

# Prednáška 3 - Regulárne výrazy, vlastnosti regulárnych jazykov

Ing. Viliam Hromada, PhD.

C-510  
Ústav informatiky a matematiky  
FEI STU

`viliam.hromada@stuba.sk`



## Regulárne gramatiky - opakovanie

Pre pripomenutie: Gramatika  $G = (N, T, P, S)$  sa nazýva regulárna, ak každé jej pravidlo spĺňa jeden z 2 tvarov:

1.  $A \rightarrow xB, A \in N, B \in N, x \in T^+$
2.  $A \rightarrow w, A \in N, w \in T^*$

Každý jazyk, ktorý je možné generovať regulárnou gramatikou sa nazýva **regulárny**.





## Regul'arna gramatika $\Rightarrow$ KA

### Veta

*Nech  $G = (N, T, P, S)$  je regul'arna gramatika. Potom existuje nedeterministicky konecny automat  $M = (Q, T, \delta, q_0, F)$  taky, ze  $L(M) = L(G)$ .*

### Veta

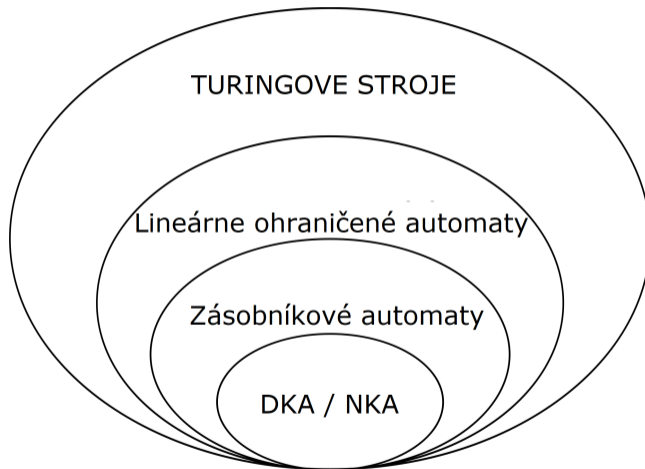
*Nech  $M = (Q, T, \delta, q_0, F)$  je DKA. Potom existuje tak' regul'arna gramatika  $G = (N, T, P, S)$ , ze  $L(G) = L(M)$ .*



- Teda každý jazyk, ku ktorému vieme nájsť regulárnu gramatiku, ktorá ho generuje, je zároveň akceptovateľný nejakým (deterministickým/nedeterministickým konečným automatom).
- A naopak, každý jazyk, ku ktorému existuje nejaký DKA/NKA, ktorý ho akceptuje, je generovateľný nejakou regulárnou gramatikou.



## Z hľadiska výpočtových zariadení



## Jazyk, ktorý nie je regulárny

- Ako naznačuje hierarchia, existujú jazyky, ktoré nie sú regulárne - a teda sa nedajú generovať regulárnou gramatikou, resp. akceptovať konečným automatom.
- Učebnicové príklady takýchto jazykov, napríklad nad abecedou  $A = \{a, b\}$  sú:
  - $L = \{a^n b^n \mid n \in \{0, 1, 2, \dots\}\}$
  - $L = \{w \in \{a, b\}^* \mid \#_a(w) = \#_b(w)\}$
  - $L = \{ww^R \mid w \in \{a, b\}^*\}$
  - $L = \{ww \mid w \in \{a, b\}^*\}$





## Regulárne jazyky

Ako už bolo spomínané, hoci sú regulárne jazyky pomerne obmedzené, používajú sa napríklad na:

- Popis ľubovoľnej konečnej množiny (napr. kľúčových slov programovacieho jazyka)
- Popis identifikátorov premenných, funkcií, atď.
- Popis číselných konštánt

Taktiež tzv. **regulárne výrazy** úzko súvisia s regulárnymi jazykmi...



# Regulárne výrazy

Regulárny výraz je popri regulárnych gramatikách a konečných automatoch iný spôsob popisu regulárnych jazykov.

