

# Úvod do SHO



Výkonnost a spolehlivost – KIV/VSP

Richard Lipka

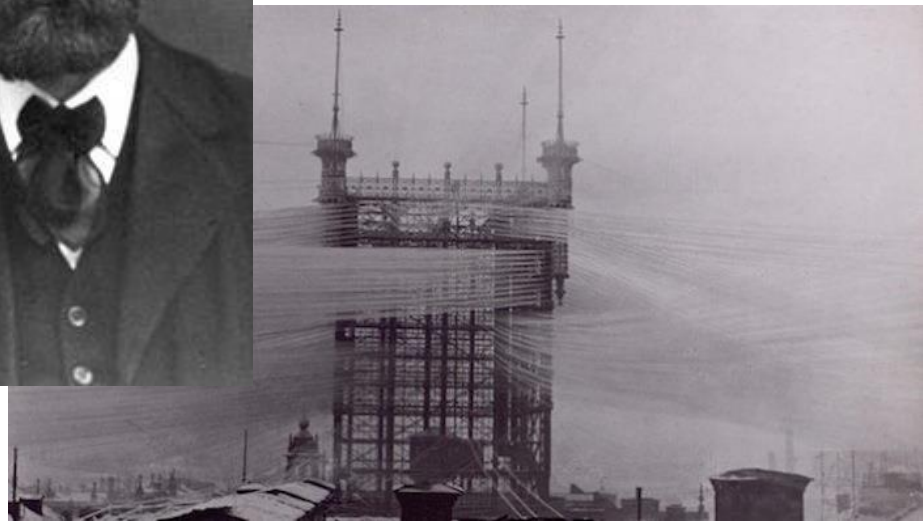
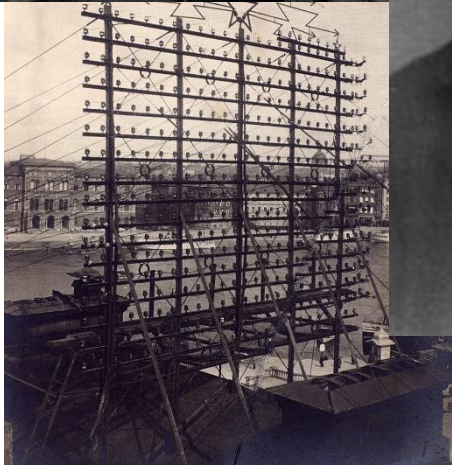
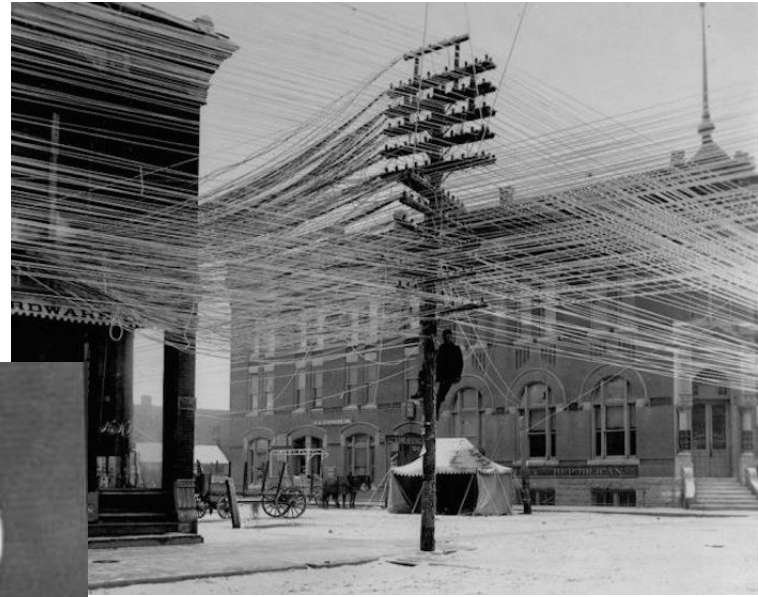
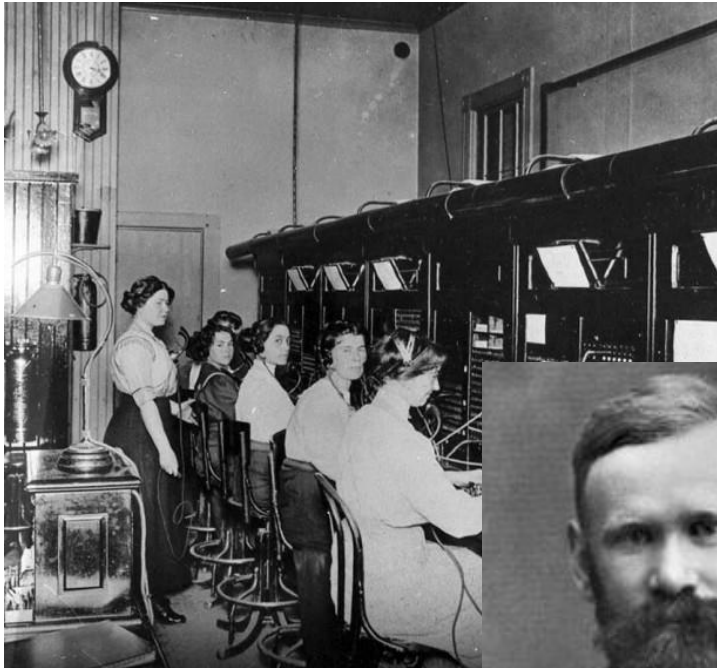
13.10.2015



# Systemy hromadné obsluhy (*Queueing theory*)

- Modelování systémů, které obsluhují větší množství požadavků
  - Telekomunikační systémy
  - Řízení dopravy
  - Plánování procesů v OS
  - Návrh výrobních linek
  - Krizový management
  - Návrh míst, kde se čeká (obchody, úřady, nemocnice ...)
- Predikce výkonu takových systémů
- První modely ze začátku 20. století (telefonní ústředna v Kodani – *Erlang, 1909*)
  - podobně staré jako **Markovské modely**
    - Řada vztahů odvozena právě přes markovské modely

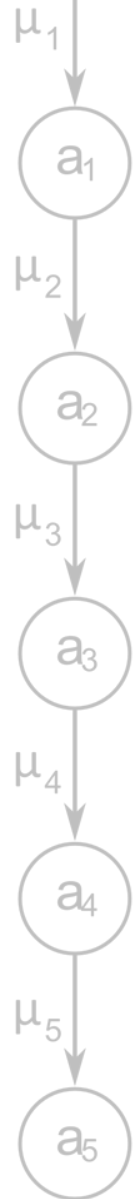




VSP - Úvod do SHO I



# SHO - základy



- Základní koncept:
  - *Systém poskytuje službu*
    - Realizována *kanálem obsluhy / serverem*
  - *Klienti posílají požadavky / zahajují transakce*
  - *Požadavky jsou buď obsluhovány, nebo čekají ve frontách*
- Teorie založena na statistice → funguje jen pro dostatečný počet požadavků
  - Abstrahuje od konkrétní realizace služby
  - Sleduje jen časové posloupnosti příchodů požadavků a jejich obsluh  
→ lze určit základní výkonové charakteristiky
- Zjednodušení umožňuje konstruovat SHO model v uzavřeném tvaru (jako vzoreček)  
(složité SHO je třeba řešit simulačně, ne vždy je možné nebo praktické najít uzavřený tvar)



# Příklady

System	Klient	Server	Požadavek
Dopravní síť	Vozidlo	Křižovatka	Průjezd křižovatkou
Google	Hledající	Vyhledávací stroj a DB	Vyhledání stránky
Letiště	Letadlo	Runway	Odlet nebo přilet
Úřad	Občan	Přepážka	Vydání dokladu
Menza	Hladový student	Výdejní okénko / pult / kasa	Vydání jídla / zaplacení jídla
Myčka aut	Vozidlo	Mycí linka	Umytí vozidla

VSP - Úvod do SHOI



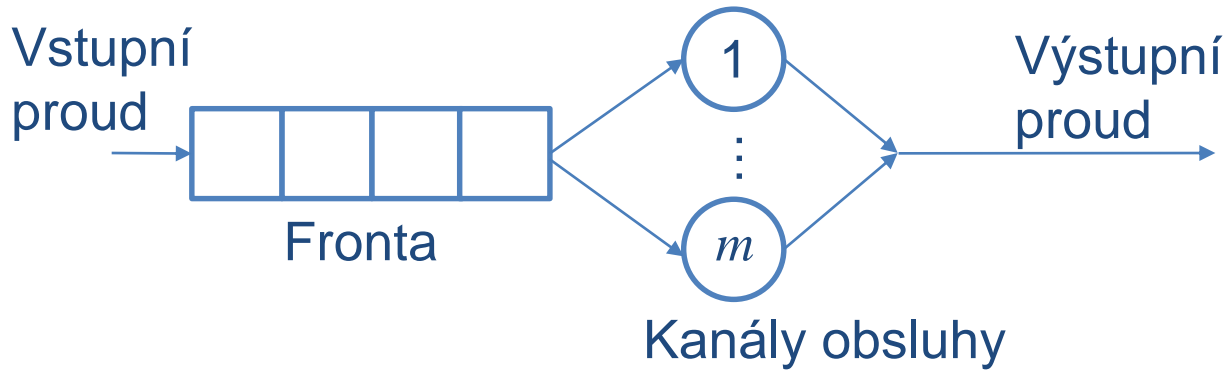
# Základní problémy

- Příliš mnoho požadavků na systém
  - Požadavky čekají ve frontě příliš dlouho (fronta může narůstat do nekonečna)
  - Požadavky nemají kde čekat a jsou zahazovány
  - potřebuji zvýšit kapacitu systému (přidat zdroje) – relativně snadné v cloudových řešeních
- Příliš málo požadavků na systém
  - Systém není vytížen – většinu času nic nedělá (ale stále spotřebovává zdroje)
  - mohu zdroje využít nějak jinak (nebo zrušit)





# Elementární SHO



- Stav systému: počet požadavků v něm
- Potřebujeme znát:
  - Charakteristiku vstupního proudu
  - Chování fronty
  - Charakteristiku obsluhy požadavků
- Předpokládáme stacionární režim činnosti
  - Charakteristiky se nemění v čase (podobně jako u Markovských modelů)  
 → modelujeme *ustálený provoz*, ne *přechodový děj*

