CS 4300: Compiler Theory

Chapter 3
Lexical Analysis

Dr. Xuejun Liang

Outlines (Sections)

1. The Role of the Lexical Analyzer
2. Input Buffering (Omit)
3. Specification of Tokens
4. Recognition of Tokens
5. The Lexical -Analyzer Generator Lex
6. Finite Automata
7. From Regular Expressions to Automata
8. Design of a Lexical-Analyzer Generator
9. Optimization of DFA-Based Pattern Matchers*

Quick Review of Last Lecture

- The Lexical-Analyzer Generator Lex
  - Structure of Lex Programs
  - Regular Expressions in Lex
  - Example Lex Specification
  - Conflict Resolution in Lex
- Finite Automata
  - Definitions of NFA and DFA
  - Transition Graph, Transition Table
  - The Language Defined by an NFA and DFA
  - Simulate a DFA

7. From Regular Expressions to Automata
Conversion of an NFA into a DFA

- The ***subset construction*** algorithm converts an NFA into a DFA using:
  - ε-*closure*(s) = $\{s\} \cup \{ t \mid s \rightarrow_\varepsilon \ldots \rightarrow_\varepsilon t\}$
  - ε-*closure*(T) = $\cup_{s \in T}\, \varepsilon$-*closure*($s$)
  - *move*(T, a) = $\{ s \mid t \rightarrow_a s$ and $t \in T\}$
- The algorithm produces:
  - ***Dstates*** -- the set of states of the new DFA consisting of sets of states of the NFA
  - ***Dtran*** -- the transition table of the new DFA

The Subset Construction Algorithm

Initially, ε-*closure*(s_0) is the only state in *Dstates* and it is unmarked

while (there is an unmarked state *T* in *Dstates)* **{**

    mark *T*

    for (each input symbol $a \in \Sigma$) {

        U = ε-*closure*(*move*(*T*,*a*))

        if (*U* is not in *Dstates)*

            add *U* as an unmarked state to *Dstates*

        Dtran[*T*,*a*] := *U*

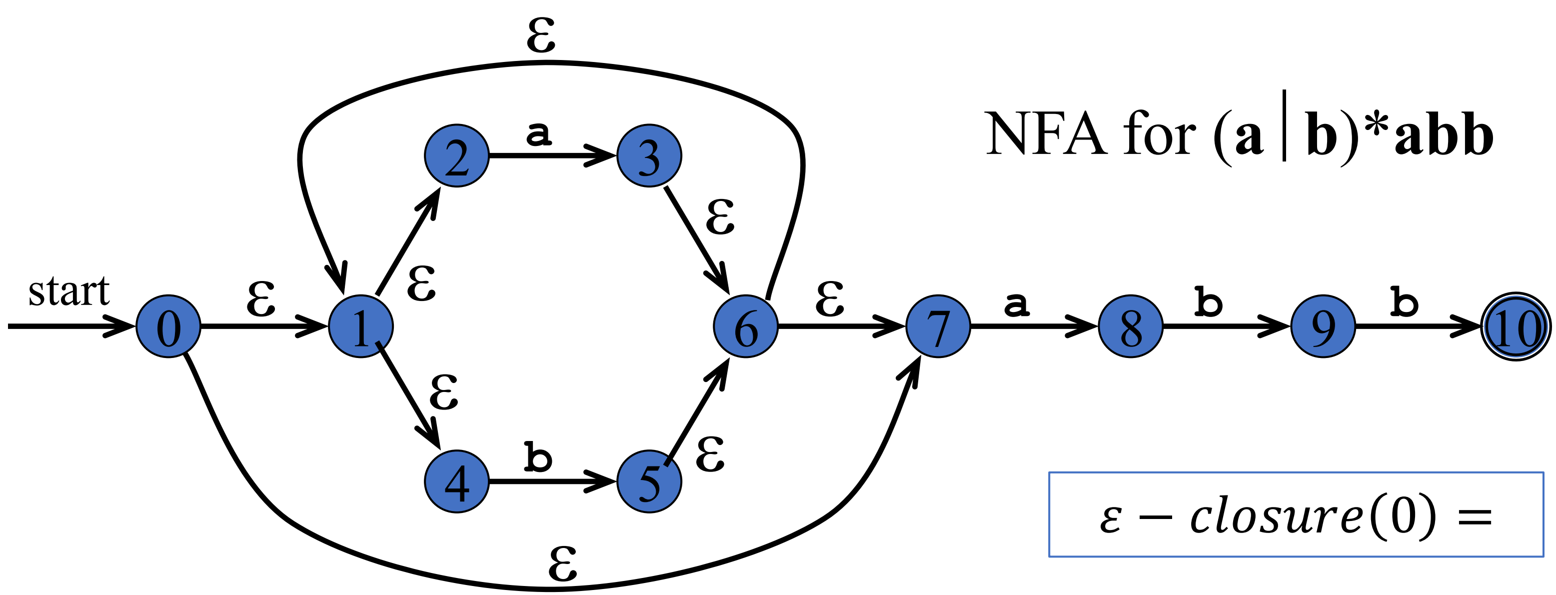
    }

}

Computing ε-closure(*T*)

```
push all states of T onto stack;
initialize ε-closure(T) to T;
while ( stack is not empty ) {
        pop t, the top element, off stack;
        for ( each state u with an edge from t to u labeled ε )
                if ( u is not in ε-closure(T) ) {
                        add u to ε-closure(T) ;
                        push u onto stack;
                }
}
```

