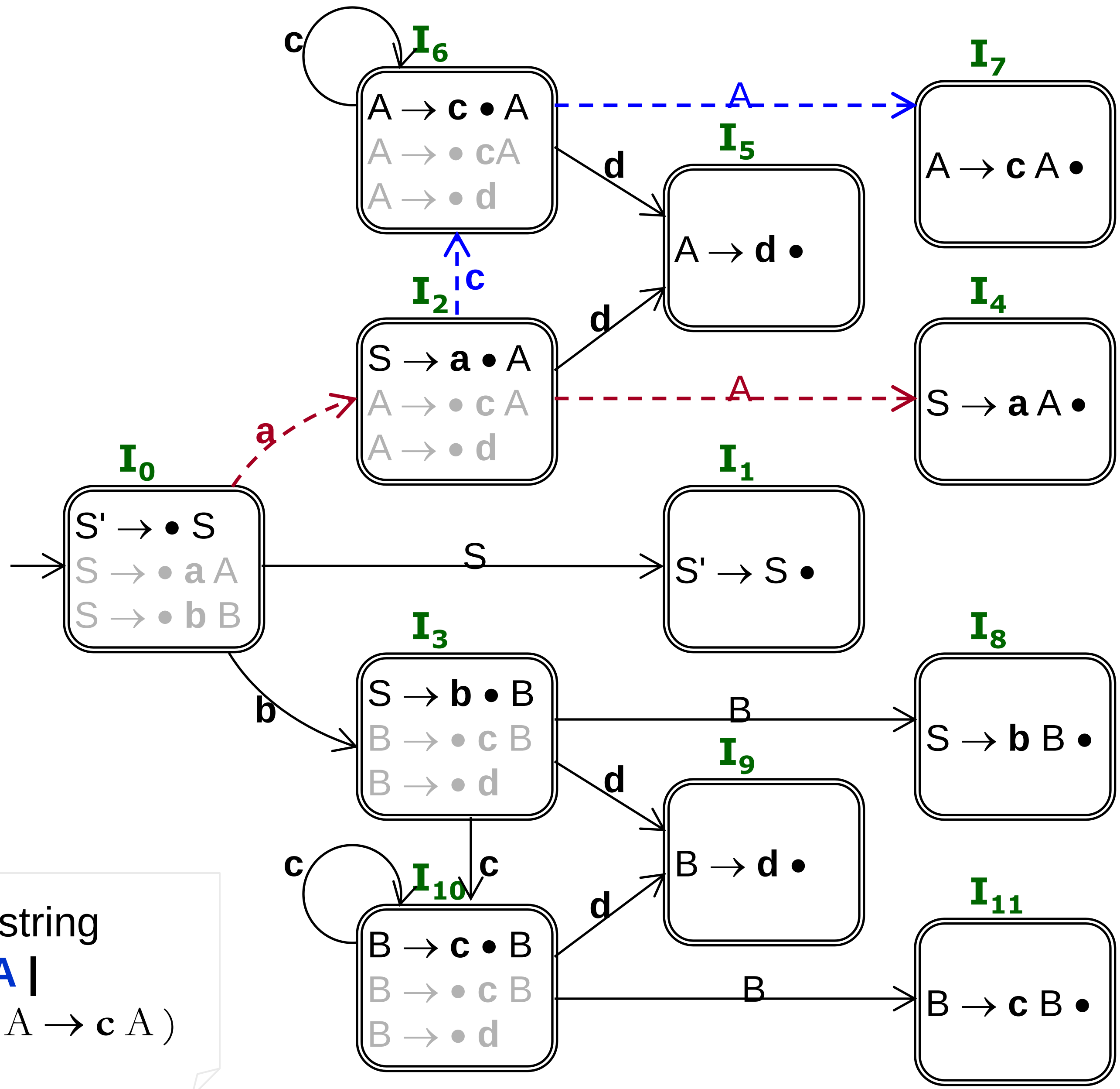
(*reduce with* $A \to \mathbf{d}$)



Input string

a **c** **A** **|**

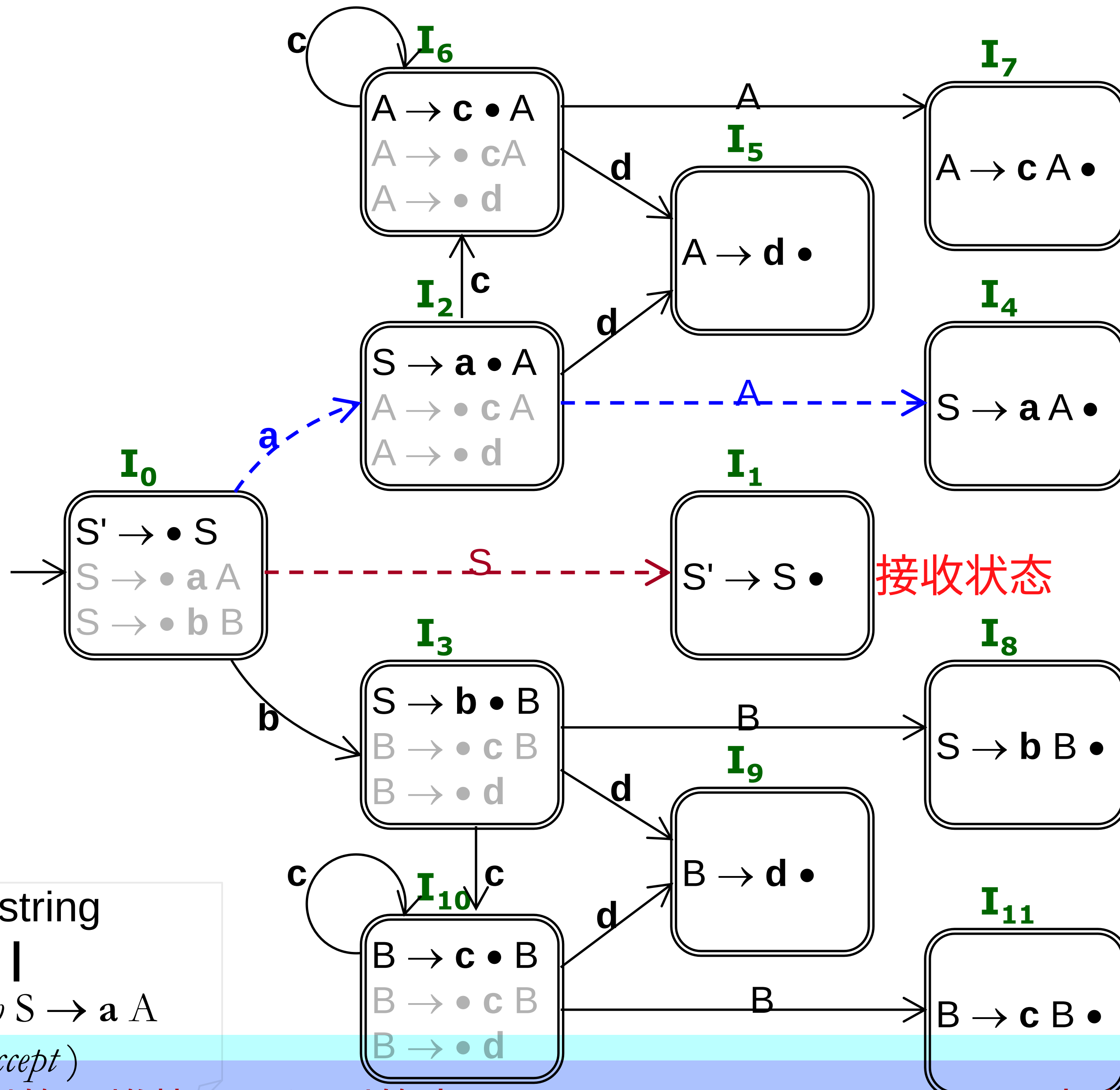
(*reduce with* A → **c** A)



Input string

a A |

(*reduce with* A → **c** A)



Input string

S |

(*reduce with* S → **a** A *and accept*)

是DFA。识别的是堆栈可以识别的串，不是contest-free grammar定义的语言，之所以能够实现是因为借助了stack。栈是状态栈（里面是状态I_0...），符号栈只是辅助理解。