# Regulárne výrazy

## Definícia

*Nech  $T$  je abeceda. Potom:*

- 1.  $\emptyset$  je regulárny výraz popisujúci prázdny jazyk,*
- 2.  $\varepsilon$  je regulárny výraz popisujúci jazyk  $\{\varepsilon\}$ ,*
- 3.  $a$ , pre  $a \in T$  je regulárny výraz popisujúci jazyk  $\{a\}$ .*
- 4. Ak  $R_1$  a  $R_2$  sú regulárne výrazy popisujúce jazyky  $L_1$  a  $L_2$ , potom  $(R_1 \mid R_2)$  je regulárny výraz popisujúci jazyk  $L_1 \cup L_2$  (zjednotenie).*
- 5. Ak  $R_1$  a  $R_2$  sú regulárne výrazy popisujúce jazyky  $L_1$  a  $L_2$ , potom  $(R_1 R_2)$  je regulárny výraz popisujúci jazyk  $L_1 L_2$  (zreťazenie).*
- 6. Ak  $R$  je regulárny výraz popisujúci jazyk  $L$ , potom  $(R^*)$  je regulárny výraz popisujúci iteráciu  $L^*$  jazyka  $L$ .*
- 7. iné regulárne výrazy ako tie, zostrojené podľa bodov 1-6, neexistujú.*

**Príklad:** Nech  $T = \{0, 1, \dots, 9, a, b, \dots, z\}$ . Potom možné regulárne výrazy:

- $(0 \mid 1) = \{0, 1\}$
- $(1(0^*)1) = \{11, 101, 1001, \dots\}$
- $((10)(0 \mid 1)^*) = \{10w \mid w \in \{0, 1\}^*\}$
- $((begin) \mid (end)) = \{begin, end\}$
- $((0 \mid 1 \mid \dots \mid 9)(0 \mid 1 \mid \dots \mid 9)^*)$  - celočíselné konštanty s prípadnými bezvýznamnými nulami zľava, bez znamienka
- $((\varepsilon \mid + \mid -)(0 \mid ((1 \mid \dots \mid 9)(0 \mid 1 \mid \dots \mid 9)^*)))$  - celočíselné konštanty bez/so znamienkom, bez bezvýznamných núl zľava
- $((a \mid \dots \mid z)^*(begin)(a \mid \dots \mid z)^*)$  - všetky textové reťazce obsahujúce *begin* ako podreťazec



## Regulárne výrazy $\Rightarrow$ NKA

### Veta

*Nech  $R$  je regulárny výraz popisujúci jazyk  $L$ . Potom existuje nedeterministický konečný automat  $M$  taký, že  $L(M) = L$ .*

Ukážeme si dôkaz tejto vety pomocou tzv. **Thompsonovej konštrukcie**.



Dôkaz -  $\emptyset$ ,  $\varepsilon$ ,  $a$ 

Stačí ukázať, že pre prvých 6 bodov z definície vieme vždy zostrojiť príslušný automat:

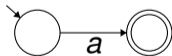
- Výraz  $\emptyset$ :



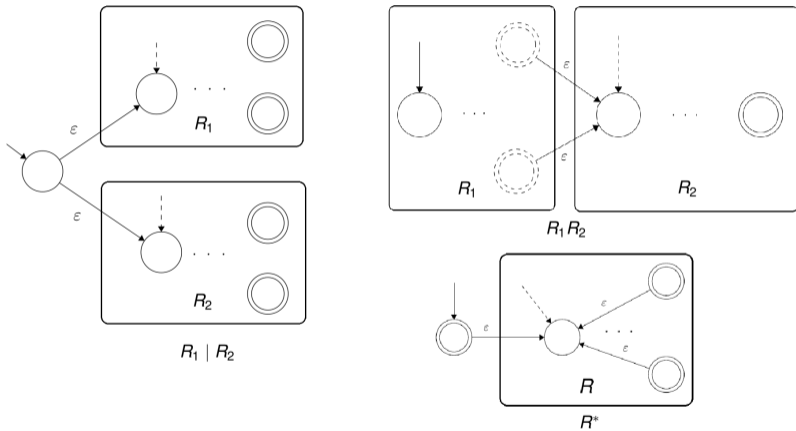
- Výraz  $\varepsilon$ :



- Výraz  $a$ :

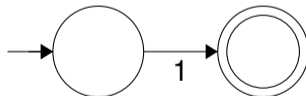
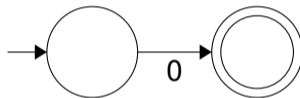


# Dôkaz - $R_1 \mid R_2, R_1 R_2, R^*$



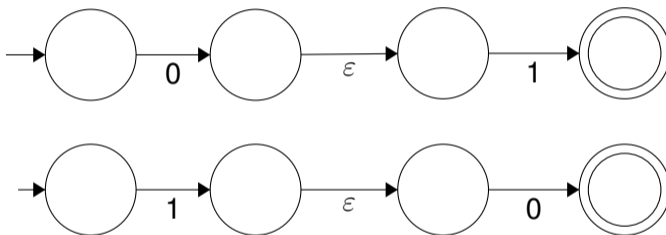
## Príklad

Nájdite DKA, ktorý akceptuje jazyk popísaný regulárnym výrazom:  
 $(01|10)^*(\varepsilon|0|1)$ . Najprv skonštruujeme NKA, ktorý rozpoznáva daný regulárny  
výraz... Začneme s elementárnymi KA pre 0 a 1.