Subset Construction Example 1



NFA for (**a** | **b**)***abb**

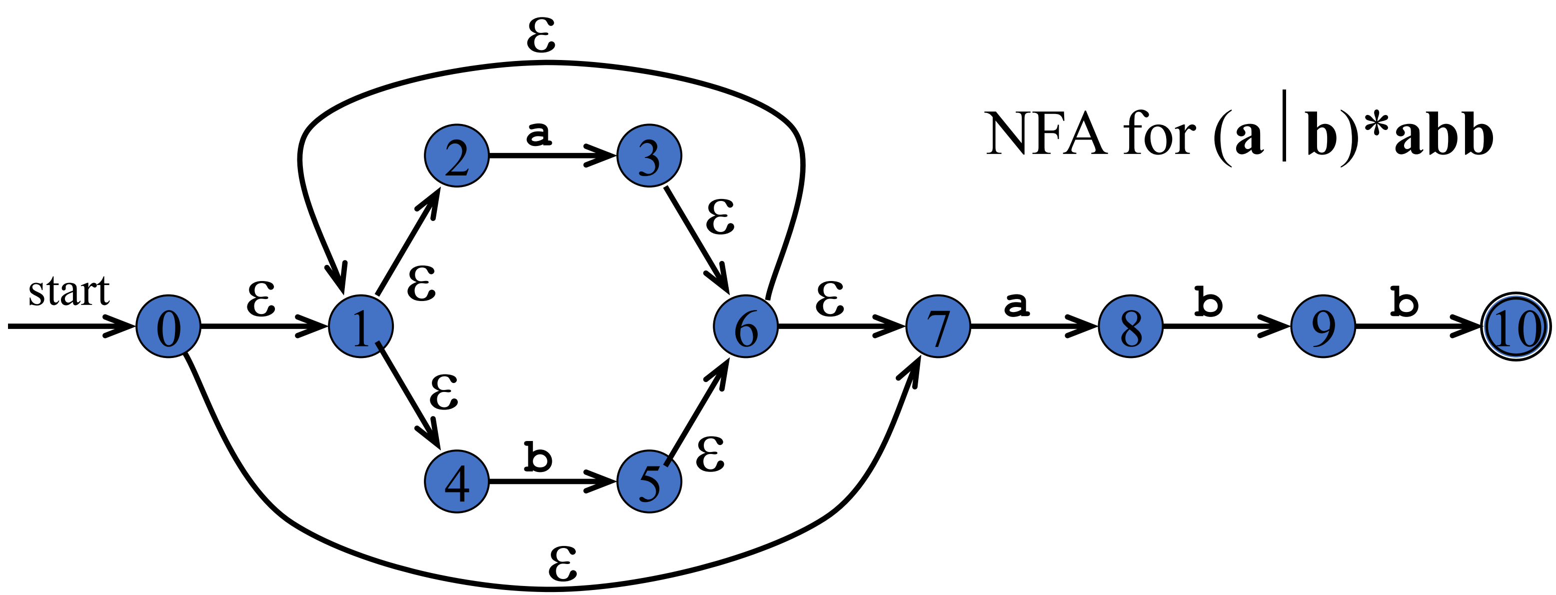
$\varepsilon - closure(0) =$

NFA STATE	DFA STATE	a	b
$\{0,1,2,4,7\}$	A	B	C
$\{1,2,3,4,6,7,8\}$	B	B	D
$\{1,2,4,5,6,7\}$	C	B	C
$\{1,2,4,5,6,7,9\}$	D	B	E
$\{1,2,3,5,6,7,10\}$	E	B	C

$move(A, a) =$

$\varepsilon - closure\big(move(A, a)\big) =$

Subset Construction Example 1



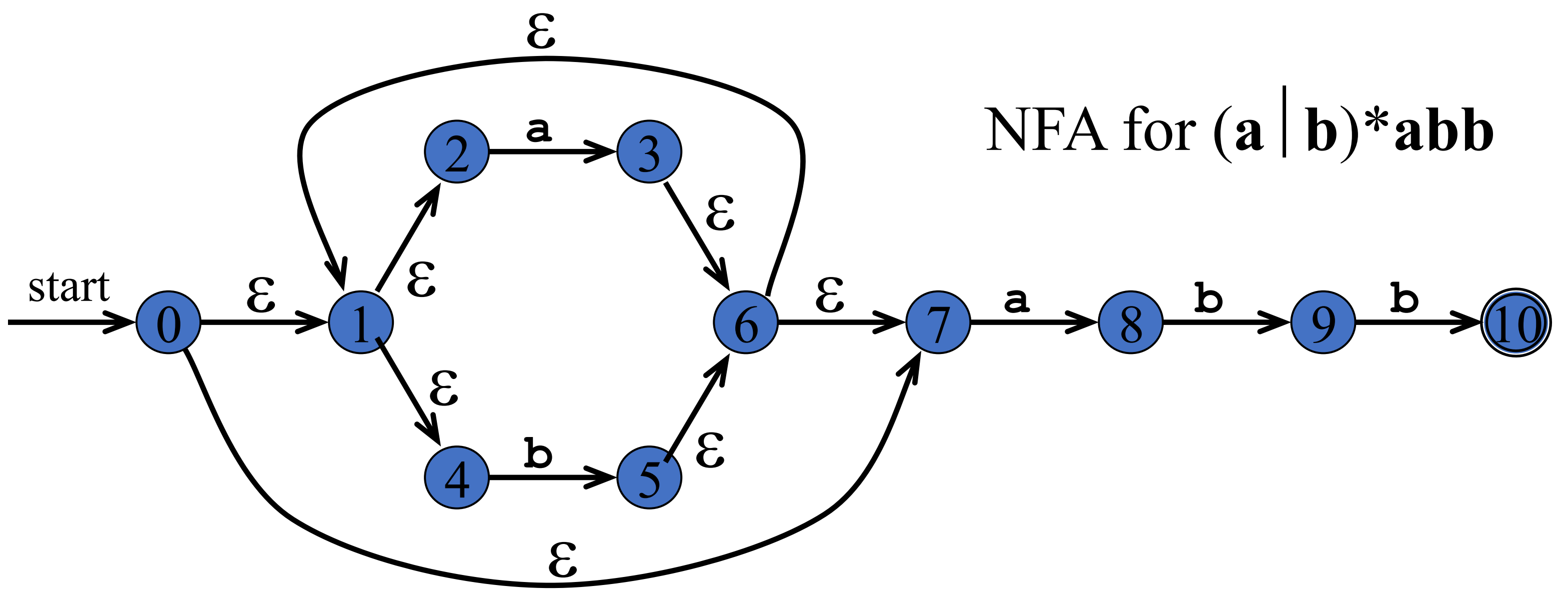
NFA for (**a** | **b**)***abb**

NFA STATE	DFA STATE	a	b
$\{0,1,2,4,7\}$	A	B	C
$\{1,2,3,4,6,7,8\}$	B	B	D
$\{1,2,4,5,6,7\}$	C	B	C
$\{1,2,4,5,6,7,9\}$	D	B	E
$\{1,2,3,5,6,7,10\}$	E	B	C