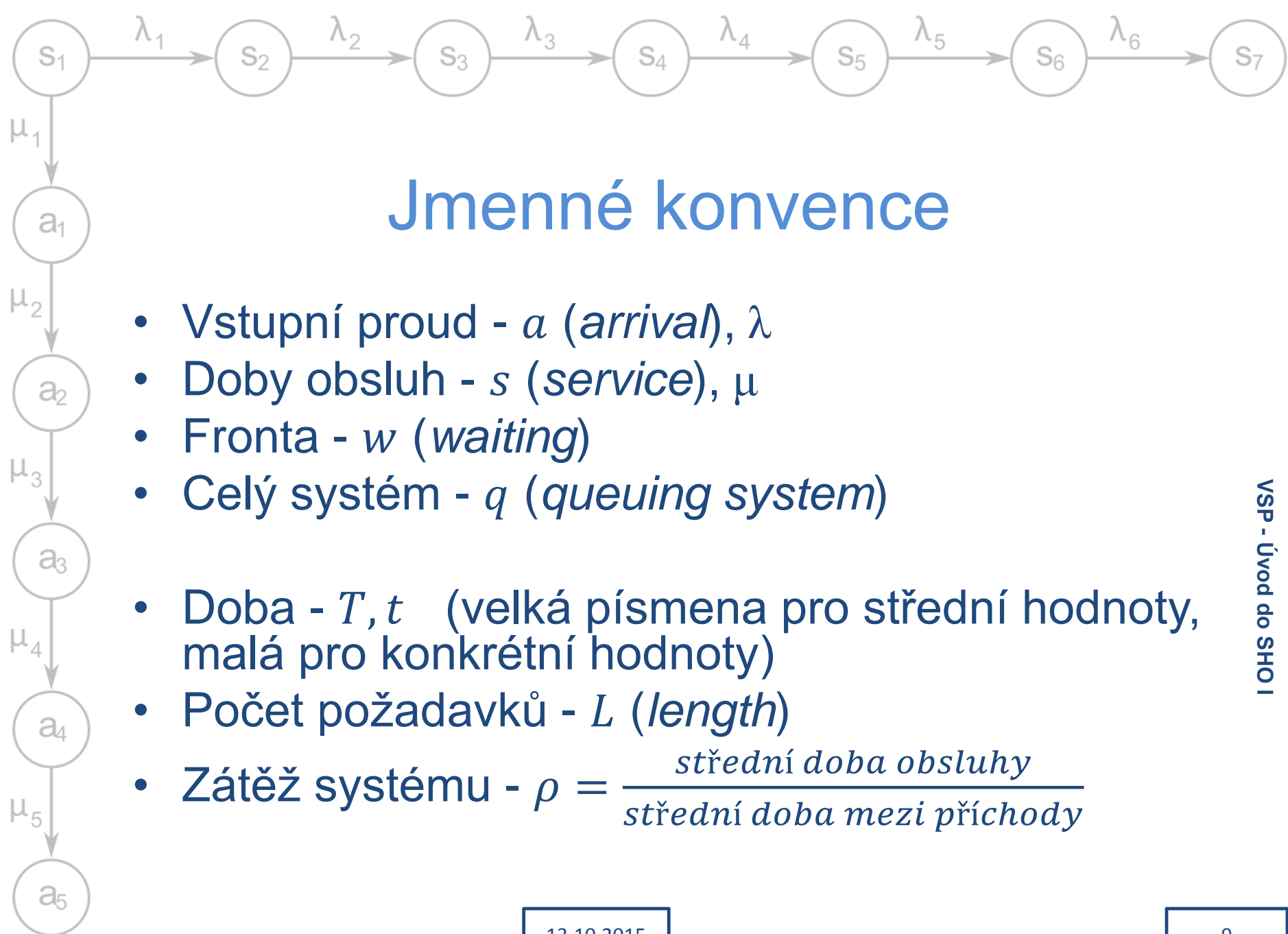
VSP - Úvod do SHO I



## Zdroje požadavků

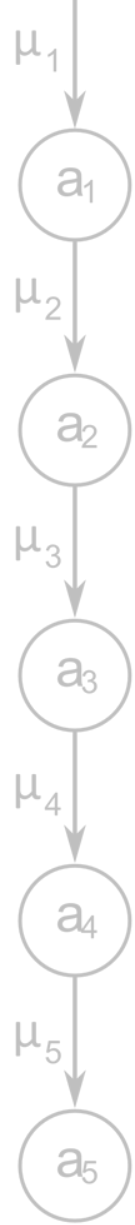
- Vkládají požadavky do fronty
  - Abstraktní, obvykle bez ekvivalentu v reálném světě, hodí se hlavně pro simulace
- Lze dělit podle počtu požadavků
  - *Omezené zdroje*
    - Předem omezená množina požadavků  
→ vstupní proud závisí na stavu SHO
    - Např. modelování procesů v kritické sekci
  - *Neomezené zdroje*
    - „Nekonečné“ množství požadavků  
→ vstupní proud nezávisí na stavu SHO
    - Např. požadavky na webový server





# Jmenné konvence

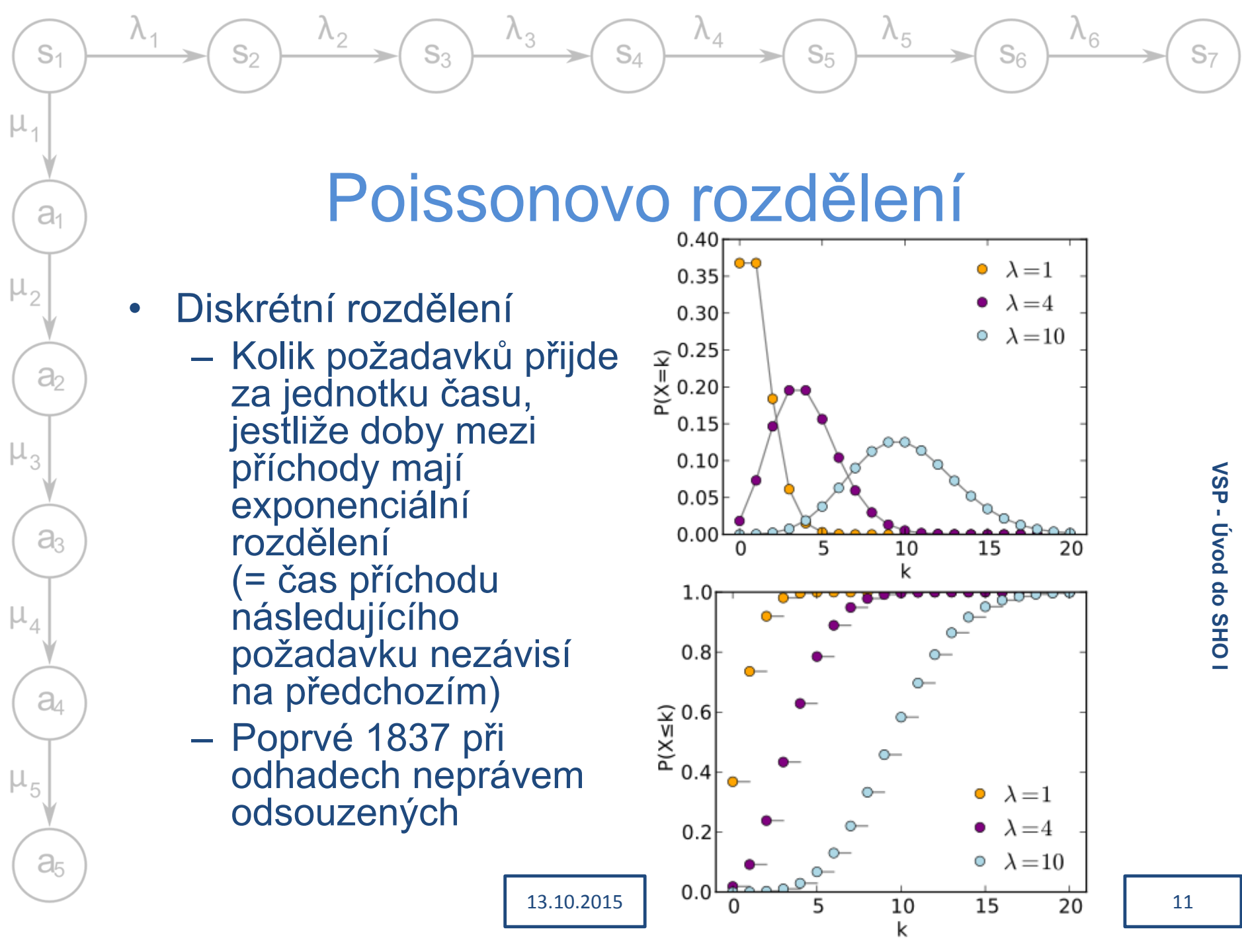
- Vstupní proud -  $a$  (*arrival*),  $\lambda$
- Doby obsluh -  $s$  (*service*),  $\mu$
- Fronta -  $w$  (*waiting*)
- Celý systém -  $q$  (*queueing system*)
- Doba -  $T, t$  (velká písmena pro střední hodnoty, malá pro konkrétní hodnoty)
- Počet požadavků -  $L$  (*length*)
- Zátěž systému -  $\rho = \frac{\textit{střední doba obsluhy}}{\textit{střední doba mezi příchody}}$



# Vstupní proud

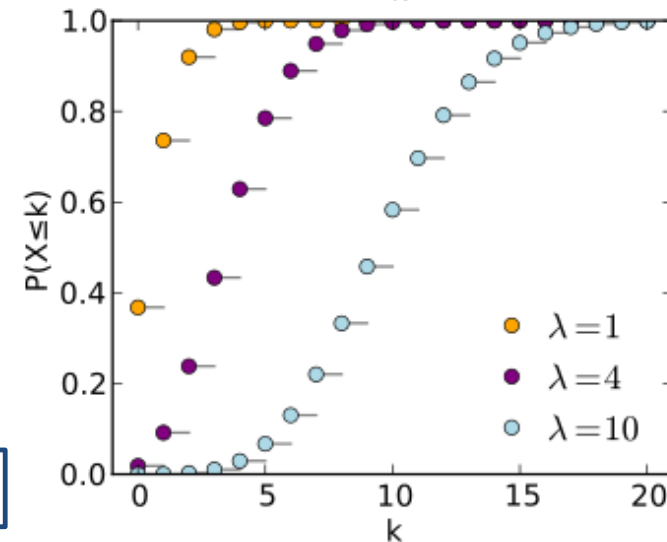
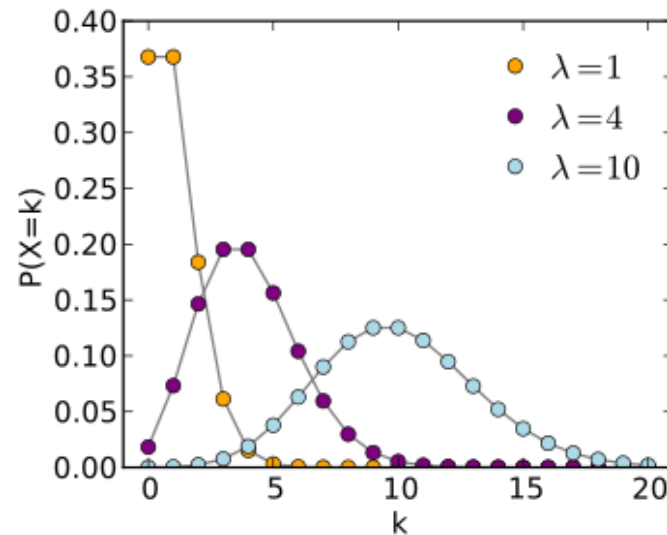
- Časová posloupnost s jakou požadavky do systému vstupují -  $\{t_1 < t_2 < t_3 \dots\}$ 
  - Popsán jako  $\tau_k = t_k - t_{k-1}$  pro  $k \geq 1$  – interval mezi příchody (*inter-arrival time*)
  - Pokud jsou jednotlivé hodnoty  $\tau_k$  statisticky nezávislé a mají stejné rozdělení, považovány za realizaci veličiny  $\tau$
- Obvyklý popis:  $F_a(t) = P\{\tau \leq t\}$  nebo odpovídající  $f_a(t) = (F_a(t))'$
- Typické vstupní proudy:
  - *Poissonovský* – exponenciální rozdělení pro doby mezi příchody požadavků (v každém okamžiku stejná pravděpodobnost příchodu dalšího požadavku – nezáleží na tom jak dlouhá doba uběhla od posledního)
  - Gaussovský
  - Rovnoměrný – bez náhody, stejné intervaly

VSP - Úvod do SHO I



# Poissonovo rozdělení

- Diskrétní rozdělení
  - Kolik požadavků přijde za jednotku času, jestliže doby mezi příchody mají exponenciální rozdělení (= čas příchodu následujícího požadavku nezávisí na předchozím)
  - Poprvé 1837 při odhadech neprávem odsouzených