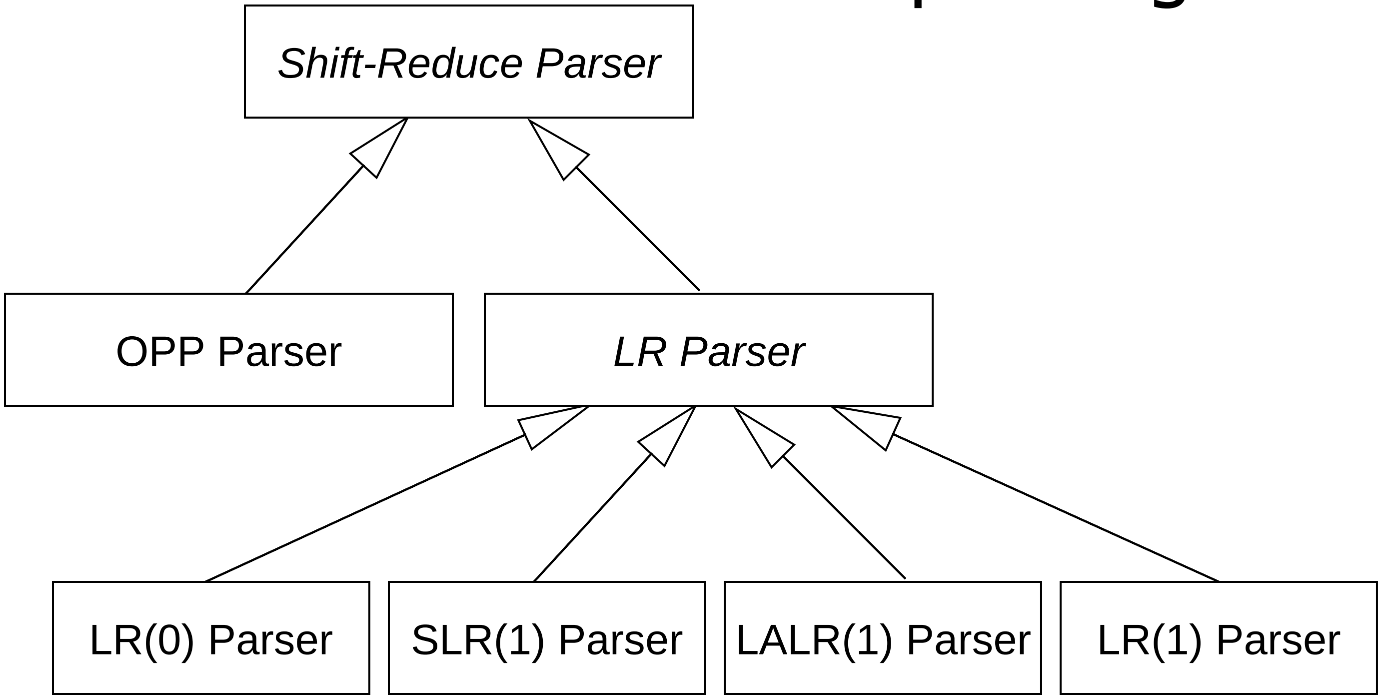
Parsing Decisions

Step	Symbol	State	Input	Reference	Action	Output
1	\$	0	**a c c d \$**	S → • **a** A	shift	
2	\$ a	0 2	**c c d \$**	A → • **c** A	shift	
3	\$ a c	0 2 6	**c d \$**	A → • **c** A	shift	
4	\$ a c c	0 2 6 6	**d \$**	A → • **d**	shift	
5	\$ a c c d	0 2 6 6 5	**\$**	A → **d** •	reduce	A → **d**
6	\$ a c c A	0 2 6 6 7	**\$**	A → **c** A •	reduce	A → **c** A
7	\$ a c A	0 2 6 7	**\$**	A → **c** A •	reduce	A → **c** A
8	\$ a A	0 2 4	**\$**	S → **a** A •	reduce	S → **a** A
9	\$ S	0 1	**\$**	S' → S •	accept	

栈底

Review

- Implementations of the abstract model for shift-reduce parsing

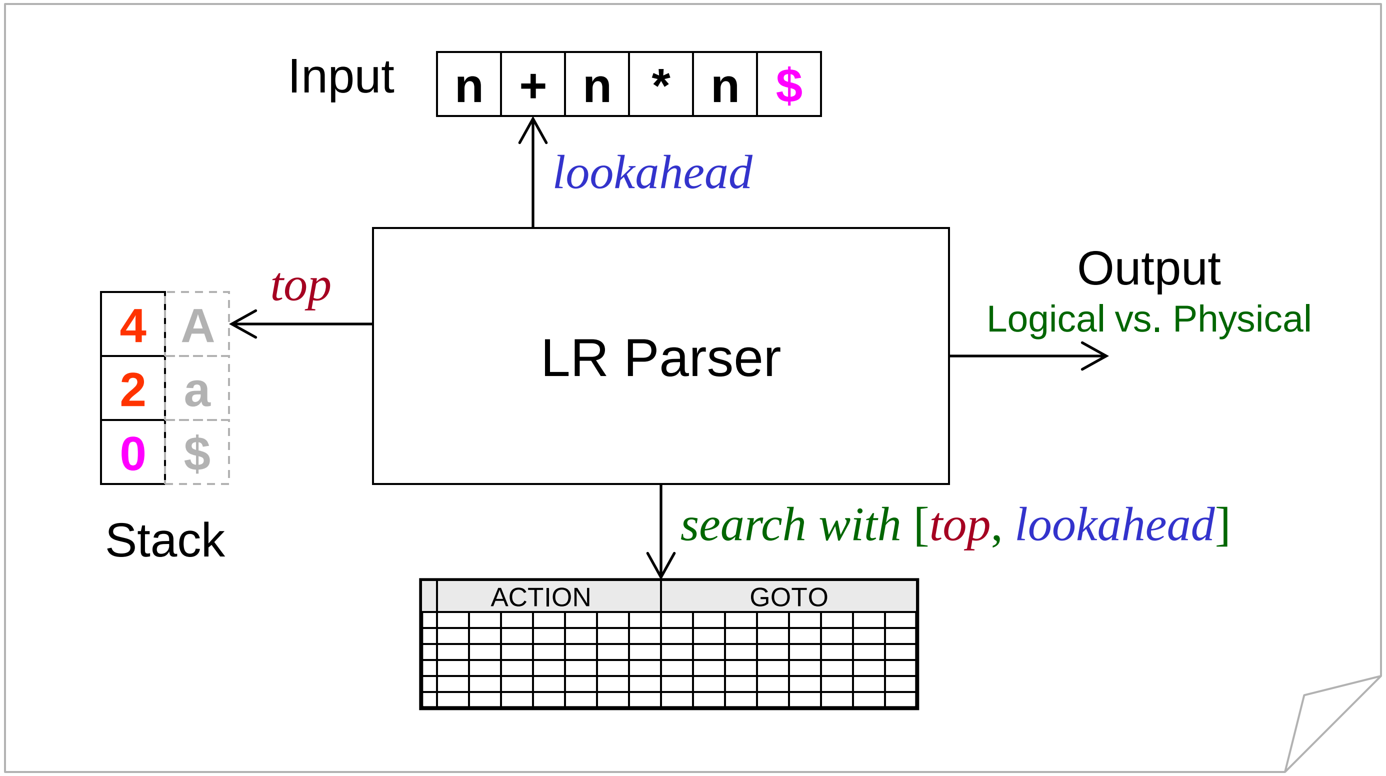


LR Parsers

- Concrete implementations
  - Parsing table
    - ACTION 状态+lookahead 状态*(T+1) $
    - GOTO 大小：状态*(T+N)
  - Explicit stack
    - States
    - Grammar symbols (optional)
  - Reducable substring
    - Handles

A Concrete Model for LR Parser

- LR Parser



What Language the DFA recognizes?

- Viable prefixes of the grammar
  - All strings of grammar symbols that can appear on the stack.
  - *A* ***viable prefix*** *is a prefix of a right-sentential form, that does not continue past the right end of the right-most handle of that sentential form.* (DBv2, pp.256, seg.4, line.2-4)

右句型

意思是：形成句柄就立刻reduce，不再shift

Learn how to build up a formulation!

Properties of Viable Prefixes

- A viable prefix must be a prefix of a right-sentential form.
  - $S \Rightarrow^*_{rm} \alpha\ \omega$, where α is the content of the stack and ω contains no nonterminals.
- Not all prefixes of a right-sentential form are viable prefixes.
  - $E \Rightarrow^*_{rm} F$ *** n** $\Rightarrow_{rm}$ **(** E **) * n**,
  - where **(**, **(E**, **(E)** are viable prefixes,
  - but **(E) *** is not, since the parser will perform reduction once the handle appears.

Definition: Valid Item

- The definition of a valid item
  - Item $A \rightarrow \beta_1 \bullet \beta_2$ is valid for a viable prefix $\alpha\beta_1$, if there exists a derivation

    $$S' \Rightarrow^*_{rm} \alpha\, A\, \omega \Rightarrow_{rm} \alpha\, \beta_1\, \beta_2\, \omega$$

    where ω contains terminals only.
  - Hints provided by valid items
    - If $\beta_2 \neq \varepsilon$, the parser should do shift() since a handle has not appear on top of parsing stack.
    - If $\beta_2 = \varepsilon$, the parser should do reduce().

合法项目，用于pp闭包项目生成

Theorem 1 on Valid Items

- Foundation of **closure()**, which is used to construct the states of DFA.
  - If $A \to \beta_1 \bullet B\ \beta_2$ is a valid item for $\alpha\beta_1$, and $B \to \gamma$ is a production, then $B \to \bullet\ \gamma$ is also a valid item for $\alpha\beta_1$.

可以推出

  - [Proof]
    - $S' \Rightarrow^*_{rm} \alpha\ A\ \omega \Rightarrow_{rm} \alpha\ \beta_1\ B\ \beta_2\ \omega$ (definition)
    - $\Rightarrow^*_{rm} \alpha\ \beta_1\ B\ \delta\ \omega$ (suppose $\beta_2\ \omega \Rightarrow^*_{rm} \delta\ \omega$)
    - $\Rightarrow_{rm} \alpha\ \beta_1\ \gamma\ \delta\ \omega$ (we have $B \to \gamma$)
    - $B \to \bullet\ \gamma$ is a valid item for $\alpha\beta_1$ (definition)

Theorem 2 on Valid Items

- Foundation of **goto()**, which is used to construct the transitions of DFA.
  - If $A \to \beta_1 \bullet X\ \beta_2$ is a valid item for $\alpha\beta_1$, then $A \to \beta_1\ X \bullet \beta_2$ is a valid item for $\alpha\beta_1 X$.
  - [Proof]
    - $S' \Rightarrow^*_{rm} \alpha\ A\ \omega \Rightarrow_{rm} \alpha\ \beta_1\ X\ \beta_2\ \omega$ (definition)
    - $S' \Rightarrow^*_{rm} \alpha\ A\ \omega \Rightarrow_{rm} \alpha\ \beta_1\ X\ \beta_2\ \omega$ (definition)
    - $A \to \beta_1\ X \bullet \beta_2$ is valid for $\alpha\beta_1 X$ (definition)

3. Simple LR Parsing

- Steps of parsing table construction
  - Augment the grammar. 开始符号不在右部　说明
    - To ensure a unique accepting state.
  - Draw the DFA recognizing all viable prefixes of the grammar.
  - Calculate FIRST() and FOLLOW() sets of all nonterminal symbols.
    - Or on-demand calculating while filling the table.
  - Fill the LR parsing table (ACTION and GOTO).
    - Using lookahead to decide reductions.