$move(A, b) =$

$\varepsilon - closure\big(move(A, b)\big) =$

Subset Construction Example 1



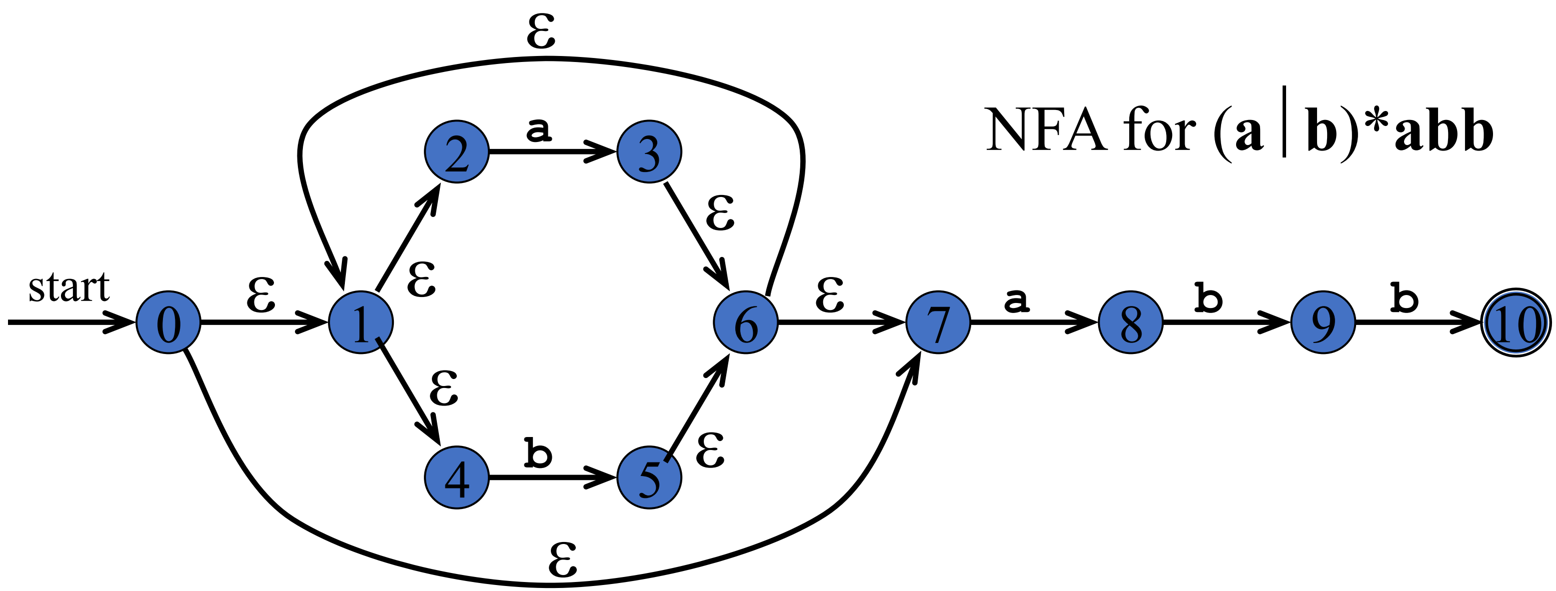
NFA for (**a** | **b**)***abb**

NFA STATE	DFA STATE	a	b
$\{0,1,2,4,7\}$	A	B	C
$\{1,2,3,4,6,7,8\}$	B	B	D
$\{1,2,4,5,6,7\}$	C	B	C
$\{1,2,4,5,6,7,9\}$	D	B	E
$\{1,2,3,5,6,7,10\}$	E	B	C

$move(B, a) =$

$\varepsilon - closure(move(B, a)) =$

Subset Construction Example 1



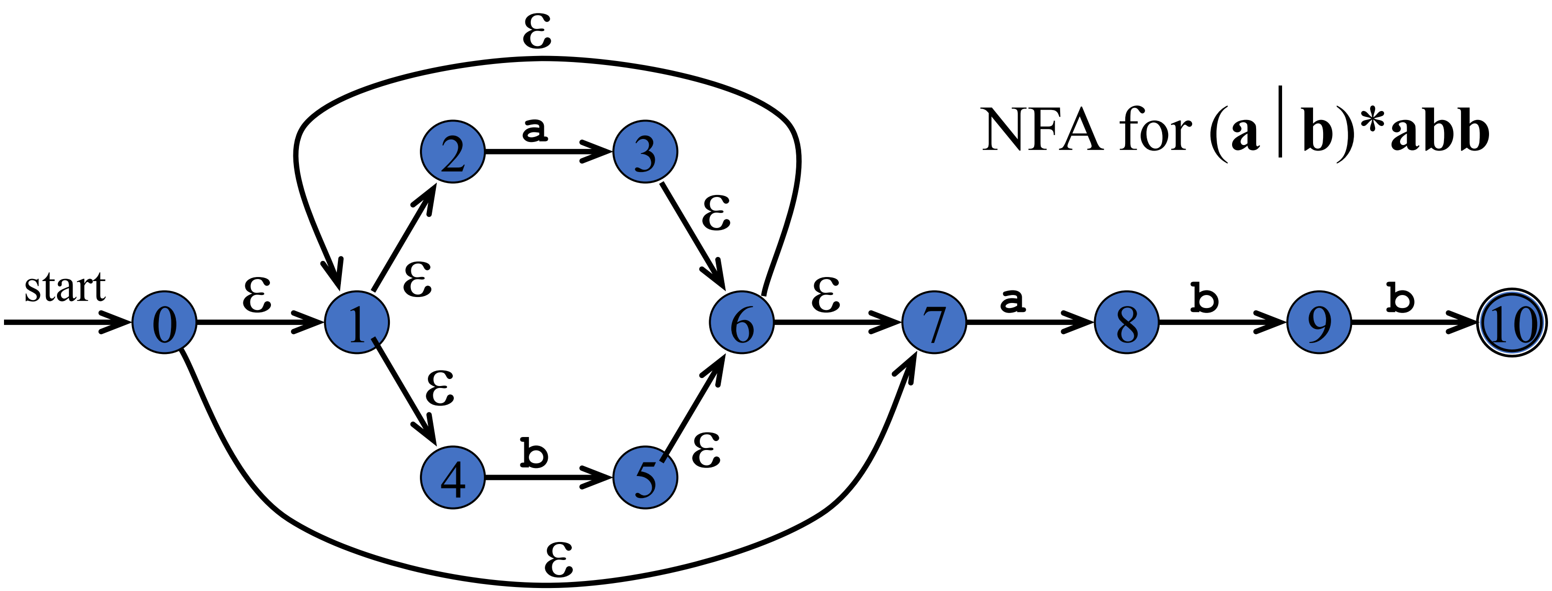
NFA for (**a** | **b**)***abb**

NFA STATE	DFA STATE	a	b
$\{0, 1, 2, 4, 7\}$	A	B	C
$\{1, 2, 3, 4, 6, 7, 8\}$	B	B	D
$\{1, 2, 4, 5, 6, 7\}$	C	B	C
$\{1, 2, 4, 5, 6, 7, 9\}$	D	B	E
$\{1, 2, 3, 5, 6, 7, 10\}$	E	B	C