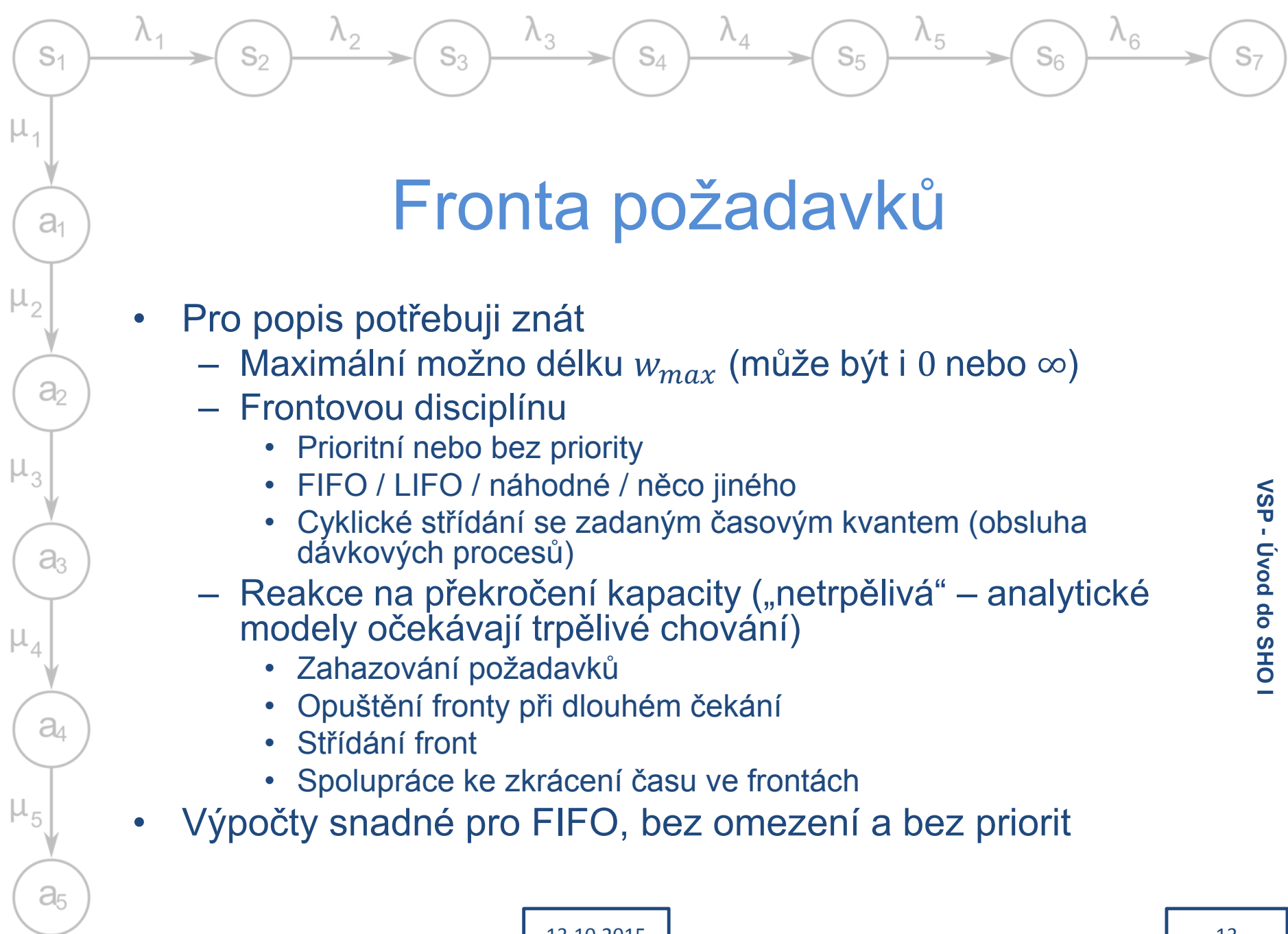


VSP - Úvod do SHOI



# Charakteristika vstupního proudu

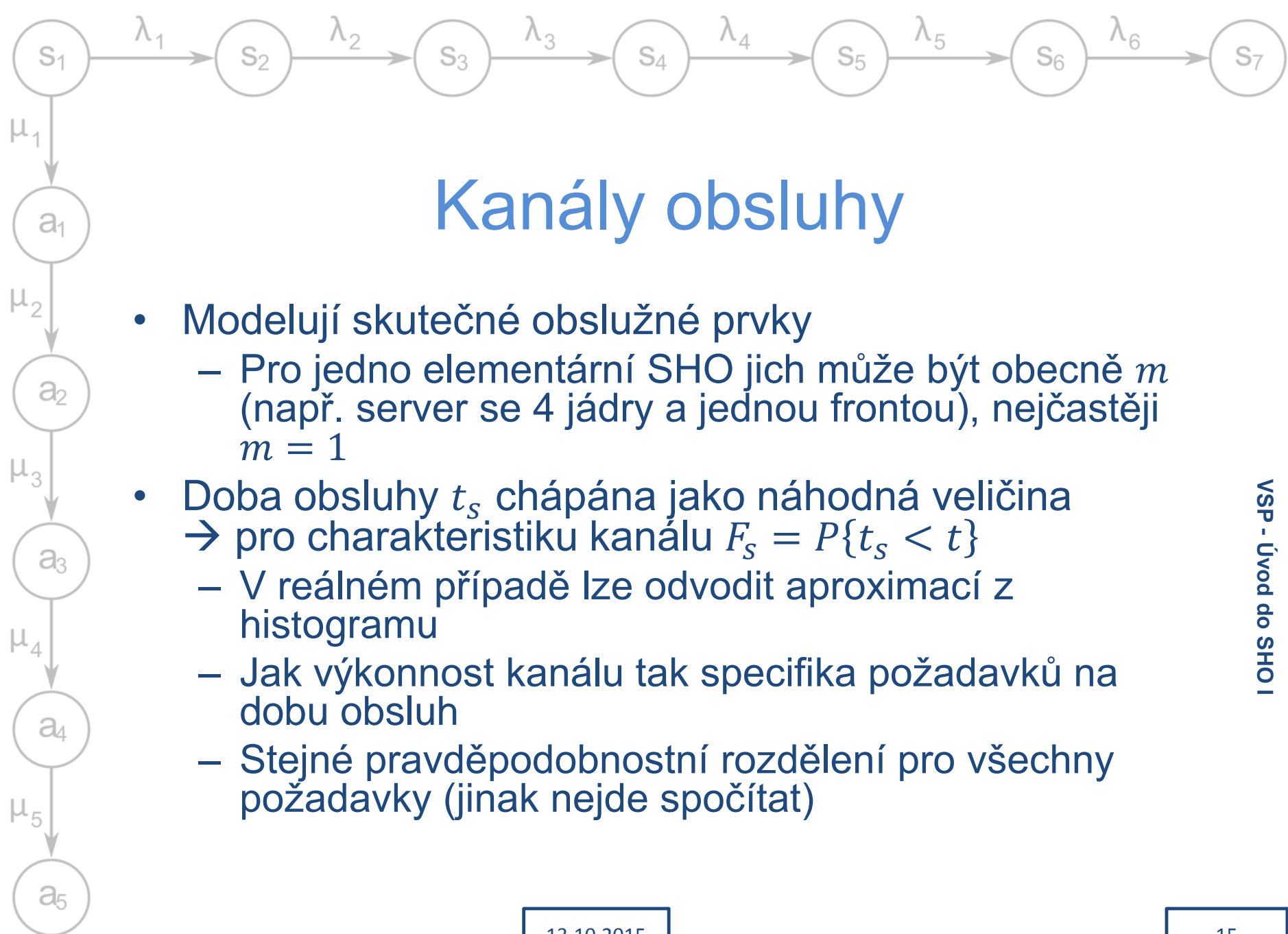
- $F_a(t)$  - distribuční funkce pro doby mezi příchody požadavků
  - Pro poissonovský proud  $F_a(t) = 1 - e^{-\lambda t}$
  - Lze ho popsat jedním parametrem  $\lambda$
- $E\{\tau\} = T_a = \frac{1}{\lambda}$  - střední doba mezi příchody  
 ( $\lambda$  – střední *frekvence* příchodů požadavků)
- $C_a = \frac{\sigma\{\tau\}}{T_a}$  - koeficient variance
  - „jak moc je proud náhodný“
  - Pravidelné příchody -  $C_a = 0$ , poissonovský proud  $C_a = 1$ , reálně někde v intervalu  $\langle 0; 1 \rangle$
  - Pokud se objevují shluky požadavků může být  $C_a > 1$





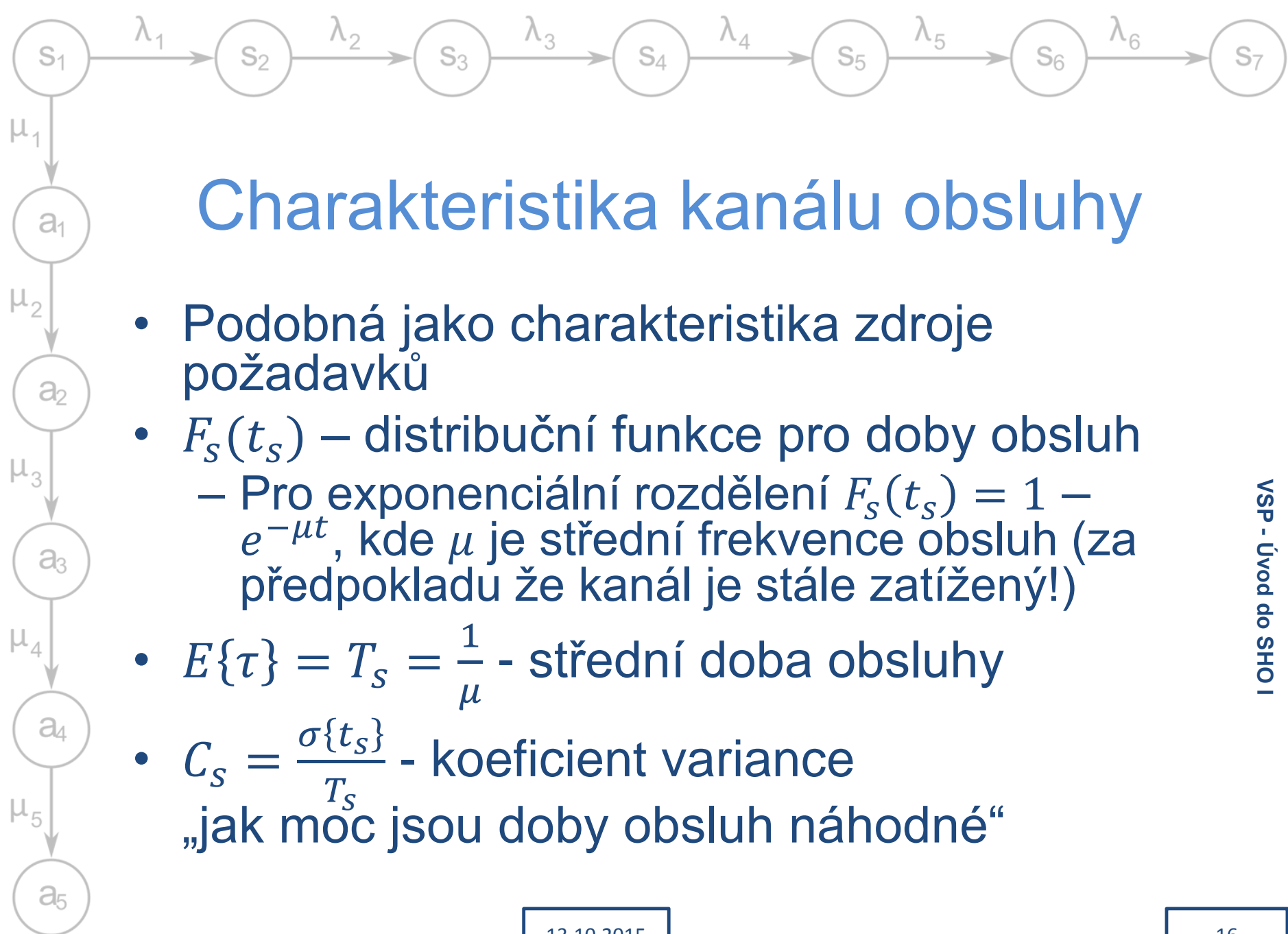
## Charakteristika fronty

- $w$  – aktuální (okamžitý) počet požadavků ve frontě ( $w(t)$  - funkce času nebo náhodná veličina)
- $E\{w\} = L_w$  – střední počet požadavků ve frontě (střední délka)
- $t_w$  – doba čekání jednoho konkrétního požadavku (náhodná veličina)
- $E\{t_w\} = T_w$  - střední doba čekání požadavku ve frontě