Augmented Grammar

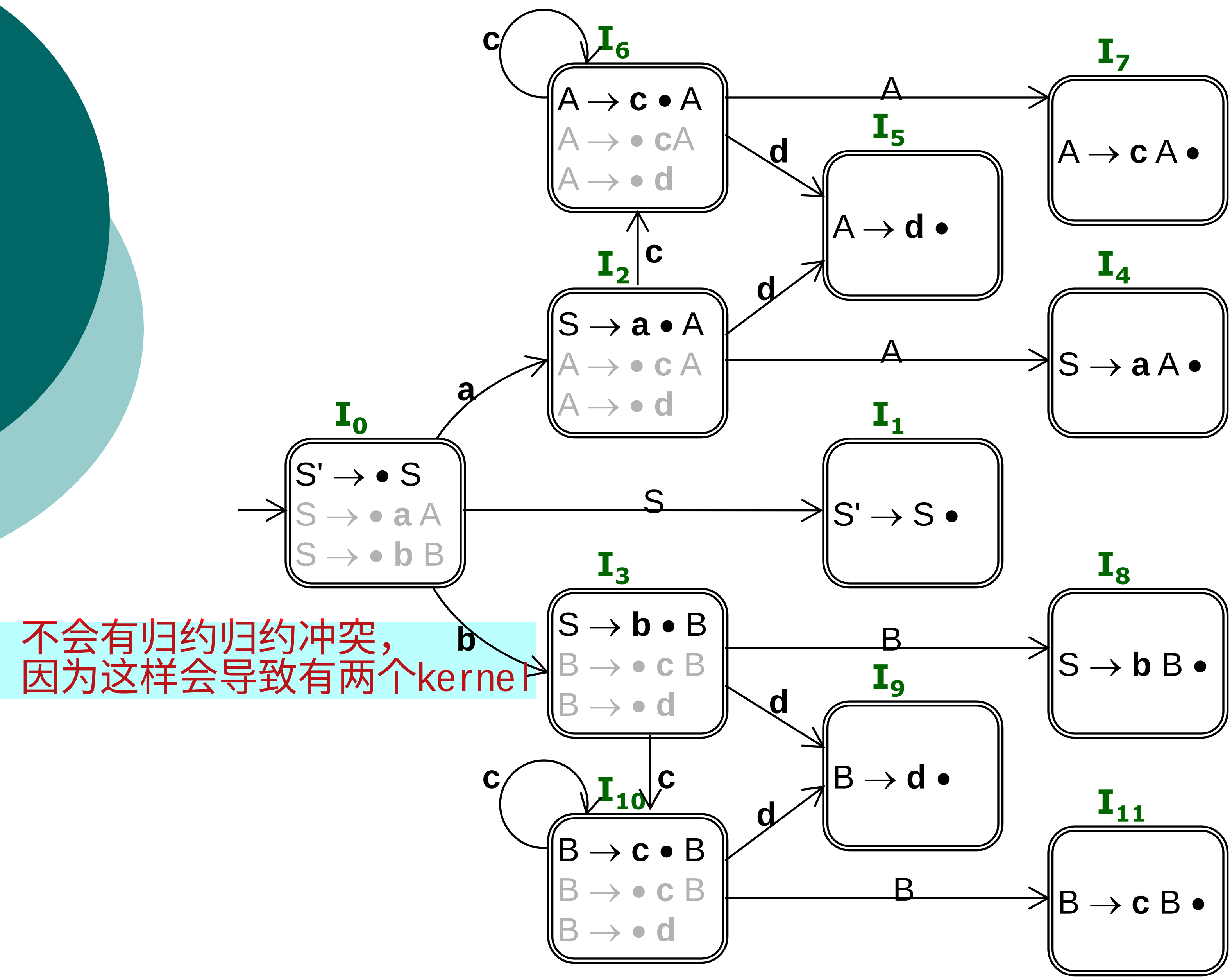
- Given the previous example augmented and numbered:

(0)	S'	$\rightarrow$	S
(1)	S	$\rightarrow$	**a** A
(2)	S	$\rightarrow$	**b** B
(3)	A	$\rightarrow$	**c** A
(4)	A	$\rightarrow$	**d**
(5)	B	$\rightarrow$	**c** B
(6)	B	$\rightarrow$	**d**

引用序号进行reduce

FIRST() and FOLLOW() Sets

- We have
  - FIRST(S) = { **a**, **b** }
  - FIRST(A) = { **c**, **d** }
  - FIRST(B) = { **c**, **d** }
  - FOLLOW(S) = { **$** }
  - FOLLOW(A) = { **$** }
  - FOLLOW(B) = { **$** }



Recall: DFA recognizing all viable prefixes

- ACTION 状态+lookahead 状态*(T+1)
- GOTO 大小：状态*(T+N)

定义：用LR(0)没有冲突

LR(0) Parsing Table

实际上是两张表画在一起

State	ACTION					GOTO		
	a	**b**	**c**	**d**	**$**	S	A	B
0	s2	s3				1		
1					acc			
2			s6	s5			4	
3			s10	s9				8
4	r1	r1	r1	r1	r1			
5	r4	r4	r4	r4	r4			
6			s6	s5			7	
7	r3	r3	r3	r3	r3			
8	r2	r2	r2	r2	r2			
9	r6	r6	r6	r6	r6			
10			s10	s9				11
11	r5	r5	r5	r5	r5			

shift后去状态2

前往的状态

top和lookahead连在一起就是右句型

SLR只是利用lookahead

*LR(0) uses **0** lookahead to decide reducing*

s:shift r：reduce r1：用产生式1来归约

有可能有conflic，及某个空格出现多个不同的操作

Overlap of Two Tables

State	ACTION					GOTO		
	a	**b**	**c**	**d**	**$**	S	A	B
0	s2	s3				1		
1					acc			
2			s6	s5			4	
3			s10	s9				8
4	r1	r1	r1	r1	r1			
5	r4	r4	r4	r4	r4			
6			s6	s5			7	
7	r3	r3	r3	r3	r3			
8	r2	r2	r2	r2	r2			
9	r6	r6	r6	r6	r6			
10			s10	s9				11
11	r5	r5	r5	r5	r5			

Overlap of Two Tables (cont')

State	ACTION					GOTO		
	a	**b**	**c**	**d**	**$**	S	A	B
0	s2	s3				1		
1					acc			
2			s6	s5			4	
3			s10	s9				8
4	r1	r1	r1	r1	r1			
5	r4	r4	r4	r4	r4			
6			s6	s5			7	
7	r3	r3	r3	r3	r3			
8	r2	r2	r2	r2	r2			
9	r6	r6	r6	r6	r6			
10			s10	s9				11
11	r5	r5	r5	r5	r5			

Decisions Based on Parsing Table

Step	Symbol	State	Input	Reference	Action	Output
1	$	0	**a c c d $**	a[0, **a**] = s2	shift	
2	$ a	0 2	**c c d $**	a[2, **c**] = s6	shift	
3	$ a c	0 2 6	**c d $**	a[6, **c**] = s6	shift	
4	$ a c c	0 2 6 6	**d $**	a[6, **d**] = s5	shift	
5	$ a c c d	0 2 6 6 5	**$**	a[5, **$**] = r4 g[6, A] = 7	reduce	A → **d**
6	$ a c c A	0 2 6 ~~6 7~~	**$**	a[7, **$**] = r3 g[6, A] = 7	reduce	A → **c** A
7	$ a c A	0 2 ~~6 7~~	**$**	a[7, **$**] = r3 g[2, A] = 4	reduce	A → **c** A
8	$ a A	0 ~~2 4~~ 3	**$**	a[4, **$**] = r1 1 g[0, S] = 1	reduce	S $\xrightarrow{2}$ **a** A
9	$ S 4	0 $\boxed{1}$ 5	**$**	a[1, **$**] = acc	accept	

在parsing时辅助理解，无实际作用。实际上用于存储结果

区别：一行有reduce就全行是reduce，依据：填的越精准，冲突越小；在SLR只填follow的

Simple LR (SLR) Parsing Table

State	ACTION					GOTO		
	a	**b**	**c**	**d**	**$**	S	A	B
0	s2	s3				1		
1					acc			
2			s6	s5			4	
3			s10	s9				8
4					r1			
5					r4			
6			s6	s5			7	
7					r3			
8					r2			
9					r6			
10			s10	s9				11
11					r5			

SLR(1) *simply uses* **1** *lookahead to decide reducing*

(Lookaheads must be in the FOLLOW *set of the reduced symbol)*

Difference between LR(0) and SLR(1)

- Parsing with the LR(0) parsing table

Step	Symbol	State	Input	Reference	Action	Output
1	$	0	**a d d $**	a[0, **a**] = s2	shift	
2	$ a	0 2	**d d $**	a[2, **d**] = s5	shift	
3	**$ a d**	0 2 5	**d $**	a[5, **d**] = r4, g[2, A] = 4	reduce	A → **d**
4	**$ a** A	0 2 4	**d $**	a[4, **d**] = r1, g[0, S] = 1	reduce	S → **a** A
5	**$** S	0 1	**d $**	a[1, **d**] = empty	error	