$move(B, b) =$

$\varepsilon - closure(move(B, b)) =$

Subset Construction Example 1



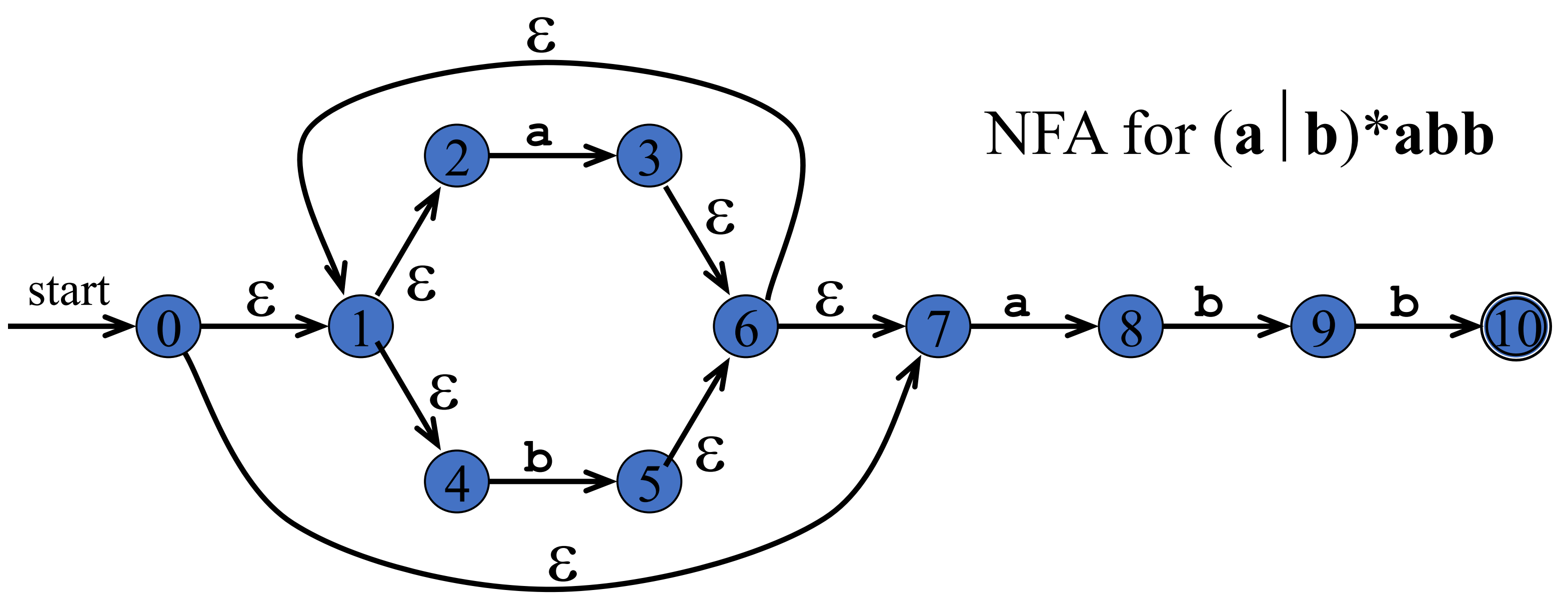
NFA for (**a** | **b**)***abb**

NFA State	DFA State	a	b
$\{0, 1, 2, 4, 7\}$	A	B	C
$\{1, 2, 3, 4, 6, 7, 8\}$	B	B	D
$\{1, 2, 4, 5, 6, 7\}$	C	B	C
$\{1, 2, 4, 5, 6, 7, 9\}$	D	B	E
$\{1, 2, 3, 5, 6, 7, 10\}$	E	B	C

$move(C, a) =$

$\varepsilon - closure\big(move(C, a)\big) =$

Subset Construction Example 1



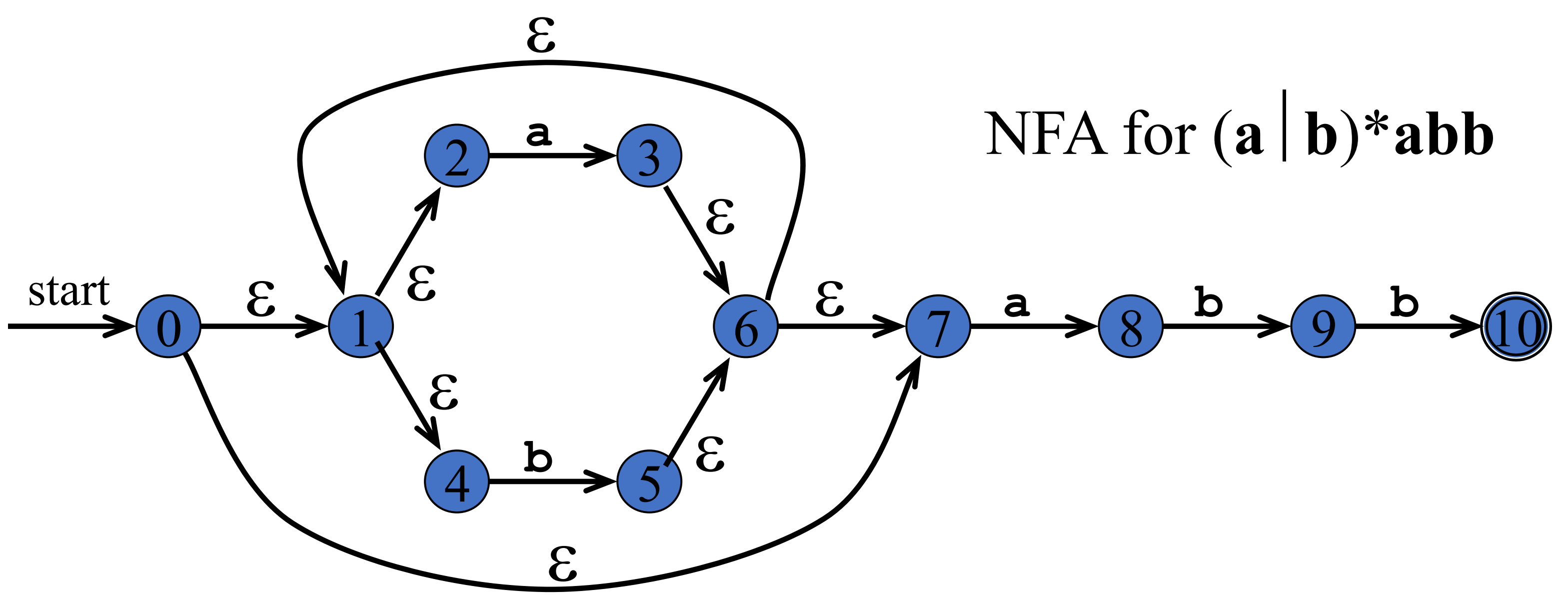
NFA for (**a** | **b**)***abb**

NFA State	DFA State	a	b
$\{0,1,2,4,7\}$	A	B	C
$\{1,2,3,4,6,7,8\}$	B	B	D
$\{1,2,4,5,6,7\}$	C	B	C
$\{1,2,4,5,6,7,9\}$	D	B	E
$\{1,2,3,5,6,7,10\}$	E	B	C

$move(C, b) =$

$\varepsilon - closure\big(move(C, b)\big) =$

Subset Construction Example 1



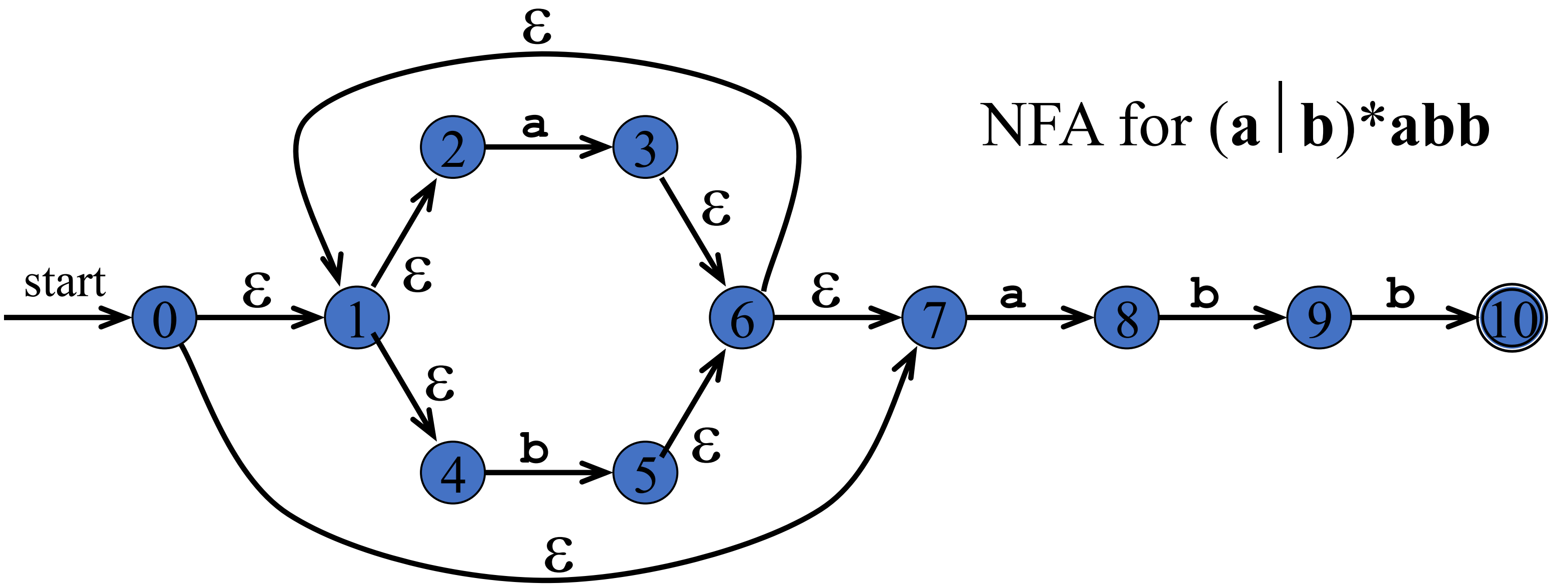
NFA for (**a** | **b**)***abb**

NFA STATE	DFA STATE	a	b
$\{0,1,2,4,7\}$	A	B	C
$\{1,2,3,4,6,7,8\}$	B	B	D
$\{1,2,4,5,6,7\}$	C	B	C
$\{1,2,4,5,6,7,9\}$	D	B	E
$\{1,2,3,5,6,7,10\}$	E	B	C