# Kanály obsluhy

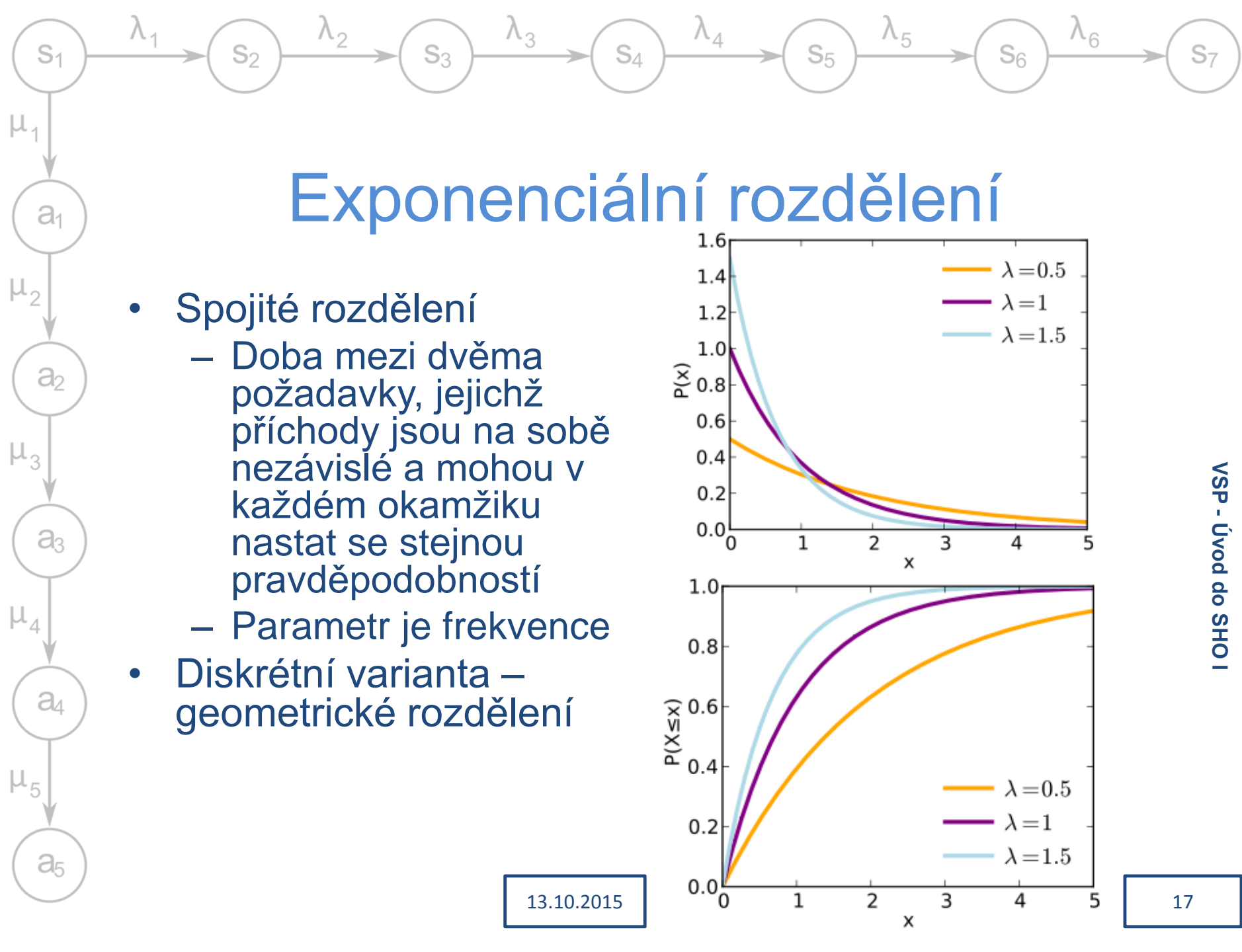
- Modelují skutečné obslužné prvky
  - Pro jedno elementární SHO jich může být obecně  $m$  (např. server se 4 jádry a jednou frontou), nejčastěji  $m = 1$
- Doba obsluhy  $t_s$  chápána jako náhodná veličina
  - pro charakteristiku kanálu  $F_s = P\{t_s < t\}$ 
    - V reálném případě lze odvodit aproximací z histogramu
    - Jak výkonnost kanálu tak specifika požadavků na dobu obsluh
    - Stejné pravděpodobnostní rozdělení pro všechny požadavky (jinak nejde spočítat)



# Charakteristika kanálu obsluhy

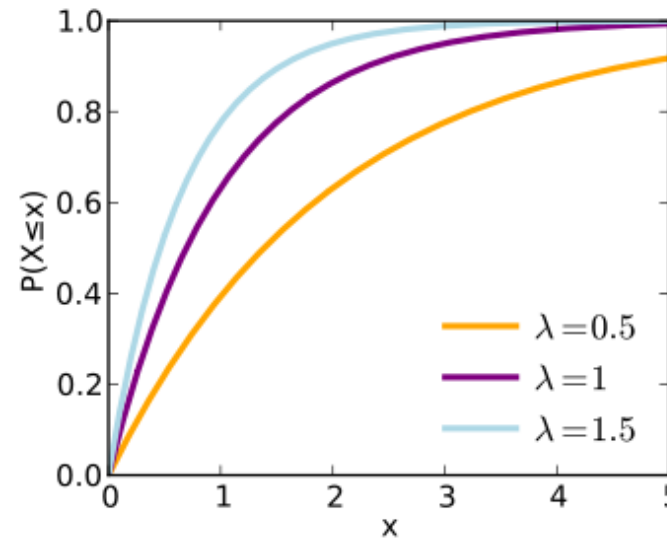
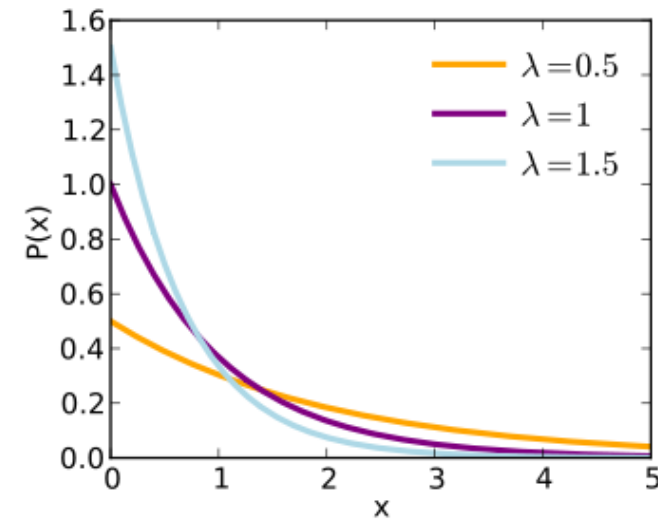
- Podobná jako charakteristika zdroje požadavků
- $F_S(t_S)$  – distribuční funkce pro doby obsluh
  - Pro exponenciální rozdělení  $F_S(t_S) = 1 - e^{-\mu t}$ , kde  $\mu$  je střední frekvence obsluh (za předpokladu že kanál je stále zatížený!)
- $E\{\tau\} = T_S = \frac{1}{\mu}$  - střední doba obsluhy
- $C_S = \frac{\sigma\{t_S\}}{T_S}$  - koeficient variance  
 „jak moc jsou doby obsluh náhodné“



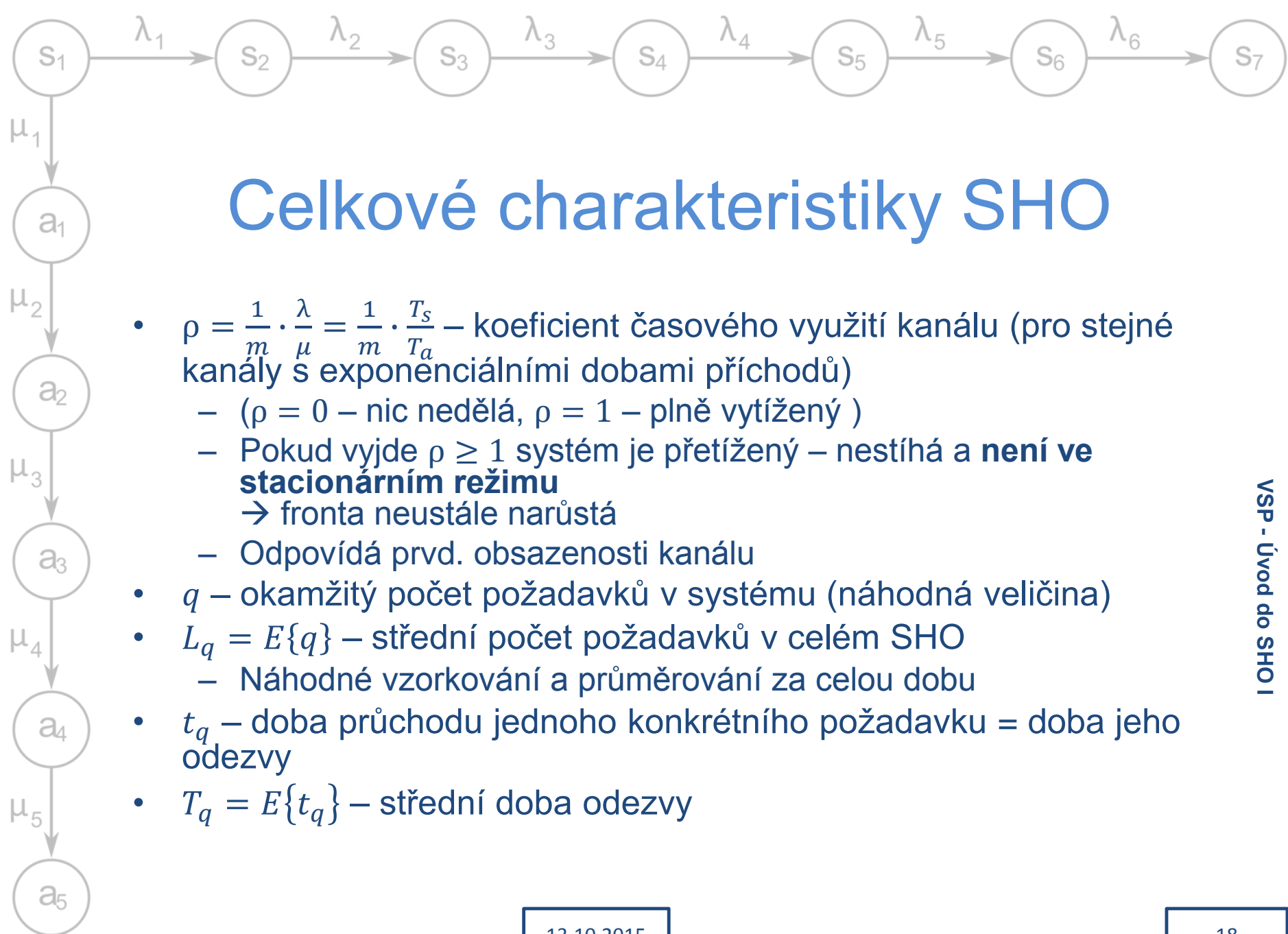


# Exponenciální rozdělení

- Spojité rozdělení
  - Doba mezi dvěma požadavky, jejichž příchody jsou na sobě nezávislé a mohou v každém okamžiku nastat se stejnou pravděpodobností
  - Parametr je frekvence
- Diskrétní varianta – geometrické rozdělení

