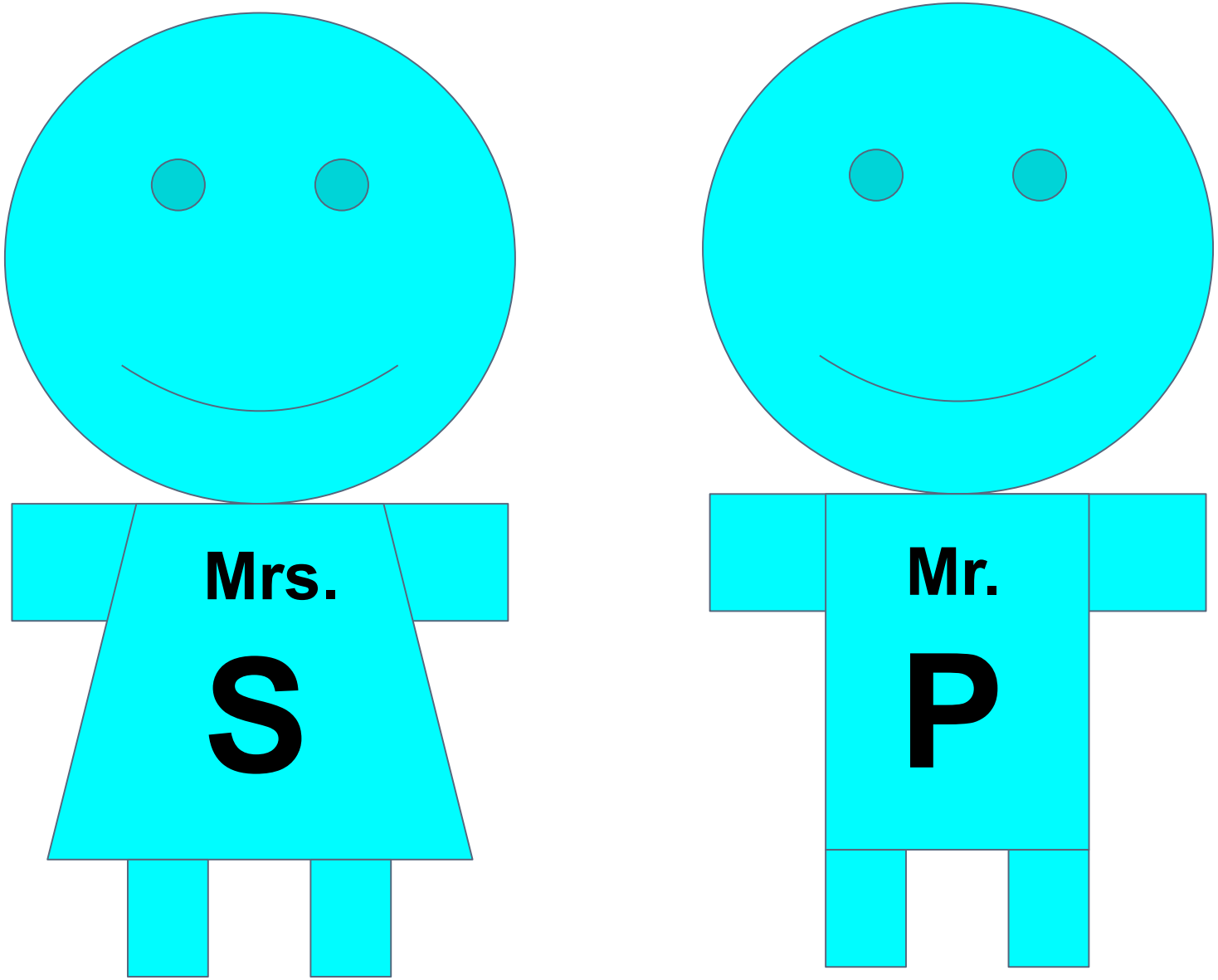
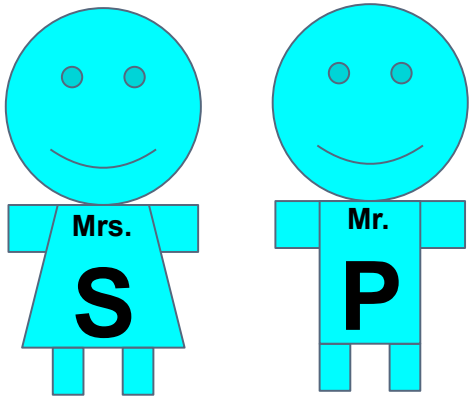
Shaping a QoS