$move(D, a) =$

$\varepsilon - closure(move(D, a)) =$

Subset Construction Example 1



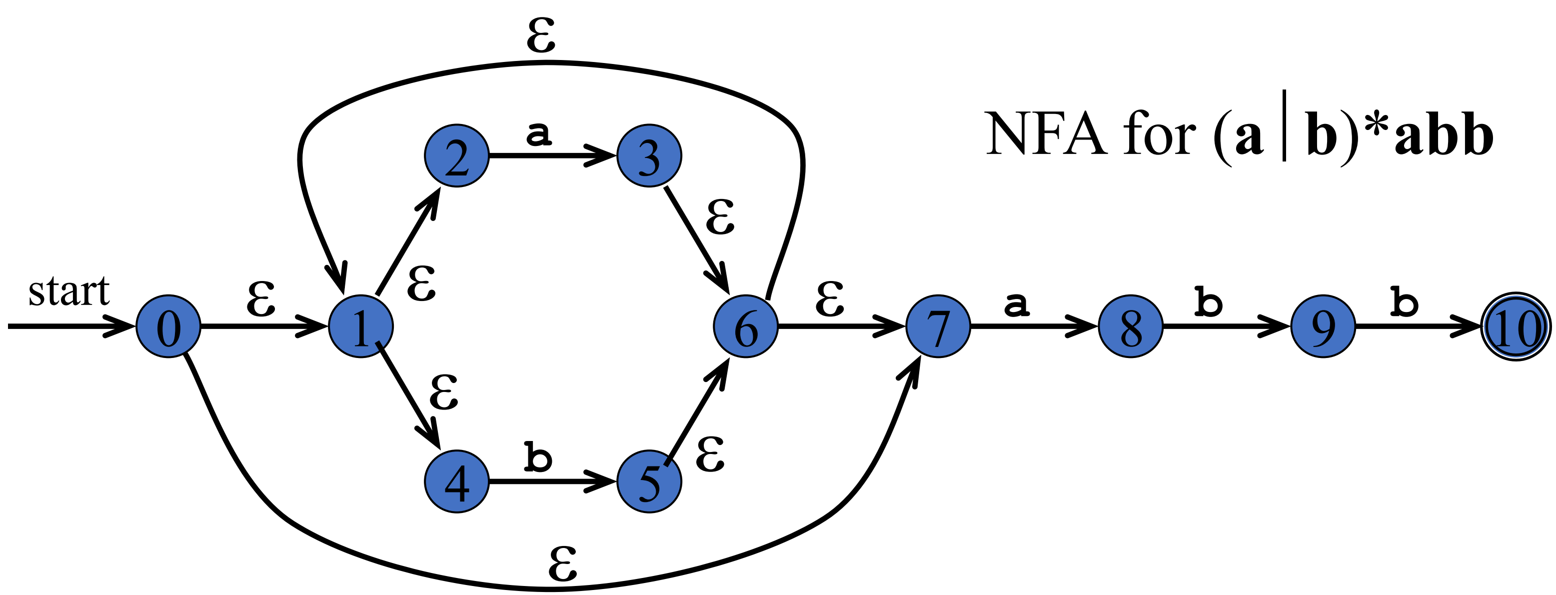
NFA for (**a** | **b**)***abb**

NFA State	DFA State	a	b
$\{0,1,2,4,7\}$	A	B	C
$\{1,2,3,4,6,7,8\}$	B	B	D
$\{1,2,4,5,6,7\}$	C	B	C
$\{1,2,4,5,6,7,9\}$	D	B	E
$\{1,2,3,5,6,7,10\}$	E	B	C

$move(D, b) =$

$\varepsilon - closure(move(D, b)) =$

Subset Construction Example 1



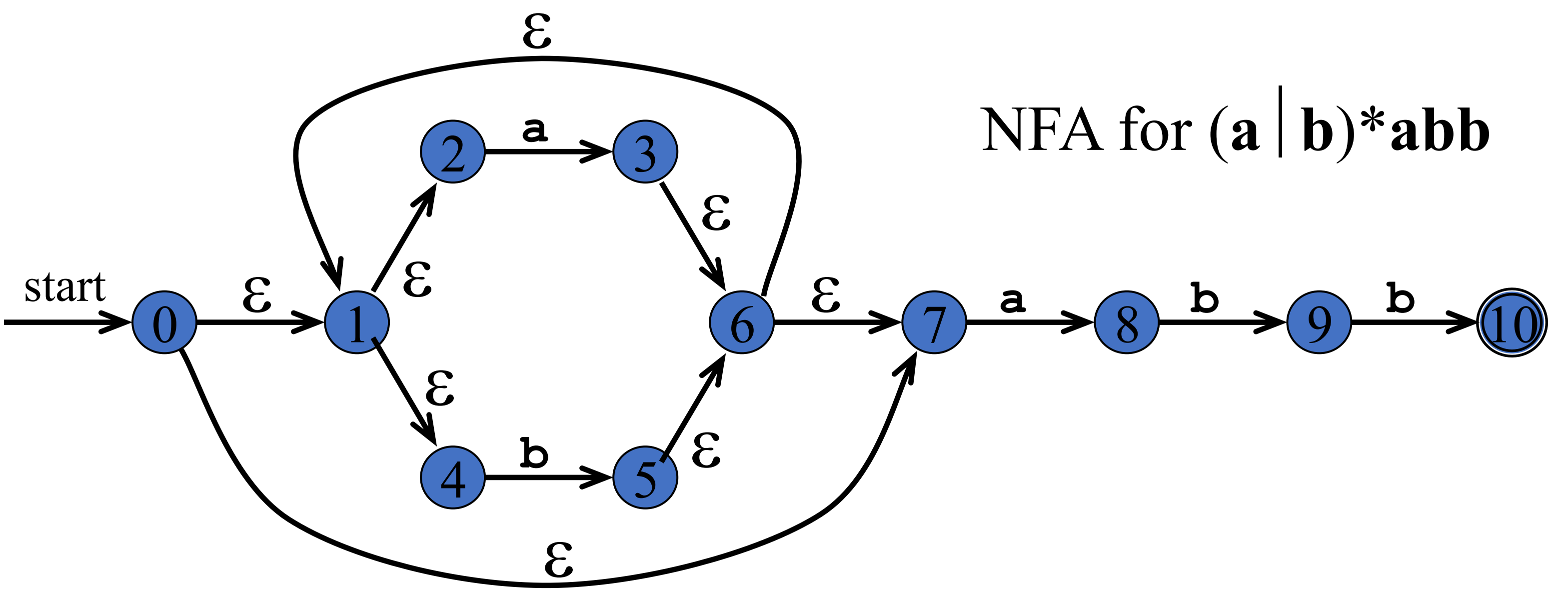
NFA for (**a** | **b**)***abb**

NFA State	DFA State	a	b
$\{0,1,2,4,7\}$	A	B	C
$\{1,2,3,4,6,7,8\}$	B	B	D
$\{1,2,4,5,6,7\}$	C	B	C
$\{1,2,4,5,6,7,9\}$	D	B	E
$\{1,2,3,5,6,7,10\}$	E	B	C