- Parsing with the SLR(1) parsing table

Step	Symbol	State	Input	Reference	Action	Output
1	$	0	**a d d $**	a[0, **a**] = s2	shift	
2	$ a	0 2	**d d $**	a[2, **d**] = s5	shift	
3	**$ a d**	0 2 5	**d $**	a[5, **d**] = empty	error	

差别是报错速度的快慢，但一定会报错；

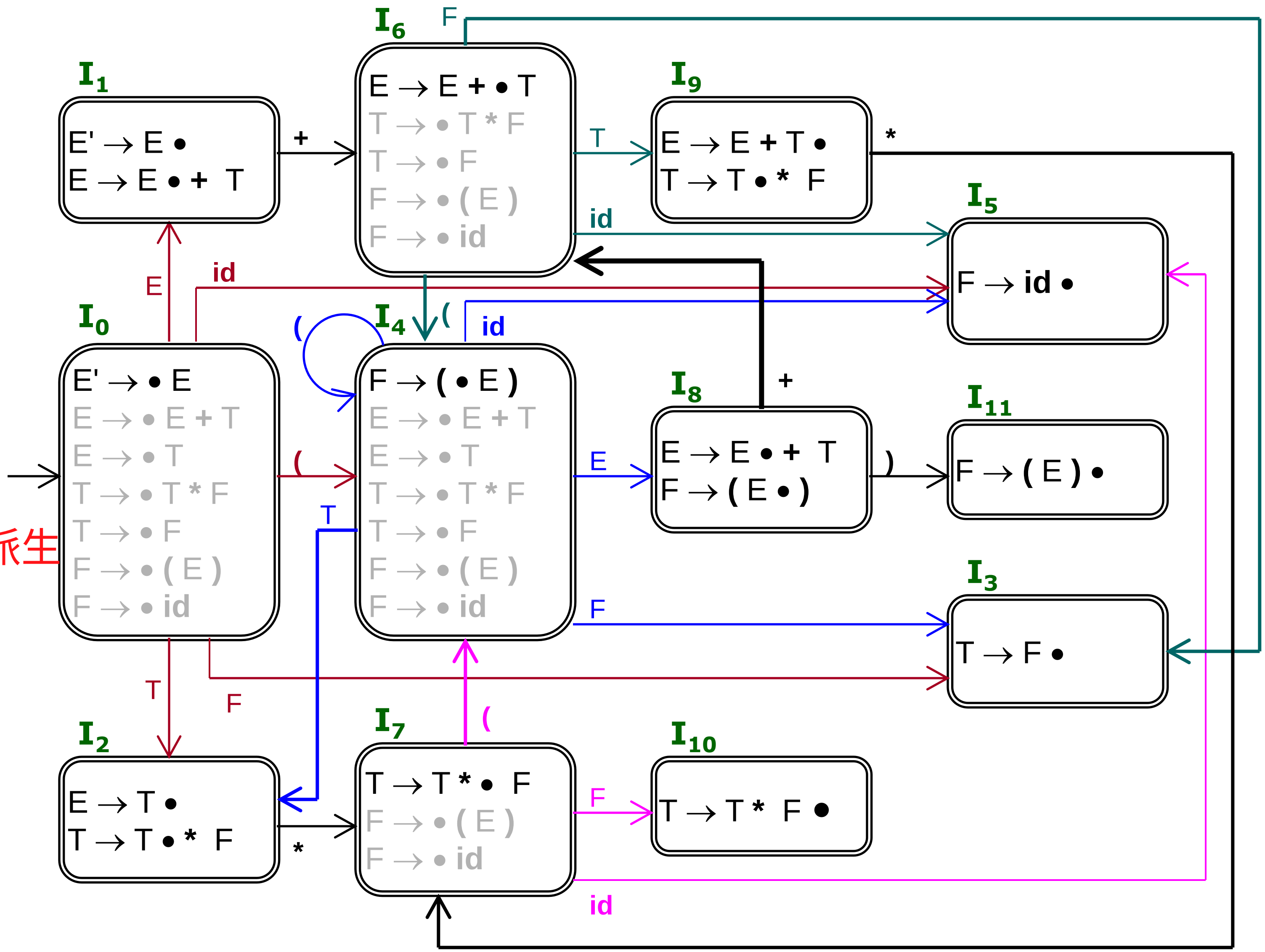
A More Complicated Example

- Consider the **unambiguous** grammar for expressions

$$E \rightarrow E + T \mid T$$
$$T \rightarrow T * F \mid F$$
$$F \rightarrow (E) \mid \mathbf{id}$$

- Augmented grammar (numbered)

$$
\begin{array}{llll}
(0) & E' & \rightarrow & E \\
(1) & E & \rightarrow & E + T \\
(2) & E & \rightarrow & T \\
(3) & T & \rightarrow & T * F \\
(4) & T & \rightarrow & F \\
(5) & F & \rightarrow & (E) \\
(6) & F & \rightarrow & \mathbf{id}
\end{array}
$$



不断进行派生

DFA recognizing all viable prefixes

FIRST() and FOLLOW() Sets

- From the previous lectures, it is easy to calculate the following sets:
  - FIRST(E) = { **(**, **id** }
  - FIRST(T) = { **(**, **id** }
  - FIRST(F) = { **(**, **id** }
  - FOLLOW(E) = { **+**, **)**, **$** }
  - FOLLOW(T) = { **+**, *****, **)**, **$** }
  - FOLLOW(F) = { **+**, *****, **)**, **$** }

考试：
增广文法、（求follow）、画出DFA、填写parsing table、问一个语句的工作过程
或者只是简单问是不是SLR，还是要重复上述过程

SLR(1) Parsing Table

State	ACTION						GOTO		
	id	**+**	*****	**(**	**)**	**$**	E	T	F
0	s5			s4			1	2	3
1		s6				acc			
2		r2	s7		r2	r2			
3		r4	r4		r4	r4			
4	s5			s4			8	2	3
5		r6	r6		r6	r6			
6	s5			s4				9	3
7	s5			s4					10
8		s6			s11				
9		r1	s7		r1	r1			
10		r3	r3		r3	r3			
11		r5	r5		r5	r5			

只填E的follow（state 2）

只填E的follow（state 9）

This Is Not an LR(0) Grammar!

State	ACTION						GOTO		
	id	**+**	*****	**(**	**)**	**$**	E	T	F
0	s5			s4			1	2	3
1		s6				acc			
2	r2	r2	s7/r2	r2	r2	r2	全行填reduce		
3	r4	r4	r4	r4	r4	r4			
4	s5			s4			8	2	3
5	r6	r6	r6	r6	r6	r6			
6	s5			s4				9	3
7	s5			s4					10
8		s6			s11				
9	r1	r1	s7/r1	r1	r1	r1			
10	r3	r3	r3	r3	r3	r3			
11	r5	r5	r5	r5	r5	r5			

4. More Powerful LR Parsing

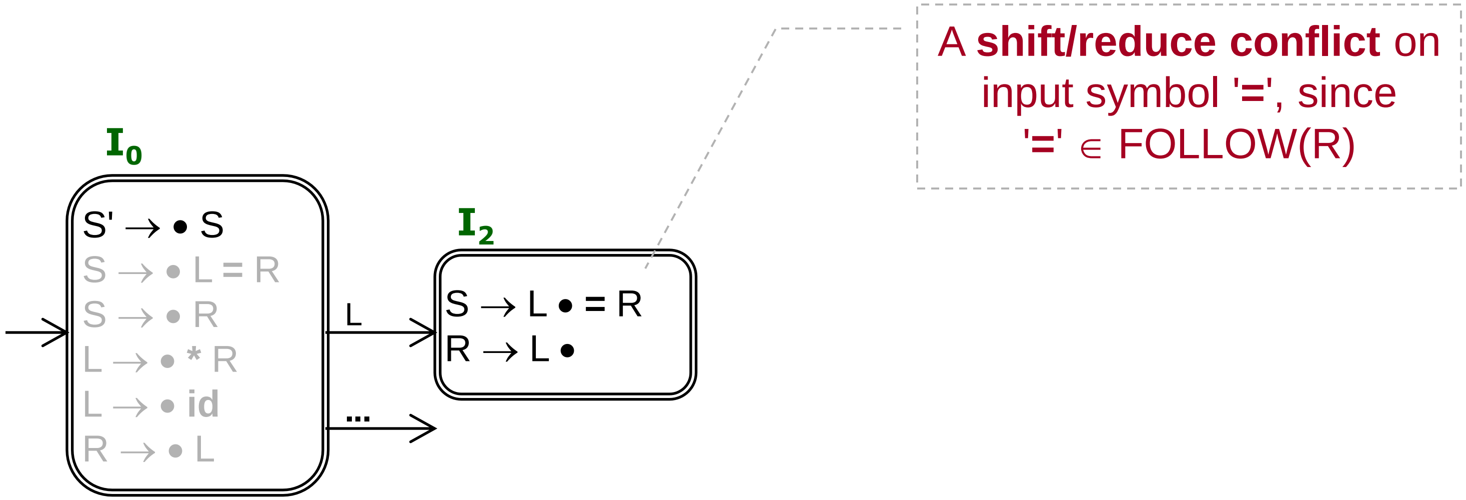
- A Grammar That Is Not SLR(1)
- LR(1) Parsing Table
- LALR(1) Parsing Table
- Conflicts in LALR(1) Parsing

An Unambiguous But Not SLR(1) Grammar

- Consider the unambiguous grammar, which is not SLR(1):

$$S \rightarrow L = R \mid R$$
$$L \rightarrow \mathbf{*}\, R \mid \mathbf{id}$$
$$R \rightarrow L$$

Conflicts in Items



Part of the DFA recognizing all viable prefixes

Motivation of LR(1) Parsing

- How to make use of the lookahead?
  - LR(0): does not use the lookahead.
    - All columns are filled with the same reduction.
  - SLR(1): simply make use of the lookahead.
    - Only columns in the FOLLOW set are filled.
    - More accurate than LR(0), thus less conflicts.
  - More powerful LR parsing – LR(1) 最精准，代价大
    - Only columns (symbols) that can follow the viable prefixes are filled. follow所有活前缀
    - More accurate than SLR(1).
  - Trade-off – LALR(1)
    - More efficient but less powerful than LR(1).
    - More powerful than SLR(1).