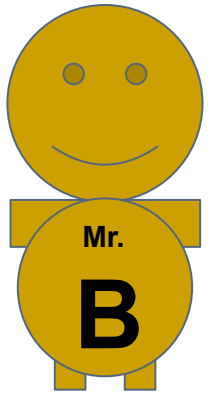
Karel Řeřicha

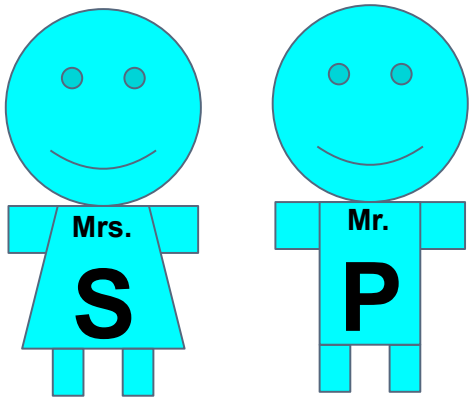
Poslední revize: 12.10.2017

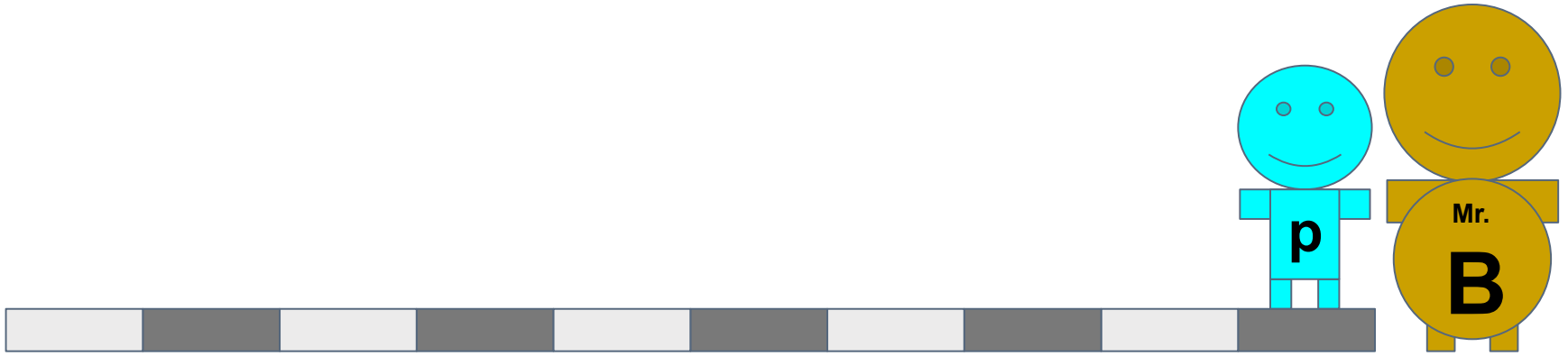

Mrs.
S
Mr.
P

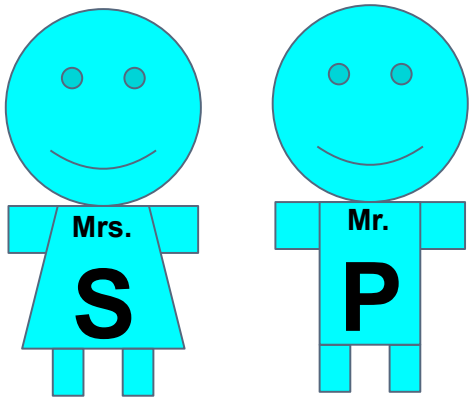

Mrs.
S