$move(E, a) =$

$\varepsilon - closure(move(E, a)) =$

Subset Construction Example 1



NFA for $(\mathbf{a} \mid \mathbf{b})^*\mathbf{abb}$

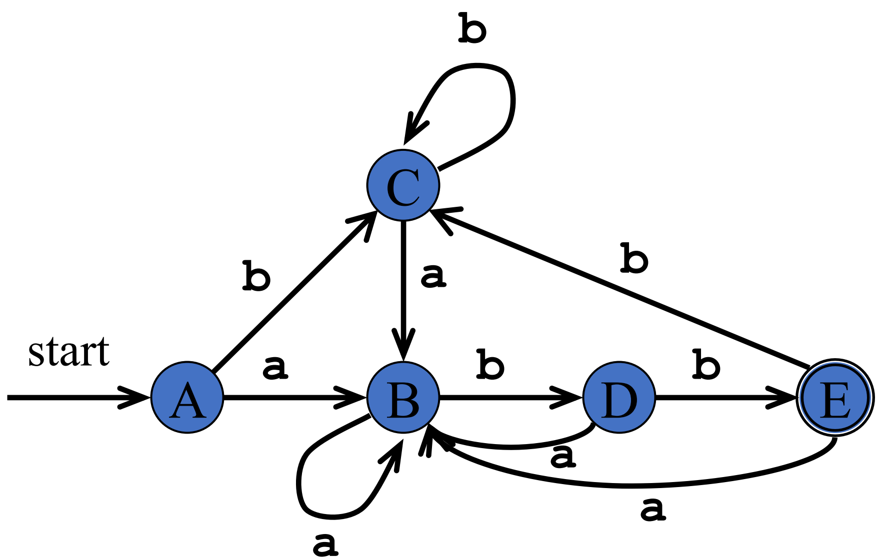
NFA STATE	DFA STATE	a	b
$\{0, 1, 2, 4, 7\}$	A	B	C
$\{1, 2, 3, 4, 6, 7, 8\}$	B	B	D
$\{1, 2, 4, 5, 6, 7\}$	C	B	C
$\{1, 2, 4, 5, 6, 7, 9\}$	D	B	E
$\{1, 2, 3, 5, 6, 7, 10\}$	E	B	C

$move(E, b) =$

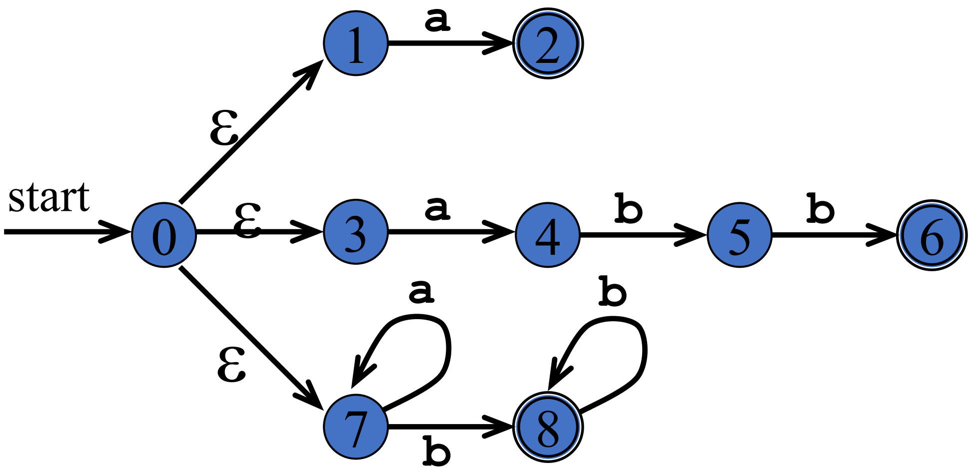
$\varepsilon - closure\big(move(E, b)\big) =$

Subset Construction Example 1 Cont.

NFA State	DFA State	a	b
$\{0,1,2,4,7\}$	A	B	C
$\{1,2,3,4,6,7,8\}$	B	B	D
$\{1,2,4,5,6,7\}$	C	B	C
$\{1,2,4,5,6,7,9\}$	D	B	E
$\{1,2,3,5,6,7,10\}$	E	B	C



Subset Construction Example 2



NFA State	DFA State	a	b
{0,1,3,7}	A	B	C
{2,4,7}	**B**	D	E
{8}	**C**	∅	C
{7}	D	D	C
{5,8}	**E**	∅	F
{6,8}	**F**	∅	C

$\varepsilon\text{-}closure(\{0\}) = \{0,1,3,7\}$ A

A

$move(\{0,1,3,7\},\mathbf{a}) = \{2,4,7\}$

$\varepsilon\text{-}closure(\{2,4,7\}) = \{2,4,7\}$ B

B

$move(\{2,4,7\},\mathbf{a}) = \{7\}$

$\varepsilon\text{-}closure(\{7\}) = \{7\}$ D

D

$move(\{7\},\mathbf{b}) = \{8\}$

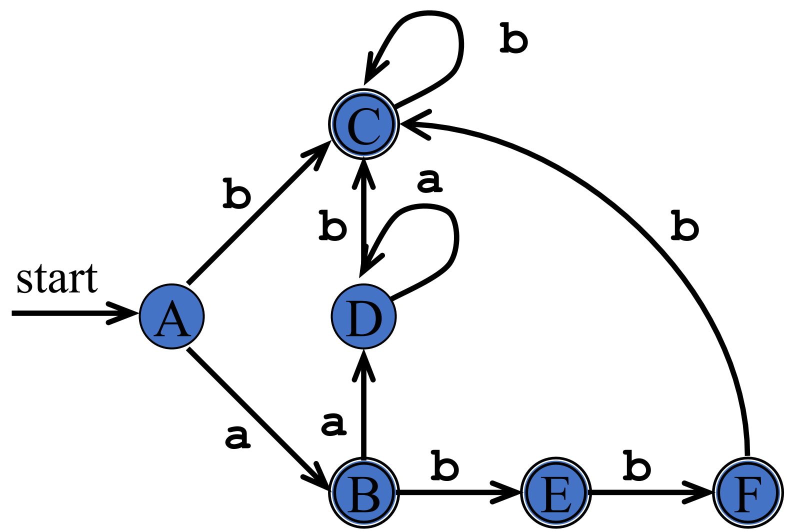
$\varepsilon\text{-}closure(\{8\}) = \{8\}$ C