Definition: Valid LR(1) Item

- The definition of a valid item
  - Item [$A \to \beta_1 \bullet \beta_2$, a] is valid for a viable prefix $\alpha\beta_1$, if there exists a derivation

$$S' \Rightarrow^*_{rm} \alpha\, A\, \omega \Rightarrow_{rm} \alpha\, \beta_1\, \beta_2\, \omega$$

    where ω starts with a, or $\omega = \varepsilon \wedge$ a = $.

区别

Theorem 1 on Valid LR(1) Items

- Foundation of closure(), which is used to construct the states of DFA.
  - If [$A \to \beta_1 \bullet B\ \beta_2$, a] is a valid item for $\alpha\beta_1$, and $B \to \gamma$ is a production, then [$B \to \bullet\ \gamma$, b] is also a valid item for $\alpha\beta_1$, where $b \in$ FIRST(β_2a).

考试易出错的是a

  - [Proof]
    - $S' \Rightarrow^*_{rm} \alpha\ A\ a\ \omega \Rightarrow_{rm} \alpha\ \beta_1\ B\ \beta_2\ a\ \omega$ (definition)
    - $\Rightarrow^*_{rm} \alpha\ \beta_1\ B\ b\ \delta$ (suppose $\beta_2\ a\ \omega \Rightarrow^*_{rm} b\ \delta$)
    - $\Rightarrow_{rm} \alpha\ \beta_1\ \gamma\ b\ \delta$ (we have $B \to \gamma$)
    - [$B \to \bullet\ \gamma$, b] is valid for $\alpha\beta_1$ (definition)

Theorem 2 on Valid LR(1) Items

- Foundation of goto(), which is used to construct the transitions of DFA.
  - If $[A \rightarrow \beta_1 \bullet X\ \beta_2,\ a]$ is a valid item for $\alpha\beta_1$, then $[A \rightarrow \beta_1\ X \bullet \beta_2\ ,\ a]$ is a valid item for $\alpha\beta_1 X$.
  - [Proof]
    - $S' \Rightarrow^*_{rm} \alpha\ A\ a\ \omega \Rightarrow_{rm} \alpha\ \beta_1\ X\ \beta_2\ a\ \omega$ (definition)
    - $S' \Rightarrow^*_{rm} \alpha\ A\ a\ \omega \Rightarrow_{rm} \alpha\ \beta_1\ X\ \beta_2\ a\ \omega$ (definition)
    - $[A \rightarrow \beta_1\ X \bullet \beta_2\ ,\ a]$ is valid for $\alpha\beta_1 X$ (definition)

An Example

- Consider the following grammar

$$
\begin{aligned}
&(0)\quad S' \rightarrow S\\
&(1)\quad S \rightarrow C\,C\\
&(2)\quad C \rightarrow \mathbf{c}\,C\\
&(3)\quad C \rightarrow \mathbf{d}
\end{aligned}
$$

FIRST() and FOLLOW() Sets

- It is easy to calculate the following sets:
  - FIRST(S) = FIRST(C) = { **c**, **d** }
  - FOLLOW(S) = { **$** }
  - FOLLOW(C) = { **c**, **d**, **$** }

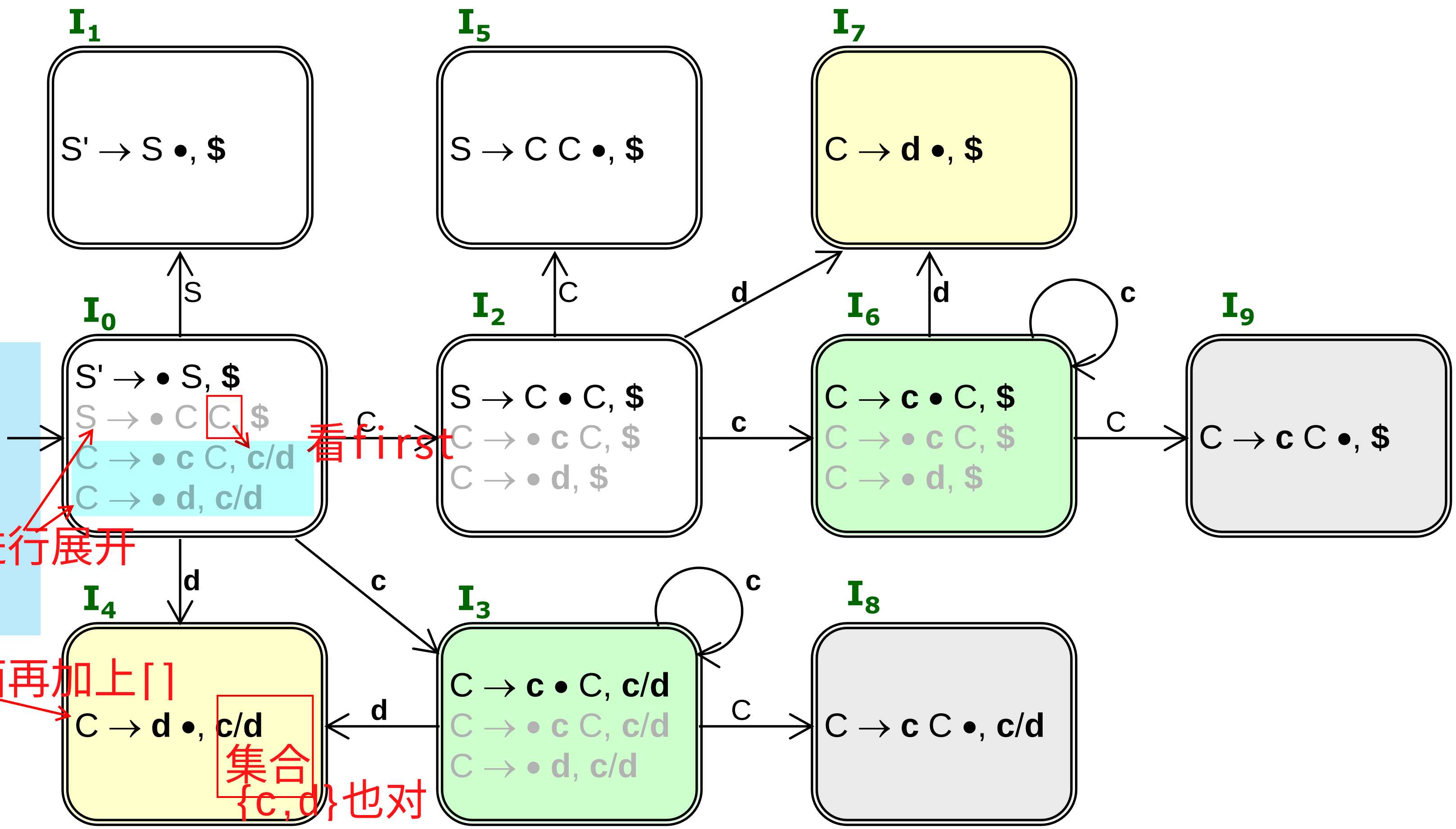
这些已经不care

这时看的是first

SLR看的是FOLLOW

- FIRST(S) = FIRST(C) = { **c**, **d** }

parsing table实际两张，只是overloading



需要不断进行展开

可以在外面再加上[]

因为augmentation，所以有$
保证S不在其他的右部

DFA recognizing all viable prefixes

活前缀

DFA可以识别ifuxilou

结构错误很严重，要注意

LR(1) Parsing Table

State	ACTION			GOTO	
	c	**d**	**$**	S	C
0	s3	s4		1	2
1			acc		
2	s6	s7			5
3	s3	s4			8
4	r3	r3			
5			r1		
6	s6	s7			9
7			r3		
8	r2	r2			
9			r2		

LALR(1) Parsing Table

- Lookahead-LR parsing in practice
  - Number of states, such as C or Pascal
    - LR(0) = SLR(1): several hundreds.
    - LR(1): several thousands (10 times).
  - Strategy for LALR(1): merge the states with the same **core**. lookahead LR(1)
    - E.g. I_3 and I_6, I_4 and I_7, I_8 and I_9
    - Lookaheads of the same item are merged.
    - GOTO() depends only on the core.

只是因为lookahead不同而分开，其他都一样，所以合并在一起

LALR(1) vs. LR(1)

- The merge will never produce **new** shift-reduce conflicts
  - Suppose in the merged state
    - [A $\rightarrow \alpha \bullet$, a] calls for a reduction
    - [B $\rightarrow \beta \bullet$ a γ, b] calls for a shift

LR1还是可能有冲突

  - Since the original states have the same core, there must be some state have
    - [A $\rightarrow \alpha \bullet$, a] calls for a reduction
    - [B $\rightarrow \beta \bullet$ a γ, c] calls for a shift (for some c)
  - Then the original state already has conflicts.