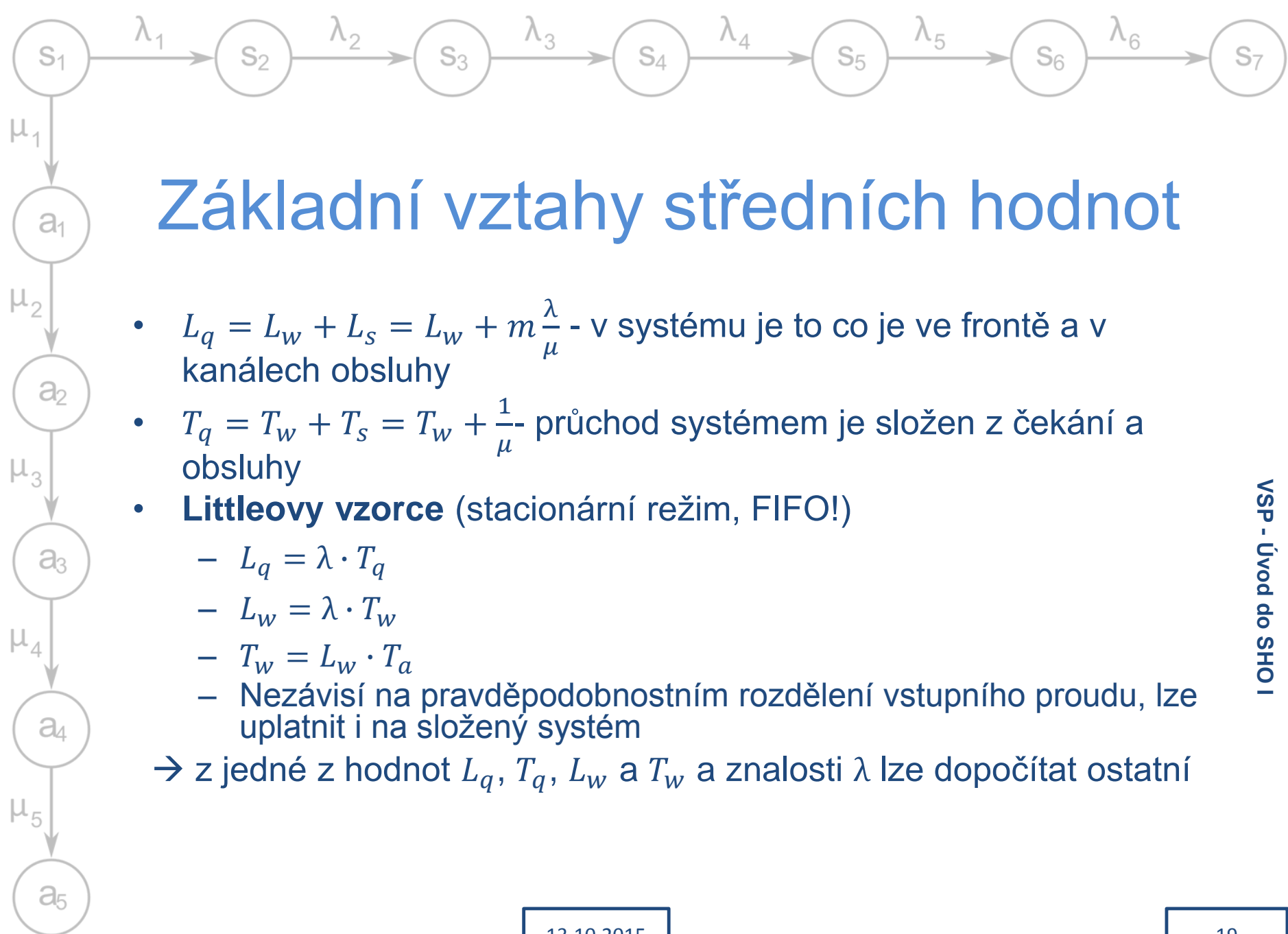


VSP - Úvod do SHO I



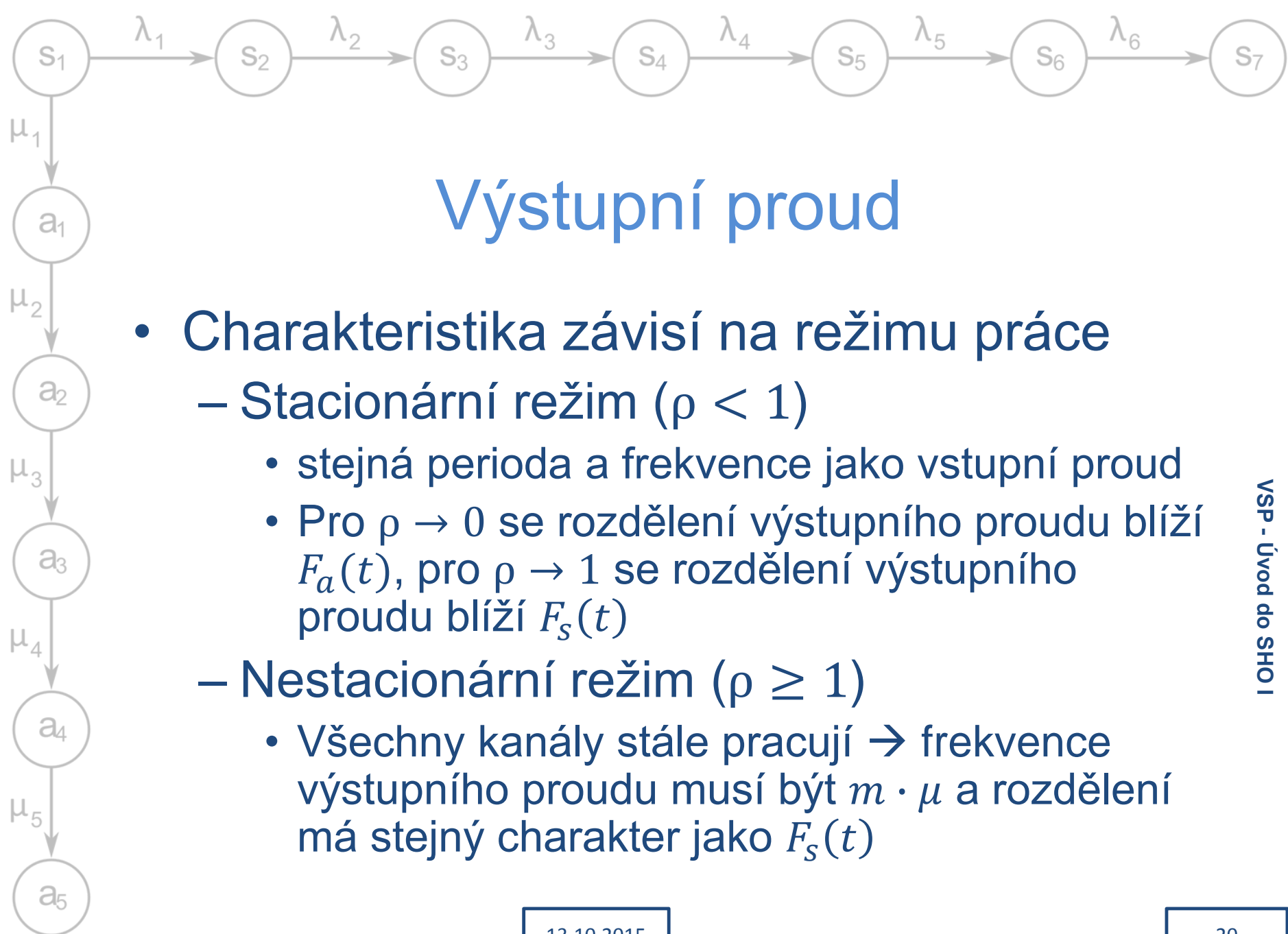
# Celkové charakteristiky SHO

- $\rho = \frac{1}{m} \cdot \frac{\lambda}{\mu} = \frac{1}{m} \cdot \frac{T_s}{T_a}$  – koeficient časového využití kanálu (pro stejné kanály s exponenciálními dobami příchodů)
  - ( $\rho = 0$  – nic nedělá,  $\rho = 1$  – plně vytížený )
  - Pokud vyjde  $\rho \geq 1$  systém je přetížený – nestíhá a **není ve stacionárním režimu**  
 → fronta neustále narůstá
  - Odpovídá prvd. obsazenosti kanálu
- $q$  – okamžitý počet požadavků v systému (náhodná veličina)
- $L_q = E\{q\}$  – střední počet požadavků v celém SHO
  - Náhodné vzorkování a průměrování za celou dobu
- $t_q$  – doba průchodu jednoho konkrétního požadavku = doba jeho odezvy
- $T_q = E\{t_q\}$  – střední doba odezvy



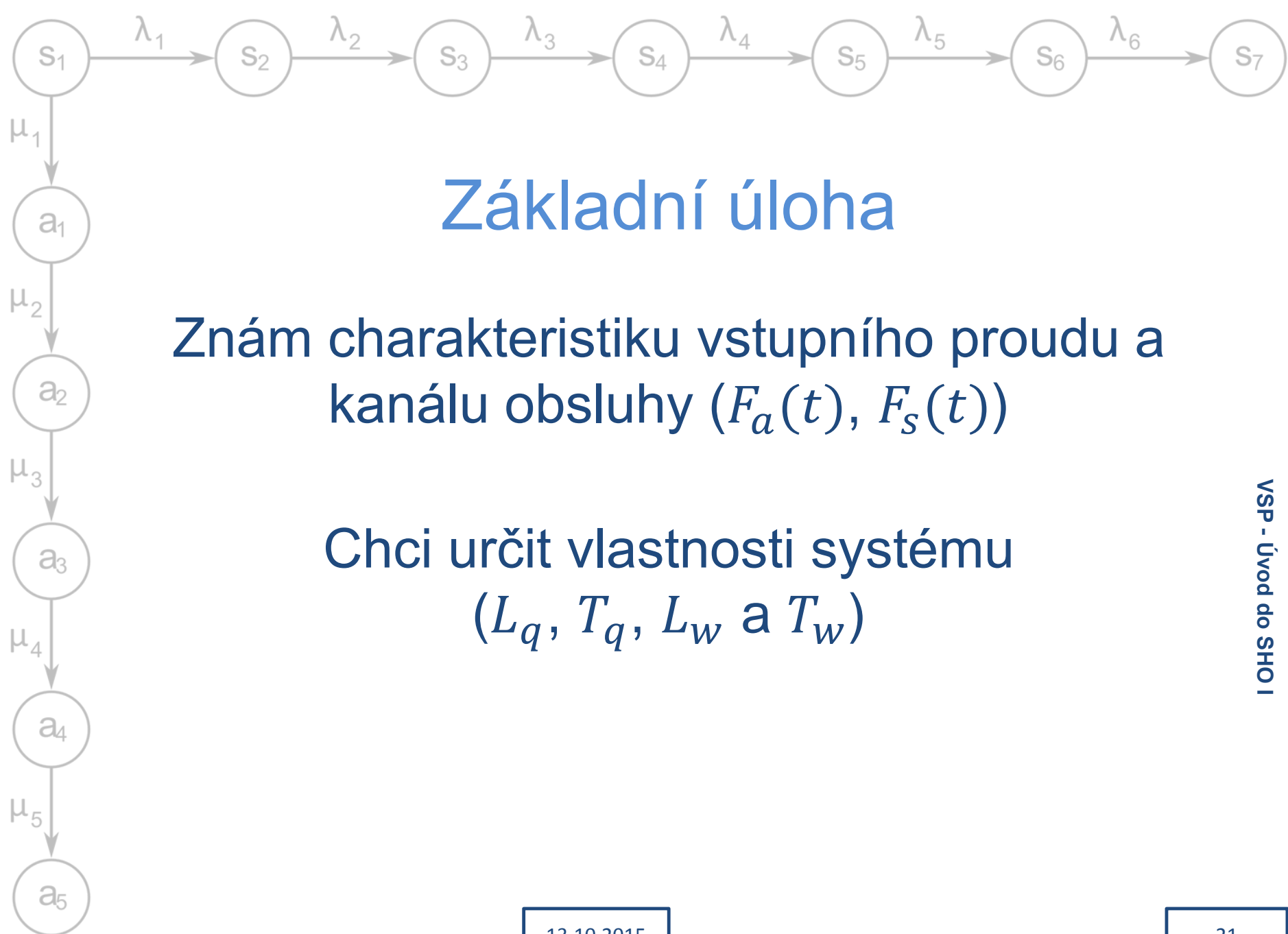
# Základní vztahy středních hodnot

- $L_q = L_w + L_s = L_w + m \frac{\lambda}{\mu}$  - v systému je to co je ve frontě a v kanálech obsluhy
  - $T_q = T_w + T_s = T_w + \frac{1}{\mu}$  - průchod systémem je složen z čekání a obsluhy
  - **Littleovy vzorce** (stacionární režim, FIFO!)
    - $L_q = \lambda \cdot T_q$
    - $L_w = \lambda \cdot T_w$
    - $T_w = L_w \cdot T_a$
    - Nezávisí na pravděpodobnostním rozdělení vstupního proudu, lze uplatnit i na složený systém
- z jedné z hodnot  $L_q$ ,  $T_q$ ,  $L_w$  a  $T_w$  a znalosti  $\lambda$  lze dopočítat ostatní