C

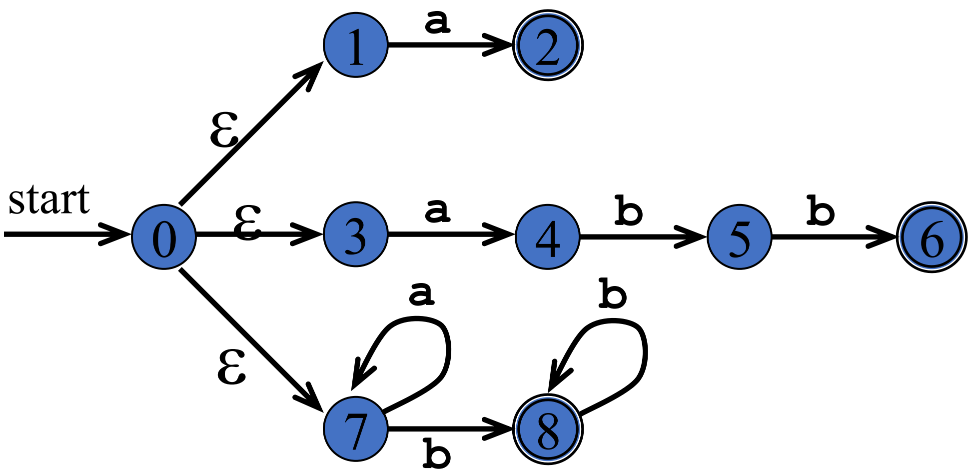
$move(\{8\},\mathbf{a}) = \emptyset$

Subset Construction Example 2 Cont.

NFA State	DFA State	a	b
{0,1,3,7}	A	B	C
{2,4,7}	**B**	D	E
{8}	**C**	Ø	C
{7}	D	D	C
{5,8}	**E**	Ø	F
{6,8}	**F**	Ø	C