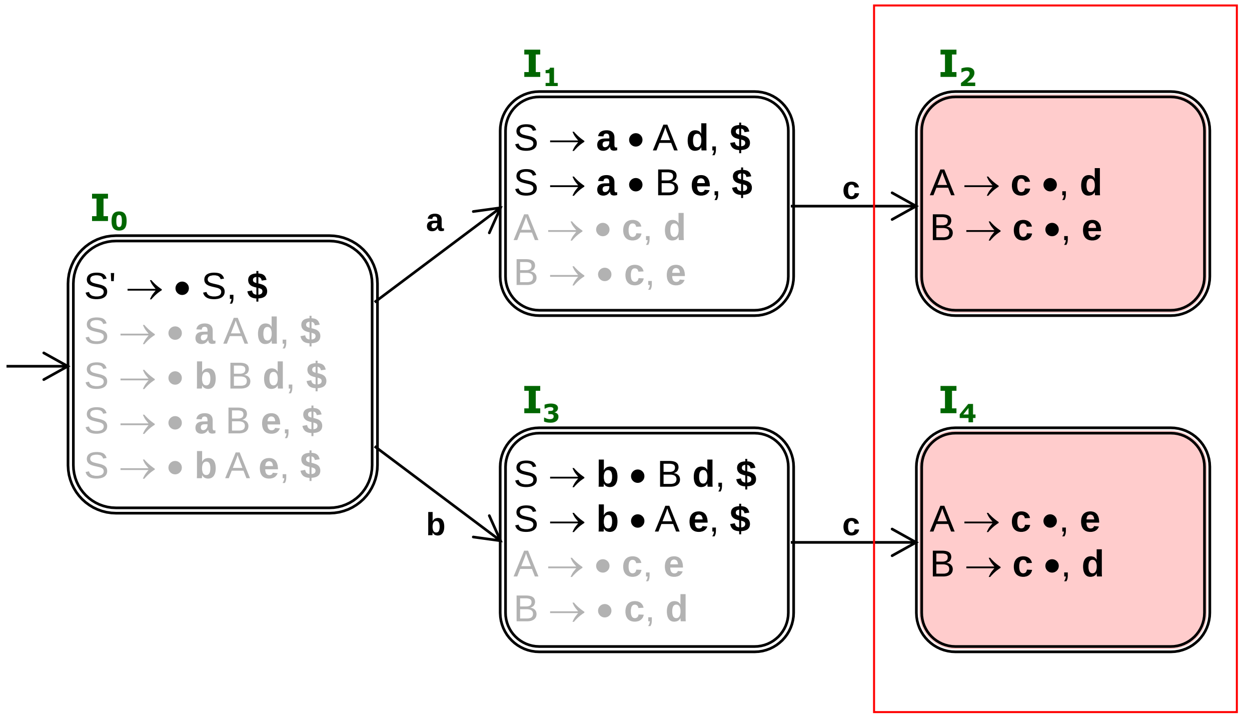
LALR(1) vs. LR(1) (cont')

- But the merge will produce **new** reduce-reduce conflicts
  - For example

$$
\begin{array}{lll}
S' & \rightarrow & S \\
S & \rightarrow & \mathbf{a}\,A\,\mathbf{d} \mid \mathbf{b}\,B\,\mathbf{d} \mid \mathbf{a}\,B\,\mathbf{e} \mid \mathbf{b}\,A\,\mathbf{e} \\
A & \rightarrow & \mathbf{c} \\
B & \rightarrow & \mathbf{c}
\end{array}
$$

  - $\{[A \rightarrow \mathbf{c} \bullet, \mathbf{d}], [B \rightarrow \mathbf{c} \bullet, \mathbf{e}]\}$ is valid for viable prefix ac, $\{[A \rightarrow \mathbf{c} \bullet, \mathbf{e}], [B \rightarrow \mathbf{c} \bullet, \mathbf{d}]\}$ is valid for viable prefix bc. But the merge has conflicts:
    - $\{[A \rightarrow \mathbf{c} \bullet, \mathbf{d/e}], [B \rightarrow \mathbf{c} \bullet, \mathbf{d/e}]\}$

Example: New Conflicts in LALR(1)



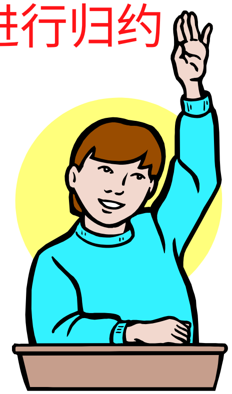
Part of the DFA recognizing all viable prefixes

More Efficient Algorithm to Construct LALR(1) Parsing Tables

- Just feel free to ignore it.

But why do you ignore it?

1.栈和输入合并后一直是右句式，规范句型
2.LR(0)：全行为reduce - rr冲突、rs冲突-》那么就不是LR(0)
3.SLR：A ->aB· ： follow (A) 才进行归约
4.上面两个表和LALR的大小一样
5.自动产生------yacc：LALR； lex



DFA：
LR0
SLR多个产生式
LR1

简洁，但又二义性

5. Ambiguity in LR Parsing

- Trade-off and consequence for ambiguity in LR parsing
  - Resolving ambiguities at the grammar level
    - Pros and cons? 简单，但是文法臃肿
  - Resolving ambiguities at the parsing table level
    - Pros and cons? 额外约束
  - Resolving ambiguities at the source code level
    - Pros and cons? 最灵活，最累

二义性有两种：
先天：语言的二义性
后天：表达式/文法的二义性

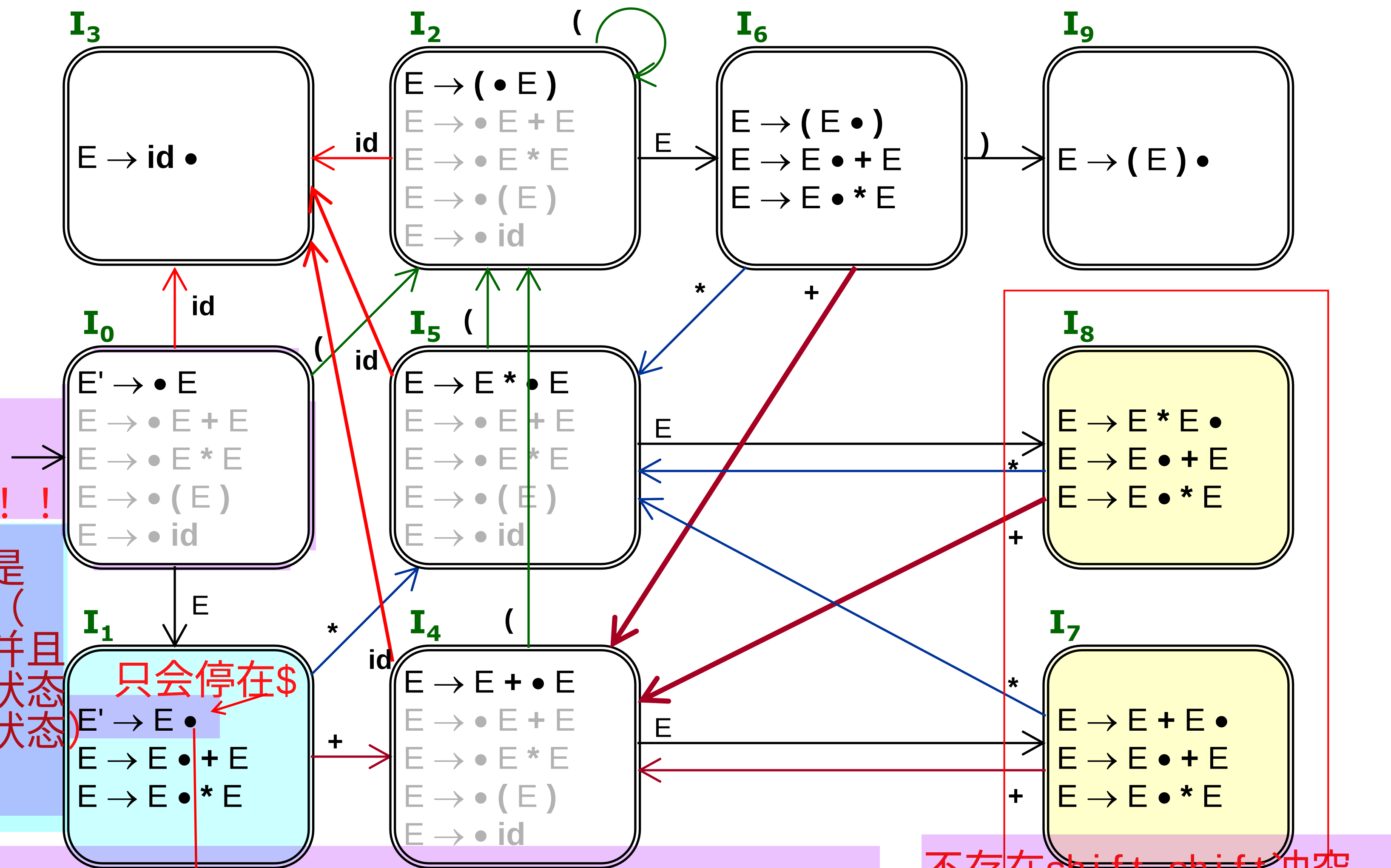
yacc：要求LALR，因此写DNF的时候需要改文法使之满足LALR文法。考试则不要求改造

Ambiguous Expression Grammar

- Given the ambiguous grammar

$$
\begin{array}{llll}
(0) & E' & \rightarrow & E \\
(1) & E & \rightarrow & E \mathbf{+} E \\
(2) & E & \rightarrow & E \mathbf{*} E \\
(3) & E & \rightarrow & \mathbf{(}\ E\ \mathbf{)} \\
(4) & E & \rightarrow & \mathbf{id}
\end{array}
$$

eg1



开始符号！！

每个圈都是画双线！（或者单线并且说明所有状态都是终结状态）

只会停在$

一般不认为存在无移进规约冲突，归约可以只填在$

不存在shift-shift冲突，因为此时就合并在一起了

DFA recognizing all viable prefixes

shift-reduce conflict

SLR(1) Parsing Table

State	ACTION						GOTO
	id	**+**	*****	**(**	**)**	**$**	E
0	s3			s2			1
1		s4	s5			acc	
2	s3			s2			6
3		r4	r4		r4	r4	
4	s3			s2			7
5	s3			s2			8
6		s4	s5		s9		
7		r1/s4	r1/s5		r1	r1	
8		r2/s4	r2/s5		r2	r2	
9		r3	r3		r3	r3	