## Výstupní proud

- Charakteristika závisí na režimu práce
  - Stacionární režim ( $\rho < 1$ )
    - stejná perioda a frekvence jako vstupní proud
    - Pro  $\rho \rightarrow 0$  se rozdělení výstupního proudu blíží  $F_a(t)$ , pro  $\rho \rightarrow 1$  se rozdělení výstupního proudu blíží  $F_s(t)$
  - Nestacionární režim ( $\rho \geq 1$ )
    - Všechny kanály stále pracují  $\rightarrow$  frekvence výstupního proudu musí být  $m \cdot \mu$  a rozdělení má stejný charakter jako  $F_s(t)$



## Základní úloha

Znám charakteristiku vstupního proudu a kanálu obsluhy ( $F_a(t), F_s(t)$ )

Chci určit vlastnosti systému ( $L_q, T_q, L_w$  a  $T_w$ )



# Kendallova klasifikace

- Charakteristika různých typů SHO, podle vlastností vstupních proudů, obsluh, front ... (z 50. let)
- Systém je popsán trojicí (pěticí)

**X / Y / m / I / disc.**

- **X** – prvd. rozdělení vstupního proudu
- **Y** – prvd. rozdělení dob obsluh
- **m** – počet kanálů obsluhy
- **I** – maximální délka fronty (obvykle  $\infty$  - pak se neuvádí)
- **disc.** – frontová disciplína (obvykle FIFO, pak se neuvádí)





# Kendallova klasifikace



- Typické charakteristiky proudů:
  - GI – obecné náhodné rozdělení, intervaly příchodů statisticky nezávislé
  - G – obecné náhodné rozdělení
  - M – exponenciální rozdělení dob obsluh nebo příchodů požadavků (*Markovské*)
  - D – deterministické intervaly (konstantní nebo jinak pravidelné)
- Typické příklady: M/M/1, M/M/m, M/M/1/n/FIFO, M/G/1



## M/M/1 (/∞/FIFO)



- Nejjednodušší varianta

- Poissonovský vstupní proud, exponenciální doby obsluhy → lze charakterizovat 2

parametry:  $\frac{\lambda}{\mu}$

- $\lambda$  – střední frekvence vstupního proudu  
( $F_a(t) = 1 - e^{-\lambda t}$ )

- $\mu$  – střední podmíněná frekvence obsluh  
pro  $\rho < 1$  bude skutečná frekvence nižší

- $\rho = \frac{\lambda}{\mu} \rightarrow$  pro stacionární režim  $\lambda < \mu$

- Lze modelovat jako markovský proces





# Markovský model M/M/1



$$0 = -\lambda p_0 + \mu p_1$$

$$p_1 = \frac{\lambda}{\mu} p_0 = \rho p_0$$

$$0 = \lambda p_0 - \lambda p_1 - \mu p_1 + \mu p_2$$

$$p_2 = \frac{\lambda}{\mu} p_1 = \rho p_1 = \rho^2 p_0$$

...

$$\rho_k = \rho^k p_0 = \rho^k (1 - \rho)$$

- Prvd. funkce rozložení náhodné veličiny k (počet požadavků v SHO)



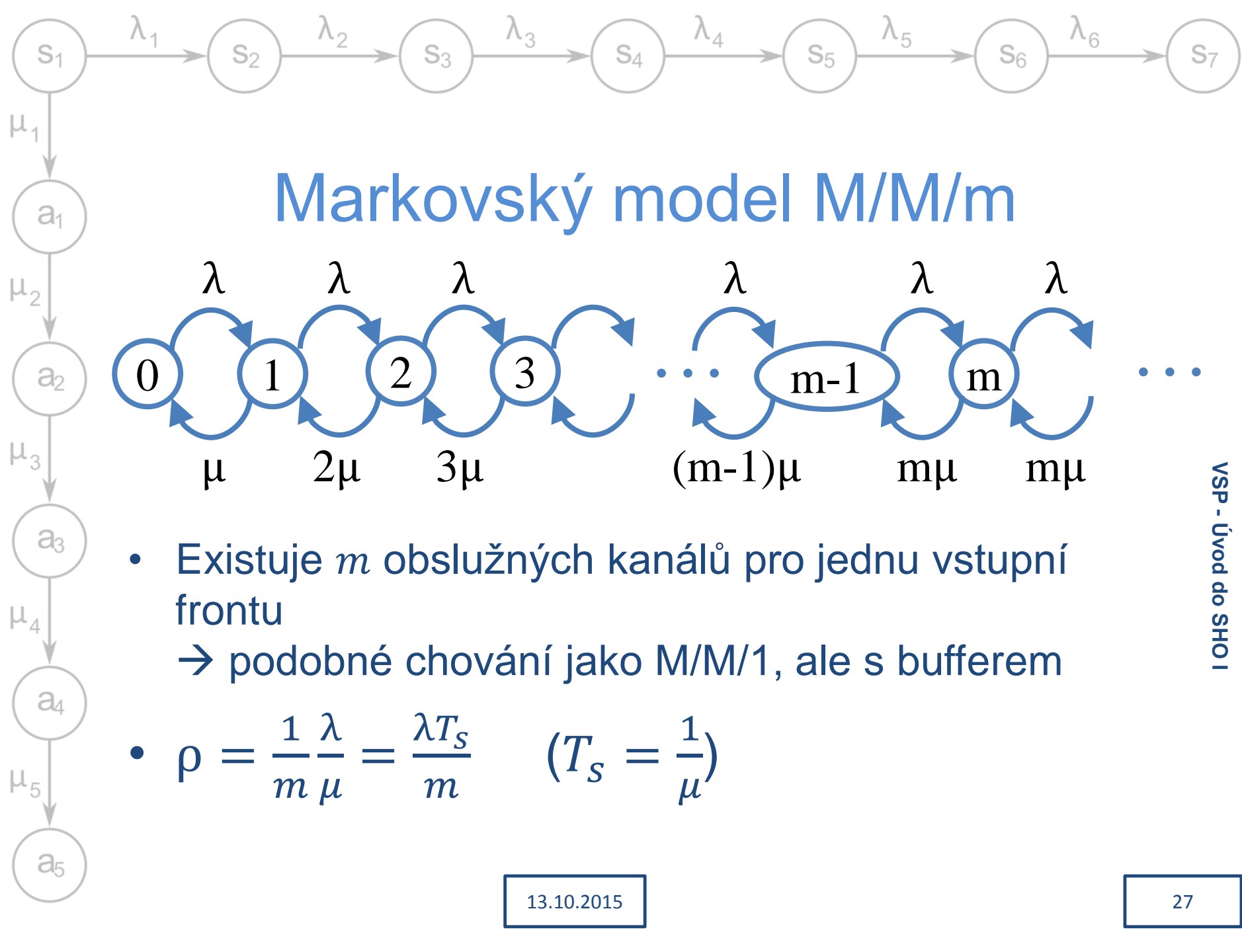
# Markovský model M/M/1



$$\rho_k = \rho^k p_0 = \rho^k (1 - \rho)$$

- $$E\{k\} = \sum_{k=0}^{\infty} k p_k = (1 - \rho) \sum_{k=0}^{\infty} k \rho^k = (1 - \rho) \frac{\rho}{(1 - \rho)^2}$$

$$= \frac{\rho}{(1 - \rho)} = L_q \text{ (pro } (\rho < 1))$$
- $T_q, L_w$  a  $T_w$  lze určit z Littleových vzorců



# Markovský model M/M/m

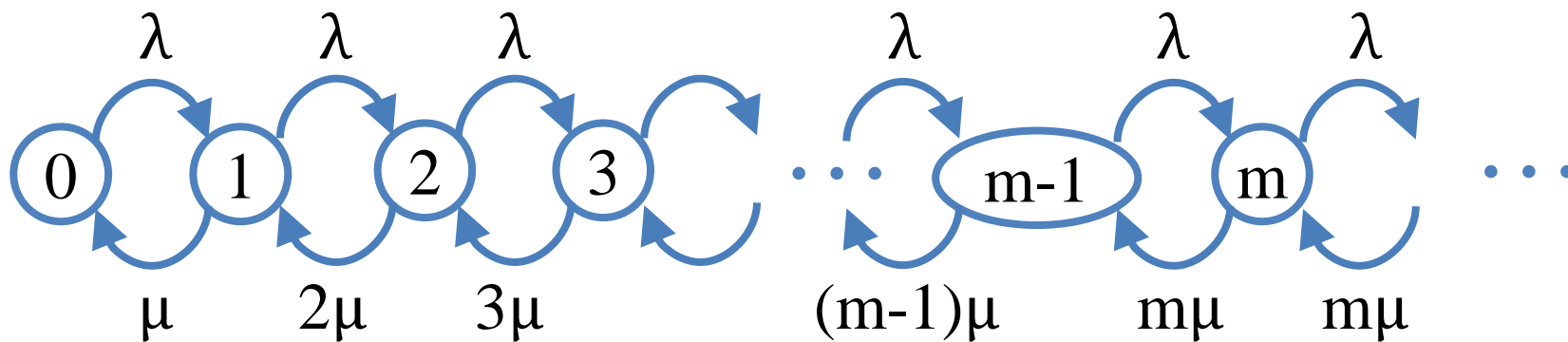
VSP - Úvod do SHO I

- Existuje  $m$  obslužných kanálů pro jednu vstupní frontu  
 → podobné chování jako M/M/1, ale s bufferem

- $$\rho = \frac{1}{m} \frac{\lambda}{\mu} = \frac{\lambda T_s}{m} \quad \left( T_s = \frac{1}{\mu} \right)$$