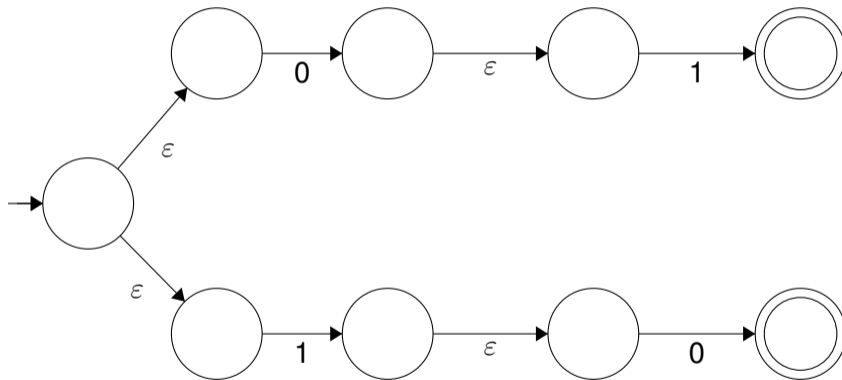




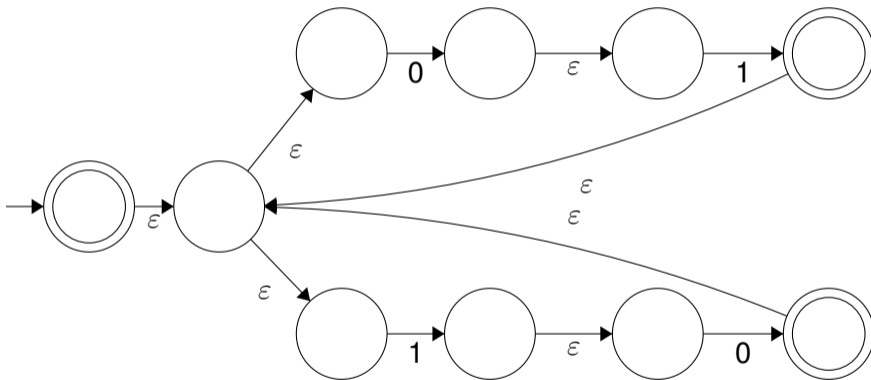
Z nich zreťazením dostávame 2 NKA - prvý pre 01 a druhý pre 10:



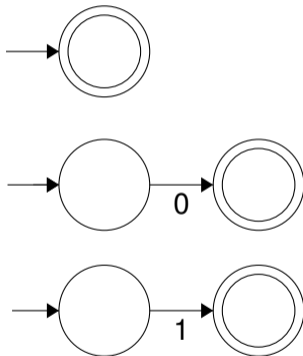
Zjednotením NKA pre 01 a NKA pre 10 vznikne NKA pre (01|10):



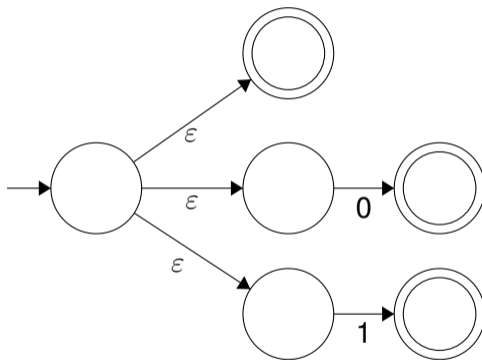
Iteráciou NKA pre  $(01|10)$  vznikne NKA pre  $(01|10)^*$ :