在分析表上消除二义性

Resolve ambiguities at the parsing table level eg2

实际上的二义性：表达式和dangling

Dangling-**else** Grammar

- Given the ambiguous grammar

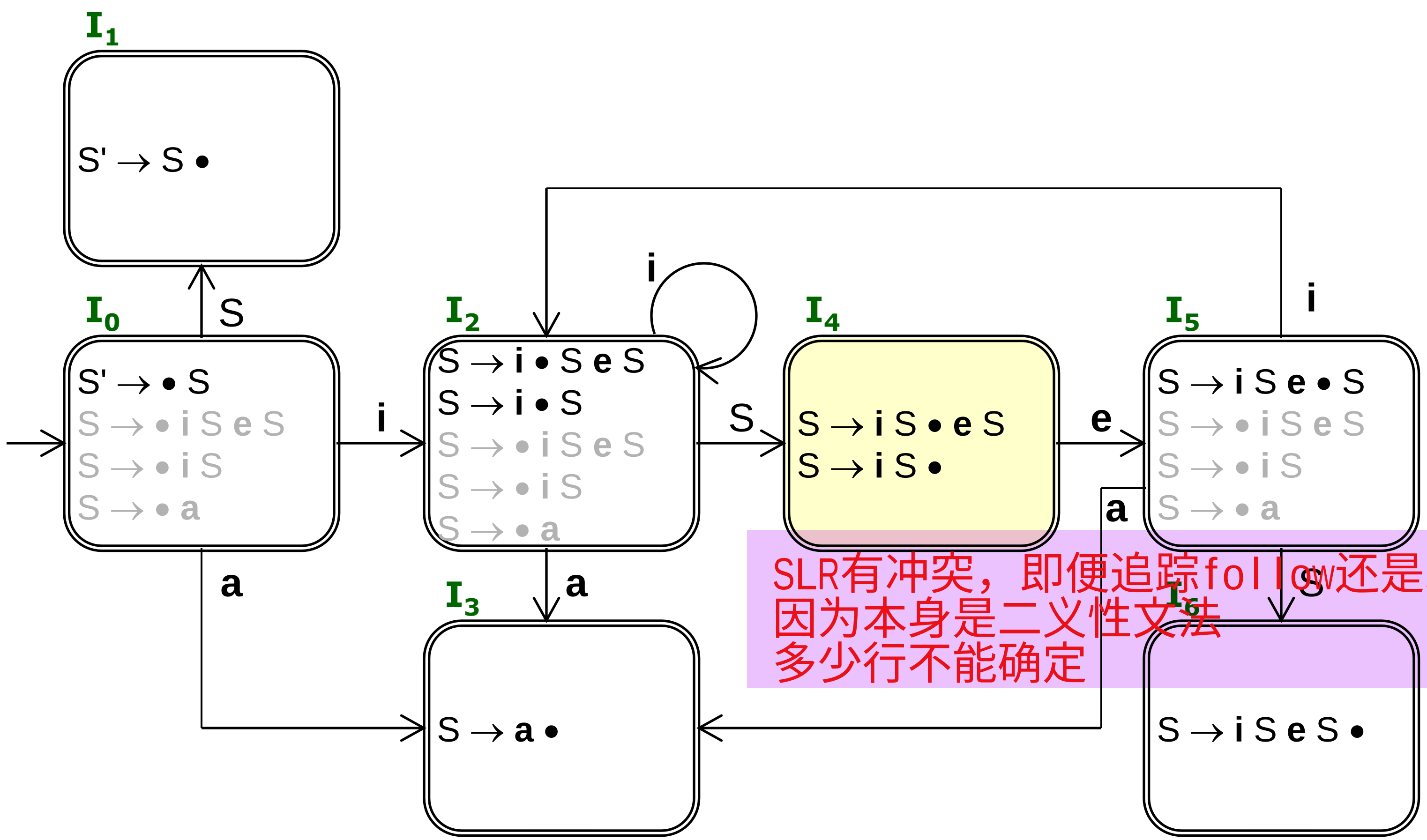
(0) $S' \rightarrow S$

(1) $S \rightarrow \mathbf{i}\ S\ \mathbf{e}\ S$

(2) $S \rightarrow \mathbf{i}\ S$

(3) $S \rightarrow \mathbf{a}$

DFA Recognizing All Viable Prefixes



SLR有冲突，即便追踪follow还是冲突，
因为本身是二义性文法
多少行不能确定

shift-reduce conflict

SLR(1) Parsing Table

State	ACTION				GOTO
	i	**e**	**a**	**$**	S
0	s2		s3		1
1				acc	
2	s2		s3		4
3		r3		r3	
4		r2/s5		r2	
5	s2		s3		6
6		r1		r1	

Resolve ambiguities
at the parsing table level

退回多少格，进入新状态的时候容易出现错误
LR(1)很重点

Parsing a Sentence

Step	Symbol	State	Input	Reference	Action	Output
1	**$**	0	**i i a e a $**	a[0, **i**] = s2	shift	
2	**$ i**	0 2	**i a e a $**	a[2, **i**] = s2	shift	
3	**$ i i**	0 2 2	**a e a $**	a[2, **a**] = s3	shift	
4	**$ i i a**	0 2 2 3	**e a $**	a[3, **e**] = r3 g[2, S] = 4	reduce	S → **a**
5	**$ i i** S	0 2 2 4	**e a $**	a[4, **e**] = s5	shift	
6	**$ i i** S **e**	0 2 2 4 5	**a $**	a[5, **a**] = s3	shift	
7	**$ i i** S **e a**	0 2 2 4 5 3	**$**	a[3, **$**] = r3 g[5, S] = 6	reduce	S → **a**
8	**$ i i** S **e** S	0 2 2 4 5 6	**$**	a[6, **$**] = r1 g[2, S] = 4	reduce	S → **i** S **e** S
9	**$ i** S	0 2 4	**$**	a[4, **$**] = r2 g[0, S] = 1	reduce	S → **i** S
10	**$** S	0 1	**$**	a[1, **$**] = acc	accept	

6. Error Recovery in LR Parsing

- For errors in an input sentence
  - All errors will be found by all LR parsers
    - If we can construct a parsing table without conflicts.
  - More reductions before reporting an error
    - LR(0) $\geq$ SLR(1) $\geq$ LALR(1) $\geq$ LR(1)
    - But all of them will never shift an erroneous input symbol onto the stack.

错误都会被找出来，
只是时间不同
是LR0就必定是后三者
是SLR就必定是后两者
是LALR就必定是LR1

Review: Error Recovery Techniques

- Panic-Mode Error Recovery
  - Skip inputs until synchronizing token found.
- Phrase-Level Error Recovery
  - Assign each empty entry a specific error routine. 内部节点就是phrase
- Error-Productions 精准报错与恢复，需要事先直到出错模式 耗时较大且性价比不高
  - Suitable for common errors but not all errors.
- Global-Correction
  - Globally analyze the input to find the error.
  - Expensive and not in practice.