## Markovský model M/M/m



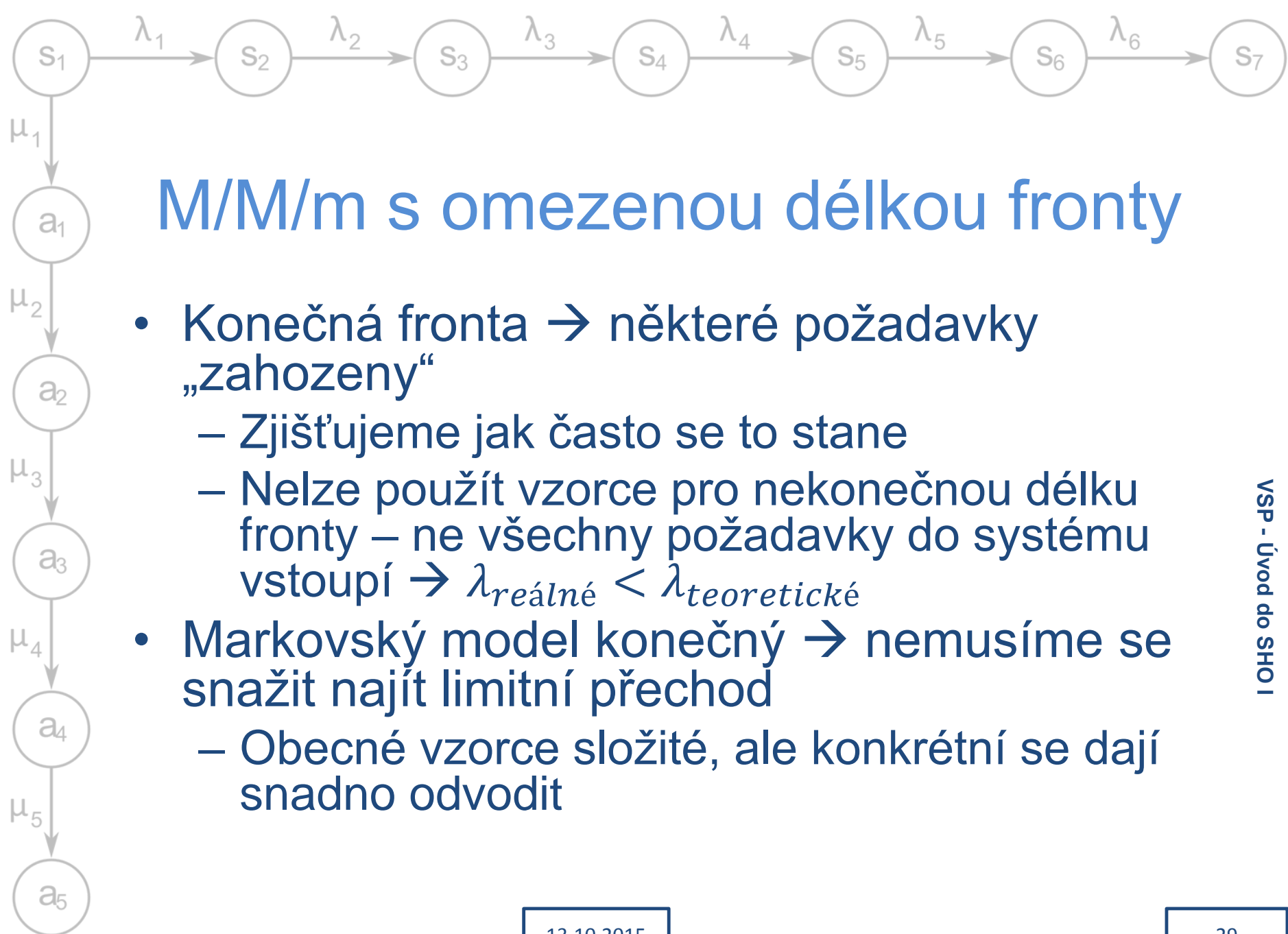
- Existuje obecné (složitě řešení), pro přibližný odhad (nebo  $m \in \{1,2\}$ ):

$$- L_q \cong \frac{m\rho}{1-\rho^m}$$

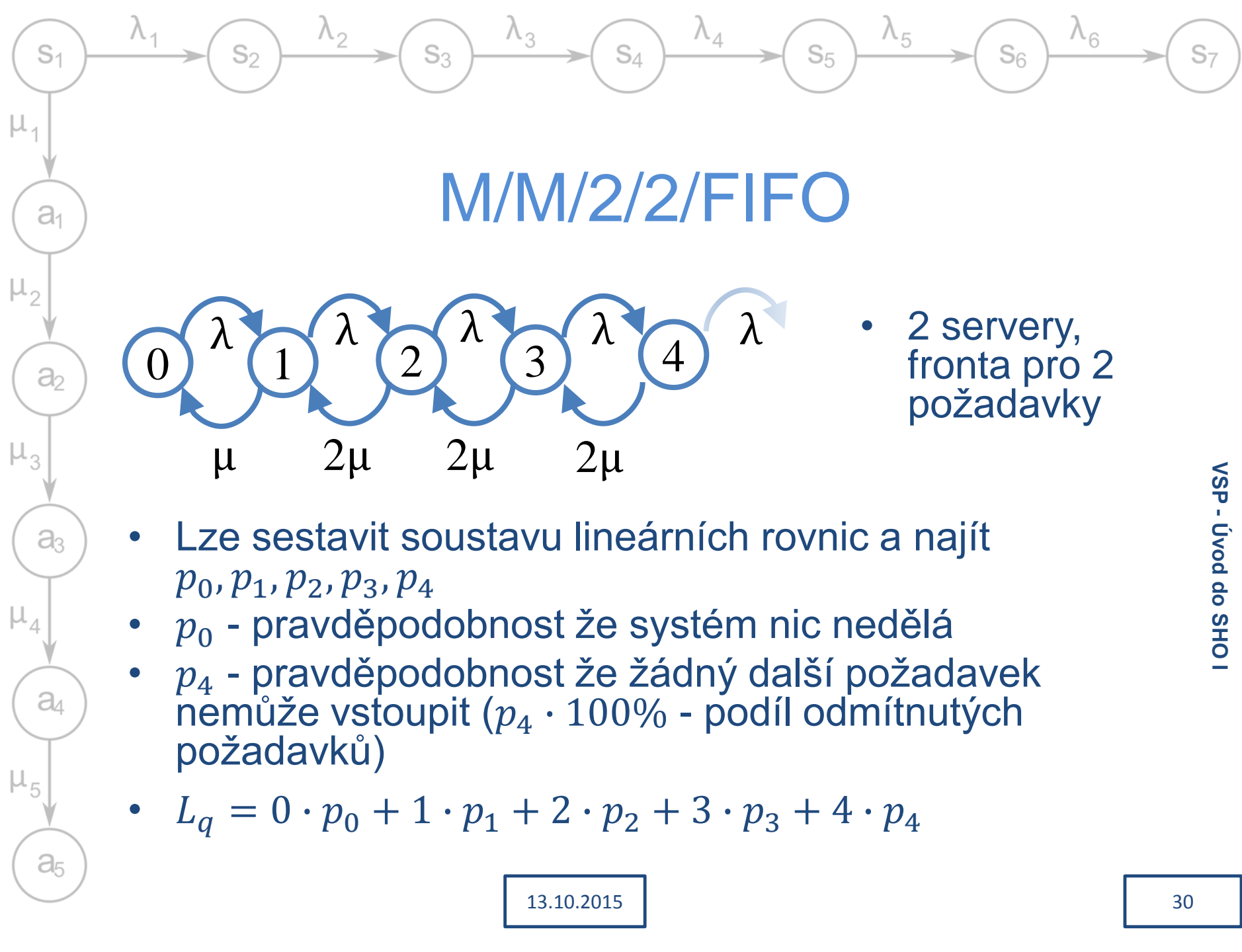
$$- T_q \cong \frac{T_s}{1-\rho^m}$$

Řešení v **Probability, Statistics, and Queuing Theory** (Second Edition) nebo na

[http://irh.inf.unideb.hu/~jsztrik/education/16/SOR\\_Main\\_Angol.pdf](http://irh.inf.unideb.hu/~jsztrik/education/16/SOR_Main_Angol.pdf) (<http://irh.inf.unideb.hu/user/jsztrik/>)



- Konečná fronta  $\rightarrow$  některé požadavky „zahozeny“
  - Zjišťujeme jak často se to stane
  - Nelze použít vzorce pro nekonečnou délku fronty – ne všechny požadavky do systému vstoupí  $\rightarrow \lambda_{reálné} < \lambda_{teoretické}$
- Markovský model konečný  $\rightarrow$  nemusíme se snažit najít limitní přechod
  - Obecné vzorce složité, ale konkrétní se dají snadno odvodit



# M/M/2/2/FIFO

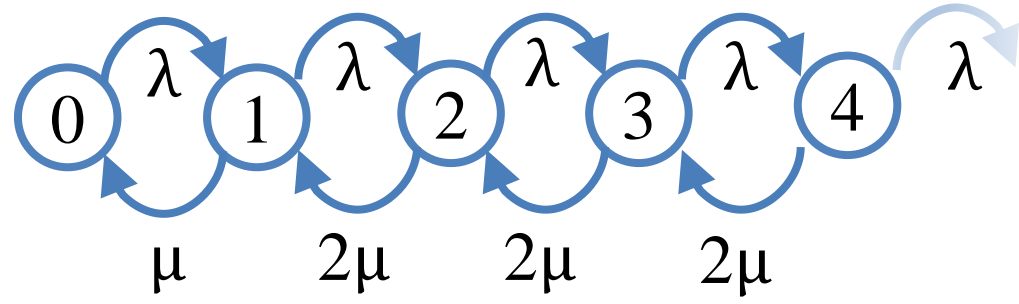
- 2 servery, fronta pro 2 požadavky

- Lze sestavit soustavu lineárních rovnic a najít  $p_0, p_1, p_2, p_3, p_4$
- $p_0$  - pravděpodobnost že systém nic nedělá
- $p_4$  - pravděpodobnost že žádný další požadavek nemůže vstoupit ( $p_4 \cdot 100\%$  - podíl odmítnutých požadavků)
- $L_q = 0 \cdot p_0 + 1 \cdot p_1 + 2 \cdot p_2 + 3 \cdot p_3 + 4 \cdot p_4$

VSP - Úvod do SHO I

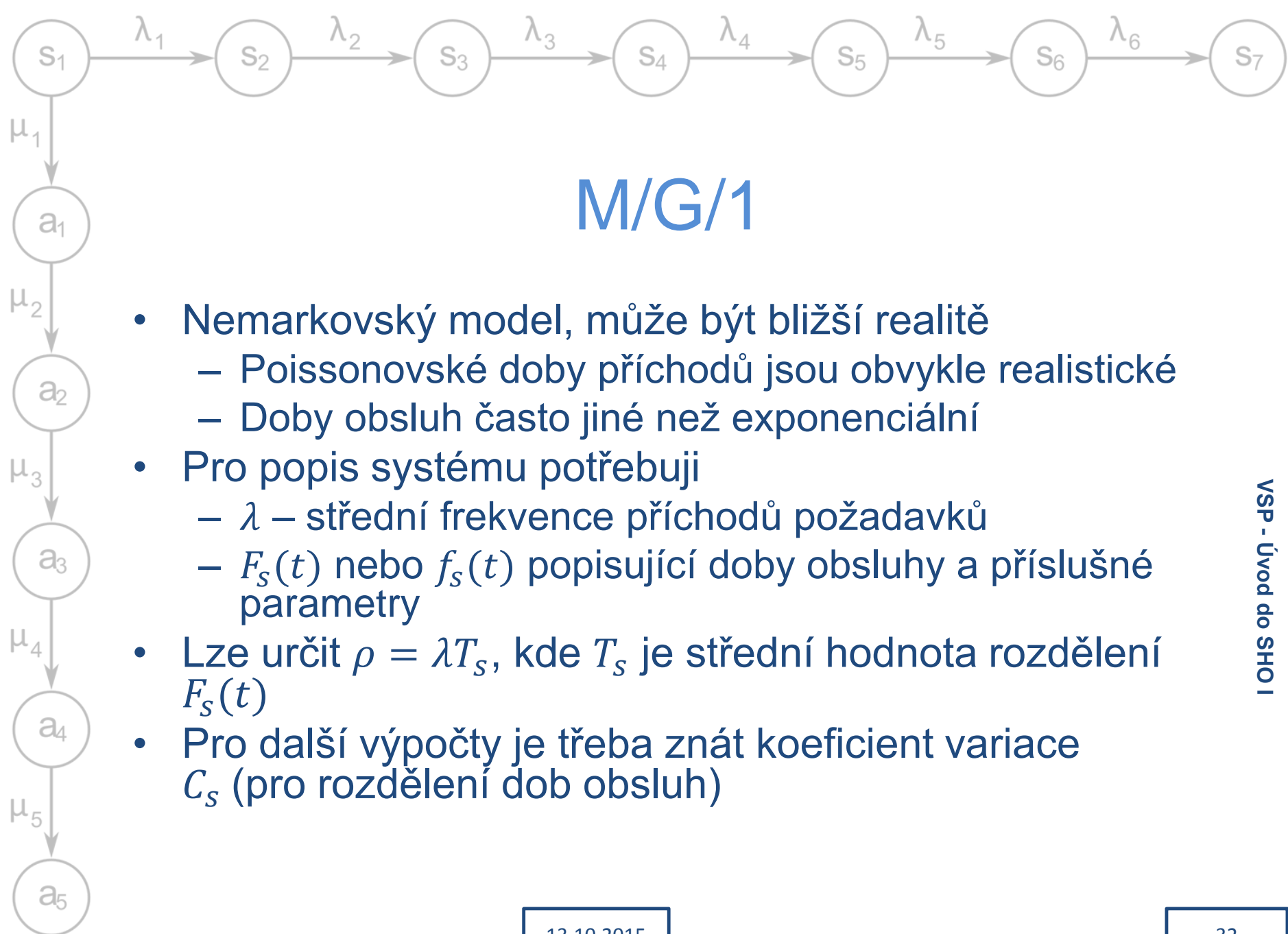


# M/M/2/2/FIFO



- 2 servery, fronta pro 2 požadavky

- $\rho = \frac{p_1}{2} + p_2 + p_3 + p_4$   
– Vážený průměr zatížení v jednotlivých stavech
- $L_w = p_3 + 2p_4$   
– Vážený průměr možných délek front
- Střední frekvence zahození požadavku:  $\lambda p_4$   
→ Frekvence přijetí:  $\lambda(1 - p_4)$  - skutečná frekvence příchodů požadavků
- $T_q = \frac{L_q}{\lambda(1-p_4)}$  (Little -  $L_q = \lambda T_q$ )