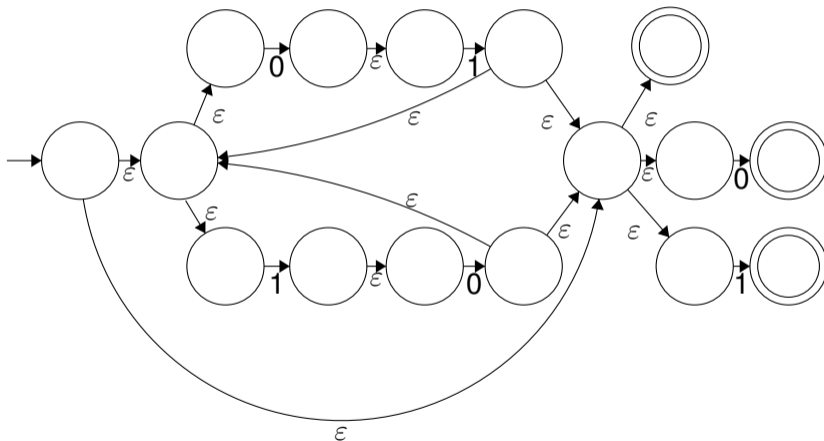
Ďalej potrebujeme zostrojiť NKA pre  $(\varepsilon|0|1)$ . Tri elementárne NKA, pre  $\varepsilon, 0, 1$  sú nasledovné:



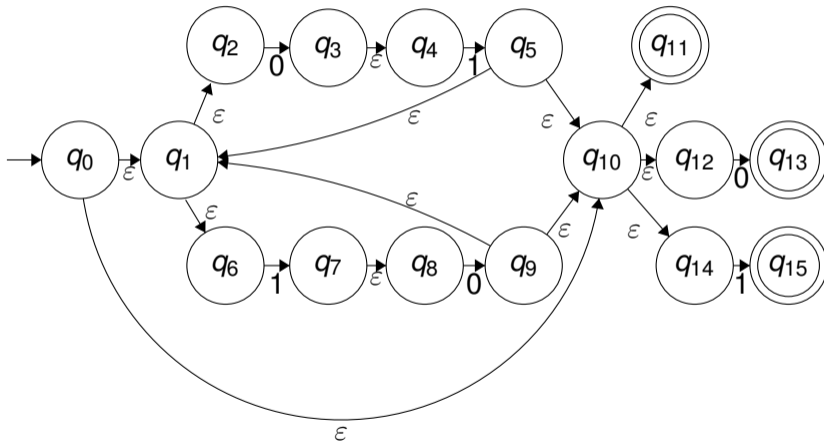
Ich zjednotením dostávame NKA pre  $\varepsilon|0|1$ :



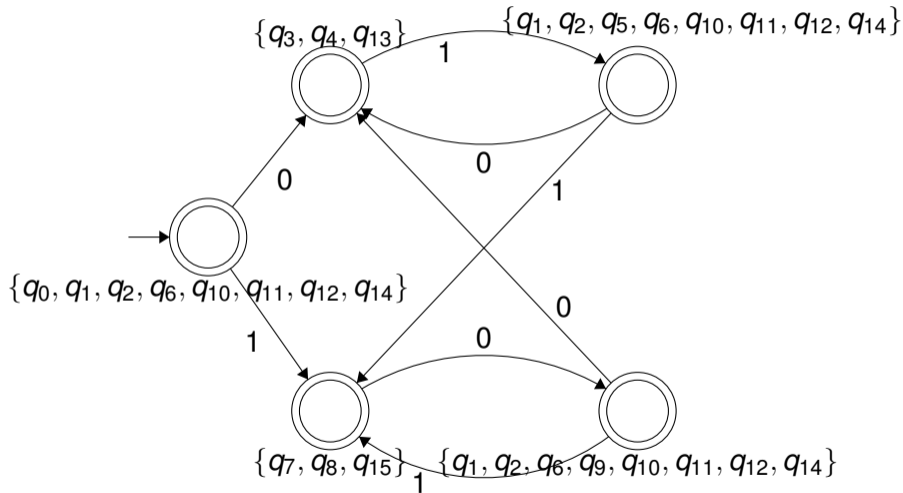
Na záver z NKA pre  $(01|10)^*$  a z NKA pre  $\epsilon|0|1$  ich zret'azením dostávame výsledný NKA pre  $(01|10)^*(\epsilon|0|1)$



Následne automat podrobíme determinizácii. Najprv však pomenujeme stavy (teraz je v princípe jedno, ako).



Ekvivalentný deterministický automat vyzerá nasledovne:







## Zhrnutie

Pre regulárne jazyky teda platí:

- Dajú sa popísať pomocou regulárnej gramatiky
- Dajú sa popísať pomocou (deterministického) konečného automatu
- Dajú sa popísať pomocou regulárneho výrazu.



## Použitá literatúra

Dedera, Ľ: Počítačové jazyky a ich spracovanie.

Linz, P.: An Introduction to Formal Languages and Automata.

Molnár, Ľ.: Gramatiky a jazyky.