考试不好考

Panic-Mode Error Recovery

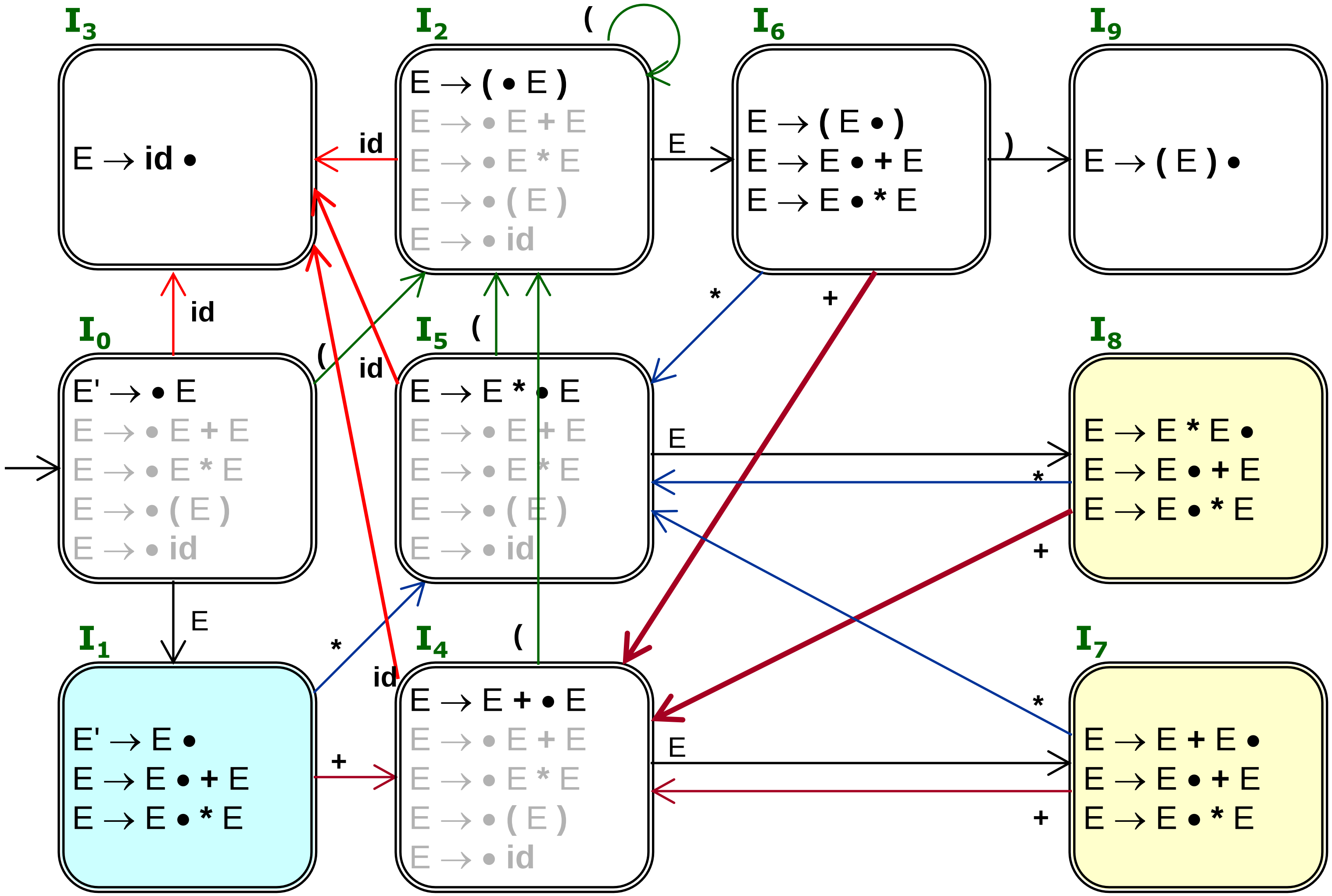
一路退到有接受A出去的状态为止

- Pop until a state **s** with a GOTO() on a particular nonterminal **A** is found.
  - Usually **A** represents major constructs, e.g. *stmt*, *expr*, or *block*.
  - It indicates that the construct A has errors.
- Push GOTO(**s**, **A**).
- 0 or more lookaheads are then discarded, until a symbol that can follow **A**.

It is easier to understand if we look from the viewpoint of DFA

Phrase-Level Error Recovery

- Examine each empty entry, and assign it a pointer to a specific error-handling routine. 见后面的表格
- Ad hoc: depending on the usage of the language.



The Previous DFA recognizing all viable prefixes phrase level

The Previous Parsing Table

n+m
除$外的共同
构成GOTO表

State	ACTION n+1						GOTO
	id	**+**	*****	**(**	**)**	**$**	E
0	s3	e1	e1	s2	e2	e1	1
1	e3	s4	s5	e3	e2	acc	
2	s3	e1	e1	s2	e2	e1	6
3	r4	r4	r4	r4	r4	r4	
4	s3	e1	e1	s2	e2	e1	7
5	s3	e1	e1	s2	e2	e1	8
6	e3	s4	s5	e3	s9	e4	
7	r1	r1	s5	r1	r1	r1	
8	r2	r2	r2	r2	r2	r2	
9	r3	r3	r3	r3	r3	r3	

全部填满去reduce还是会报错

Postpone error detection until one or more reductions are made

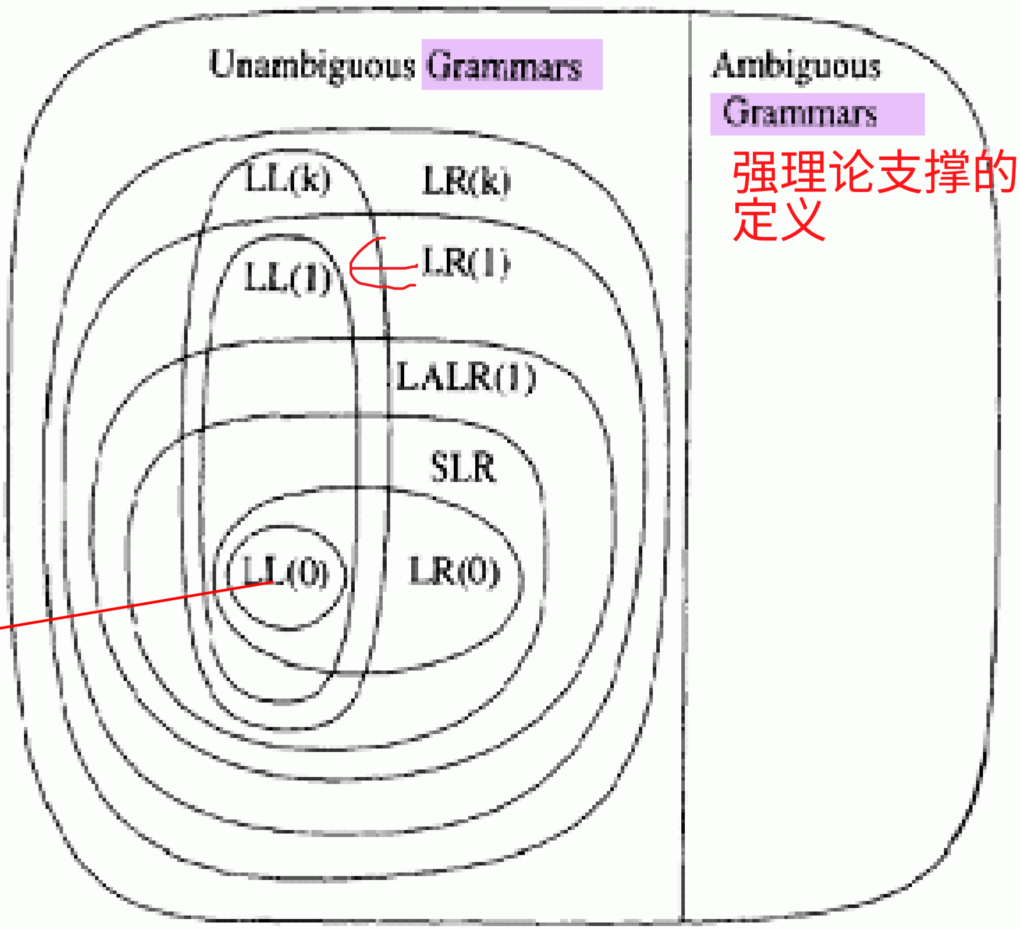
Error-Handling Routines

- e1: an operand ('id' or '(') is expected.
  - `push state 3;  // add a symbol 'id'`
- e2: unbalanced right parenthesis.
  - `drop one lookahead;  // remove ')'`
- e3: an operator is expected.
  - `push state 4;  // add a symbol '+'`
- e4: a right parenthesis is expected.
  - `push state 9;  // add a symbol ')'`

Parsing an Erroneous Input

Step	Symbol	State	Input	Reference	Action	Output
1	**$**	0	**id +) $**	a[0, **id**] = s3	shift	
2	**$ id**	0 3	**+) $**	a[3, **+**] = r4 g[0, E] = 1	reduce	E → **id**
3	**$** E	0 1	**+) $**	a[1, **+**] = s4	shift	
4	**$** E **+**	0 1 4	**) $**	a[4, **)**] = e2	drop	Unbalanced ')'
5	**$** E **+**	0 1 4	**$**	a[4, **$**] = e1	push 3	Operand expected
6	**$** E **+ id**	0 1 4 3	**$**	a[3, **$**] = r4 g[4, E] = 7	reduce	E → **id**
7	**$** E **+** E	0 1 4 7	**$**	a[7, **$**] = r1 g[0, E] = 1	reduce	E → E **+** E
8	**$** E	0 1	**$**	a[1, **$**] = acc	end	

Conclusions:
Context-Free Grammar Classification



强理论支撑的都是易错点定义

LR能处理的更多2
LL中没有状态的概念

用固定的产生式derive

Exercise 7.1

- Consider the grammar

$$S \rightarrow \mathbf{(}\ S\ R \mid \mathbf{a}$$

$$R \rightarrow \mathbf{,}\ S\ R \mid \mathbf{)}$$

  - Try to construct an SLR(1) parsing table for the grammar, and see if there are conflicts in the parsing table.

要求画才画

Exercise 7.2

- Consider the grammar

$$S \rightarrow S\,\mathbf{a}\,\mathbf{b} \mid \mathbf{a}\,R$$

$$R \rightarrow S \mid \mathbf{a}$$

Is the grammar an SLR(1) grammar? and why?

DFA画出说哪里有就行
画局部 ->...
哪里有冲突

Exercise 7.3

- Consider the grammar

$$S \rightarrow A$$

$$A \rightarrow B\,A \mid \varepsilon$$

$$B \rightarrow \mathbf{a}\,B \mid \mathbf{b}$$

  - Prove that the grammar is an LR(1) grammar.
  - Construct an LR(1) parsing table for the grammar.
  - Show the detailed parsing procedure for the sentence **abab**, following the style in slides of this lecture.

Exercise 7.4*

- (DBv2, pp.278, ex.4.7.4) Show that the grammar

  $$S \rightarrow A\,\mathbf{a} \mid \mathbf{b}\,A\,\mathbf{c} \mid \mathbf{d}\,\mathbf{c} \mid \mathbf{b}\,\mathbf{d}\,\mathbf{a}$$

  $$A \rightarrow \mathbf{d}$$

  is LALR(1) but not SLR(1).

  不用画表，局部DFA给出结论就好

最核心的内容结束

11-12 test全部内容，题目会难的，出个难的

Further Reading

- Dragon Book, 2nd Edition (DBv2)
  - Comprehensive Reading:
    - Section 4.6 on SLR(1) parsing.
    - Section 4.7 on LR(1) and LALR(1) parsing.
    - Section 4.8.1-4.8.2 on ambiguities in LR parsing.
  - Skip Reading:
    - Section 4.8.3 on error recovery in LR parsing.

Enjoy the Course!