- Nemarkovský model, může být bližší realitě
  - Poissonovské doby příchodů jsou obvykle realistické
  - Doby obsluh často jiné než exponenciální
- Pro popis systému potřebují
  - $\lambda$  – střední frekvence příchodů požadavků
  - $F_s(t)$  nebo  $f_s(t)$  popisující doby obsluhy a příslušné parametry
- Lze určit  $\rho = \lambda T_s$ , kde  $T_s$  je střední hodnota rozdělení  $F_s(t)$
- Pro další výpočty je třeba znát koeficient variace  $C_s$  (pro rozdělení dob obsluh)





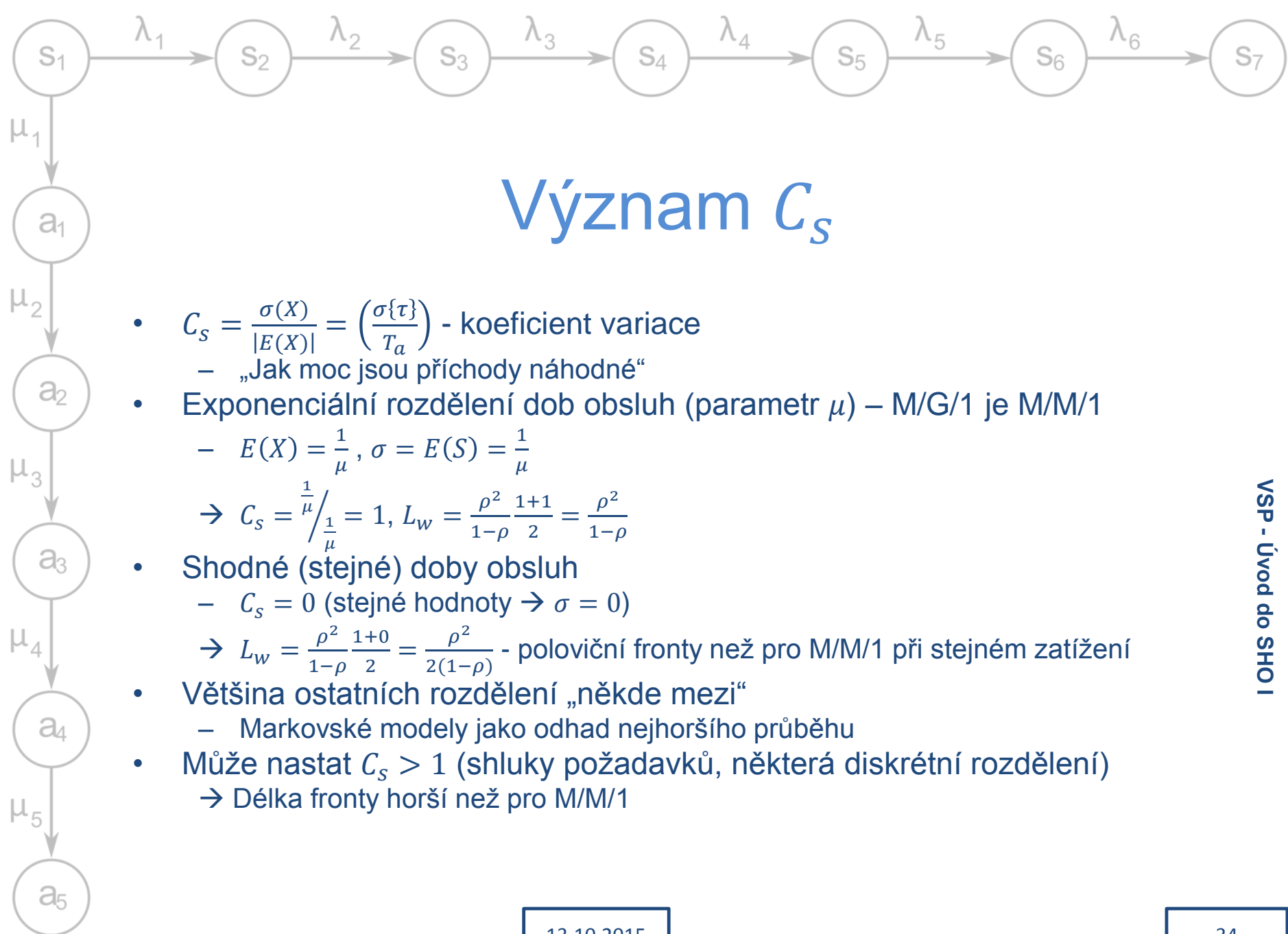
# M/G/1

- $L_w = L_{w(M/M/1)} \frac{1+C_s}{2} = \frac{\rho^2}{1-\rho} \frac{1+C_s}{2}$  (střední délka fronty)

$$- L_q = L_w + L_s = L_w + \frac{\lambda}{\mu} = L_w + \rho$$

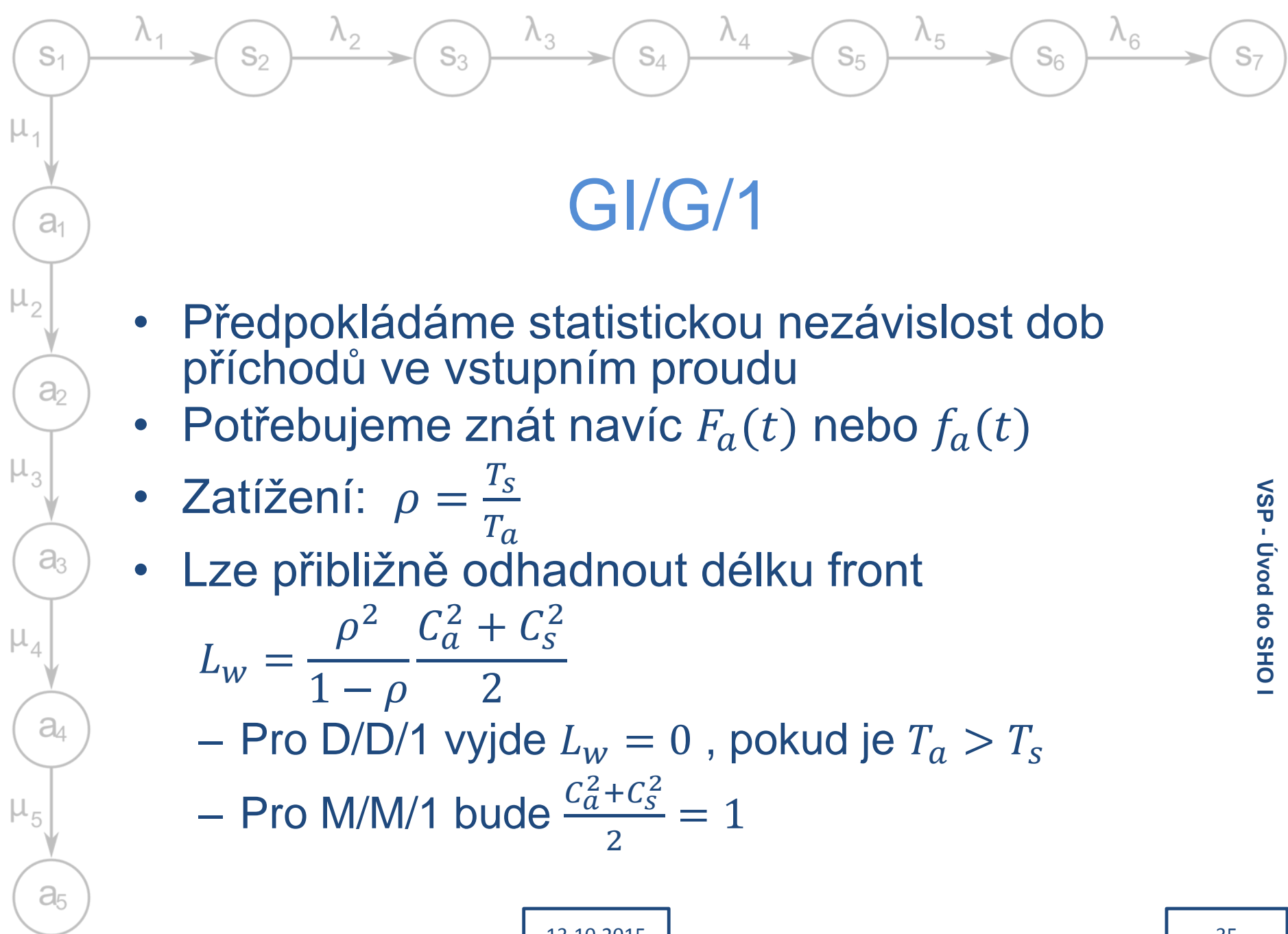
$$- T_q = \frac{L_q}{\lambda}$$

$$- T_w = \frac{L_w}{\lambda}, \quad T_w = T_q - T_s$$



## Význam $C_s$

- $C_s = \frac{\sigma(X)}{|E(X)|} = \left( \frac{\sigma\{\tau\}}{T_a} \right)$  - koeficient variace
  - „Jak moc jsou příchody náhodné“
- Exponenciální rozdělení dob obsluh (parametr  $\mu$ ) – M/G/1 je M/M/1
  - $E(X) = \frac{1}{\mu}$ ,  $\sigma = E(S) = \frac{1}{\mu}$
  - $C_s = \frac{\frac{1}{\mu}}{\frac{1}{\mu}} = 1$ ,  $L_w = \frac{\rho^2}{1-\rho} \frac{1+1}{2} = \frac{\rho^2}{1-\rho}$
- Shodné (stejně) doby obsluh
  - $C_s = 0$  (stejně hodnoty →  $\sigma = 0$ )
  - $L_w = \frac{\rho^2}{1-\rho} \frac{1+0}{2} = \frac{\rho^2}{2(1-\rho)}$  - poloviční fronty než pro M/M/1 při stejném zatížení
- Většina ostatních rozdělení „někde mezi“
  - Markovské modely jako odhad nejhoršího průběhu
- Může nastat  $C_s > 1$  (shluky požadavků, některá diskrétní rozdělení)
  - Délka fronty horší než pro M/M/1



# GI/G/1

- Předpokládáme statistickou nezávislost dob příchodů ve vstupním proudu
- Potřebujeme znát navíc  $F_a(t)$  nebo  $f_a(t)$
- Zatížení:  $\rho = \frac{T_s}{T_a}$
- Lze přibližně odhadnout délku front

$$L_w = \frac{\rho^2}{1 - \rho} \frac{C_a^2 + C_s^2}{2}$$

– Pro D/D/1 vyjde  $L_w = 0$  , pokud je  $T_a > T_s$

– Pro M/M/1 bude  $\frac{C_a^2 + C_s^2}{2} = 1$



Děkuji za pozornost

- Příště generátory náhodných čísel a metoda Monte Carlo