ε-*closure* and *move* Examples



$\varepsilon\text{-}closure(\{0\}) = \{0,1,3,7\}$ A

$move(\{0,1,3,7\},\mathbf{a}) = \{2,4,7\}$

$\varepsilon\text{-}closure(\{2,4,7\}) = \{2,4,7\}$ B

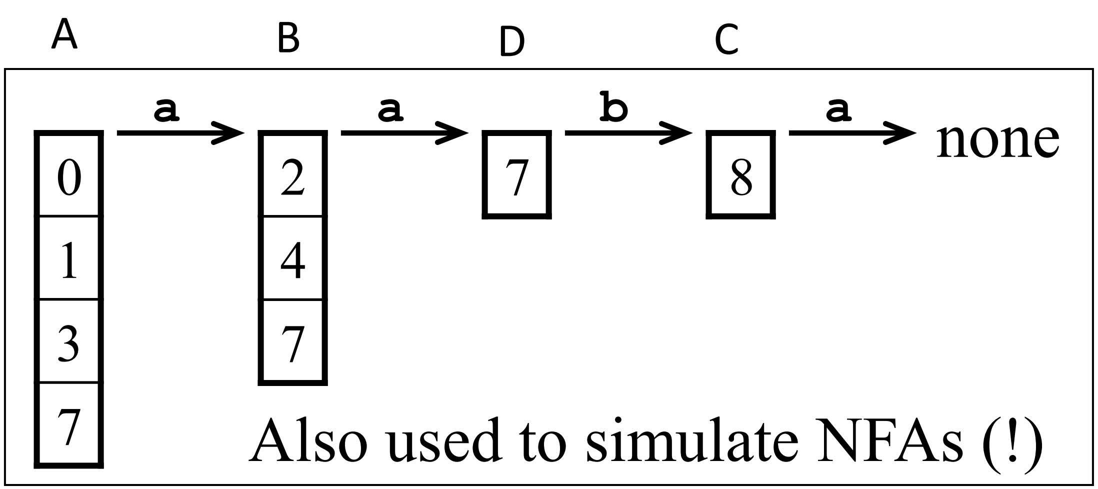
$move(\{2,4,7\},\mathbf{a}) = \{7\}$

$\varepsilon\text{-}closure(\{7\}) = \{7\}$ D

$move(\{7\},\mathbf{b}) = \{8\}$

$\varepsilon\text{-}closure(\{8\}) = \{8\}$ C

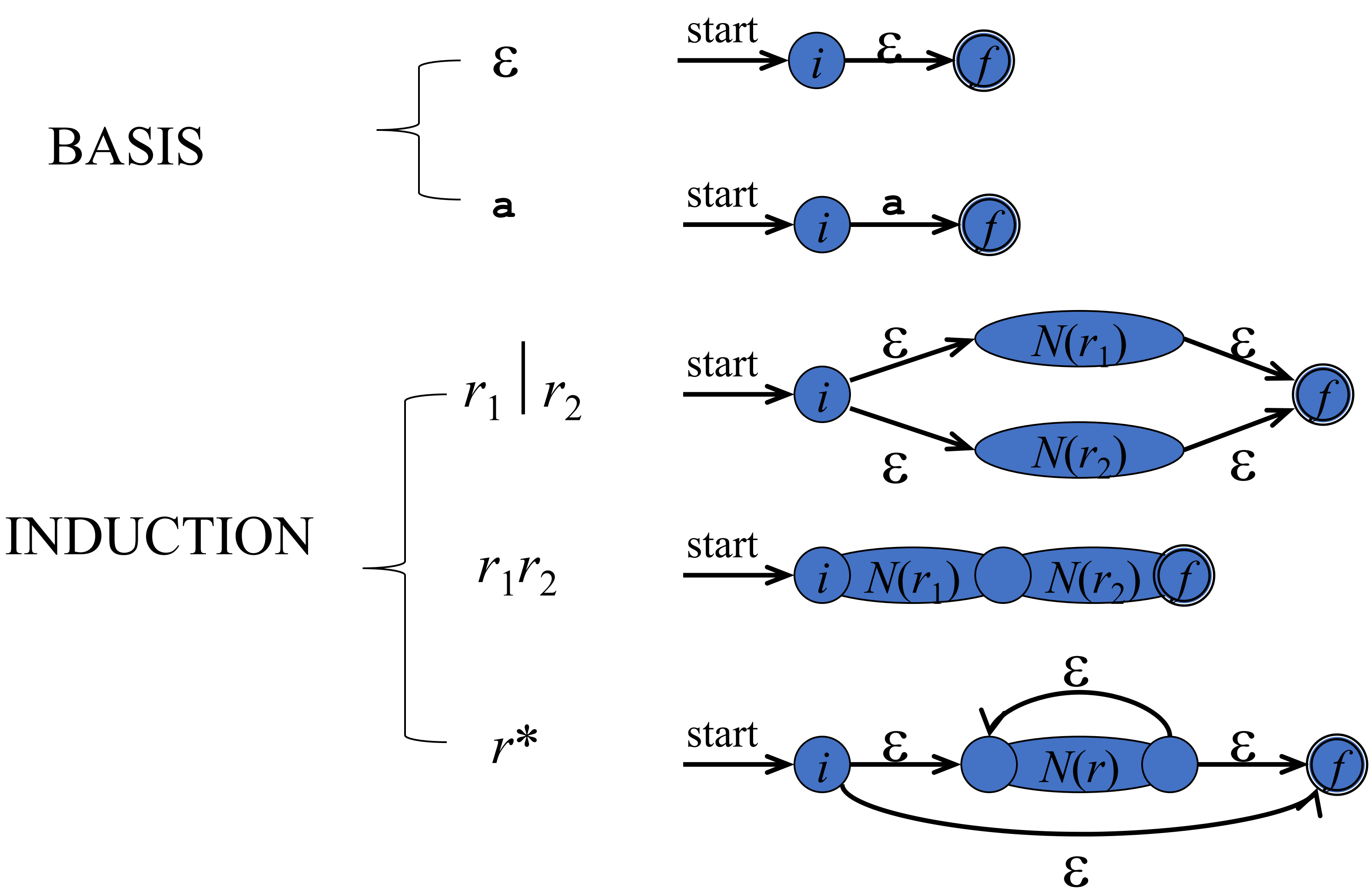
$move(\{8\},\mathbf{a}) = \varnothing$



Simulating an NFA Using ε-*closure* and *move*

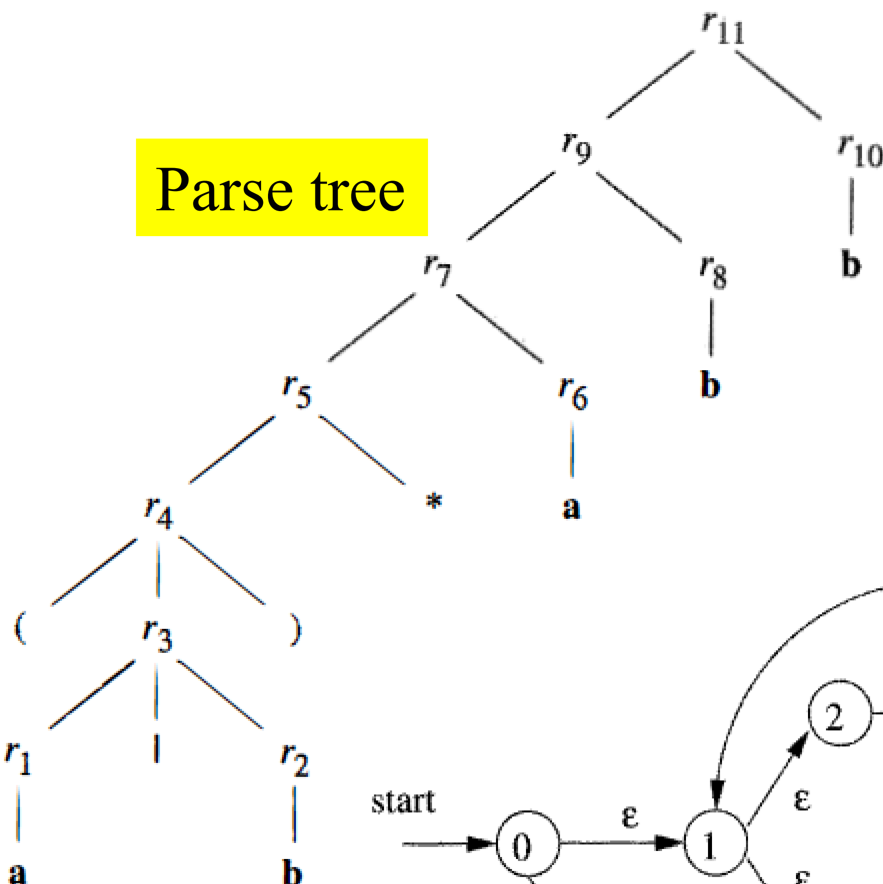
$$
\begin{array}{l}
S = \epsilon\text{-}closure(s_0); \\
c = nextChar(); \\
\textbf{while} \ (\ c \ != \ \textbf{eof} \ ) \ \{ \\
\qquad S = \epsilon\text{-}closure\big(move(S, c)\big); \\
\qquad c = nextChar(); \\
\} \\
\textbf{if} \ (\ S \cap F \ != \ \emptyset \ ) \ \textbf{return "yes"}; \\
\textbf{else return "no"};
\end{array}
$$

From Regular Expression to NFA (Thompson’s Construction)



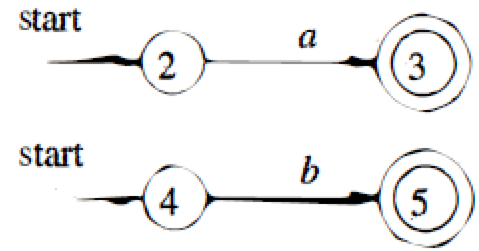
Construct an NFA for r = (a|b)*abb

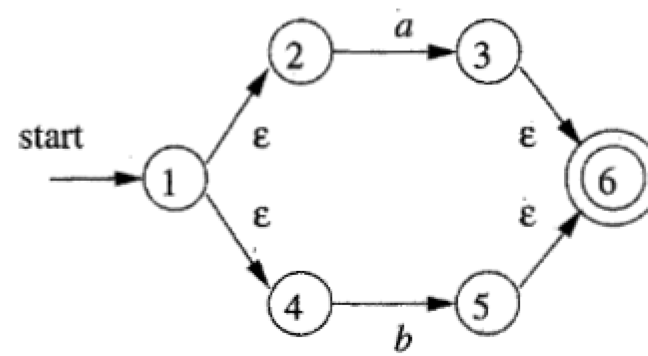
Parse tree



$r_1 = a$

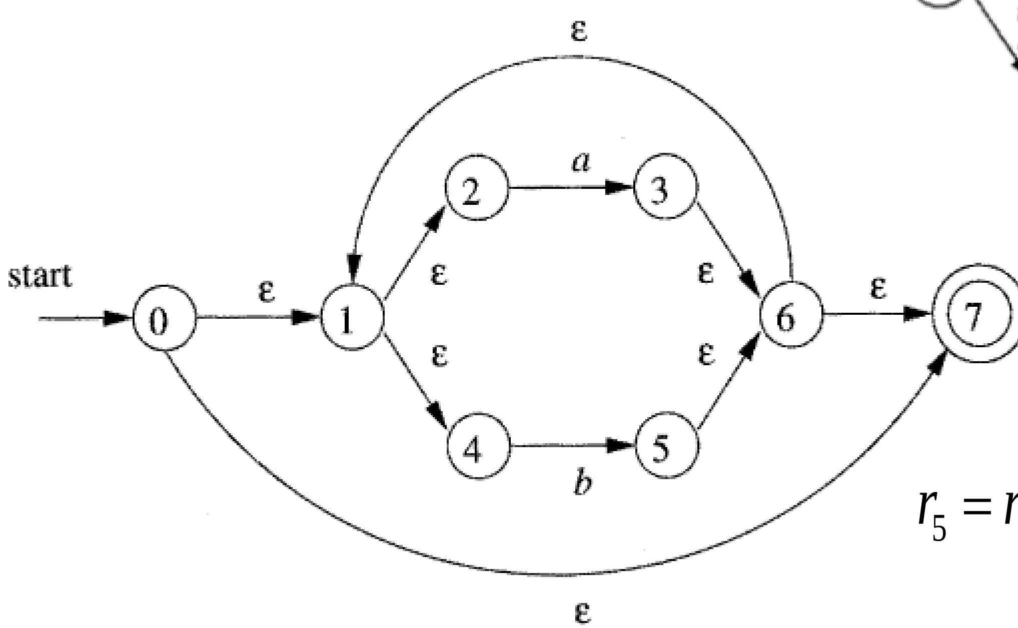
$r_2 = b$





$r_3 = r_1 / r_2$

$r_4 = (r_3)$



$r_5 = r_4 *$