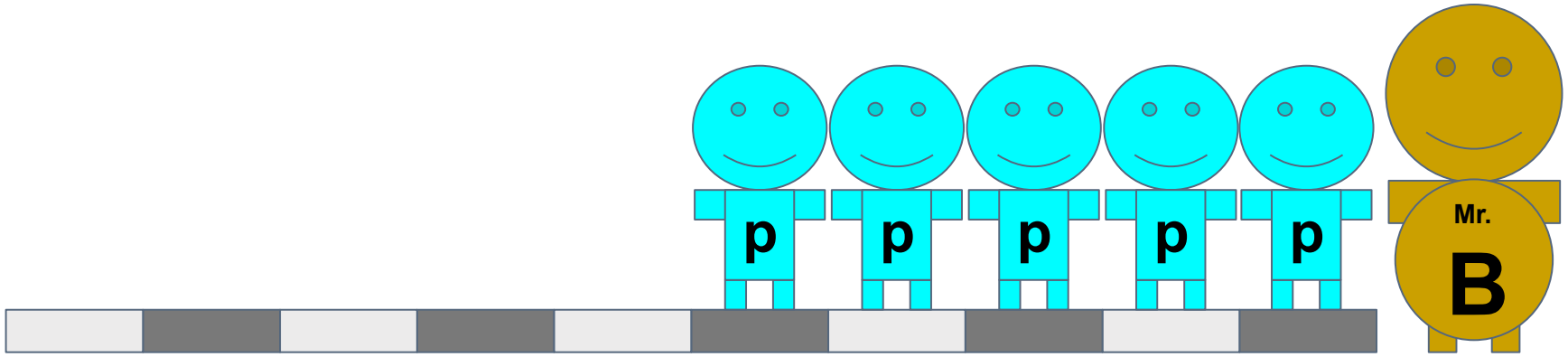
Mr.
P

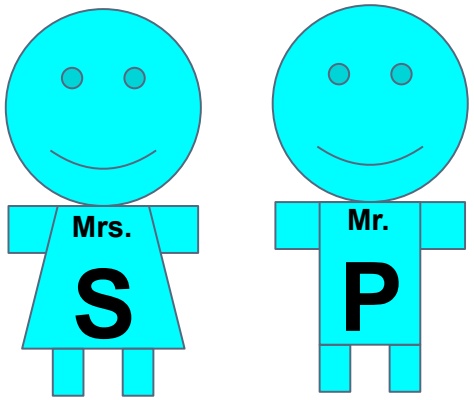

Mr.
B

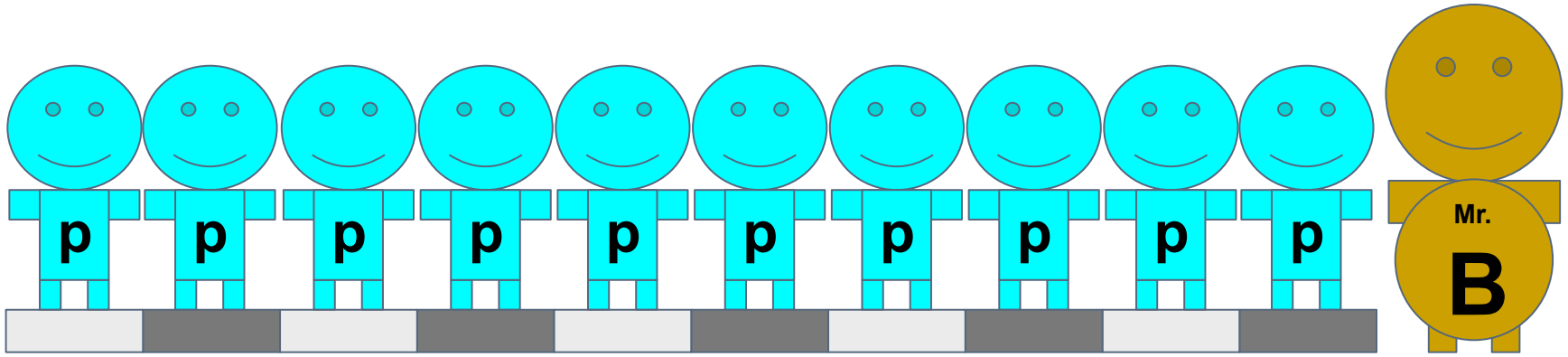

Mrs.
S
Mr.
P

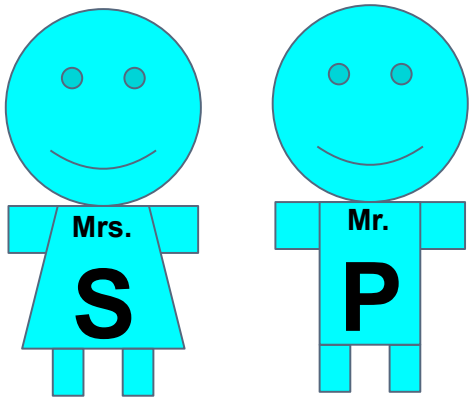

p
Mr.
B

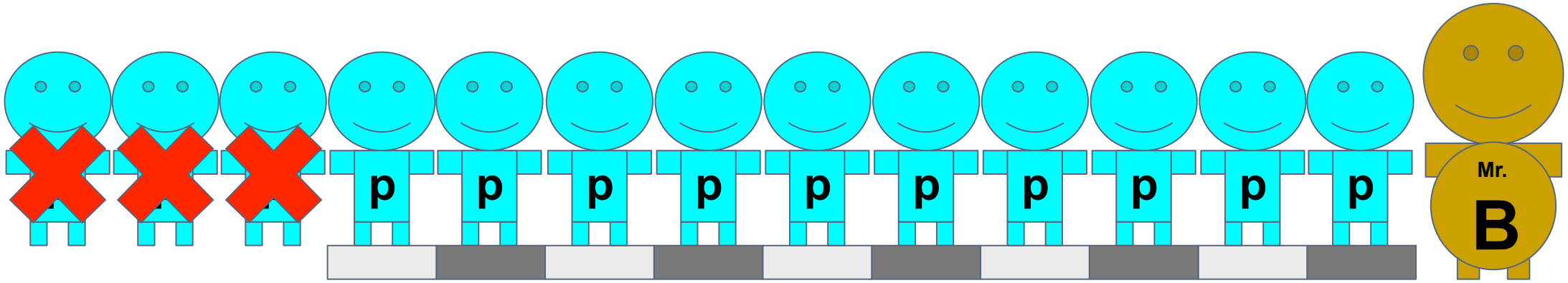

Mrs.
S
Mr.
P

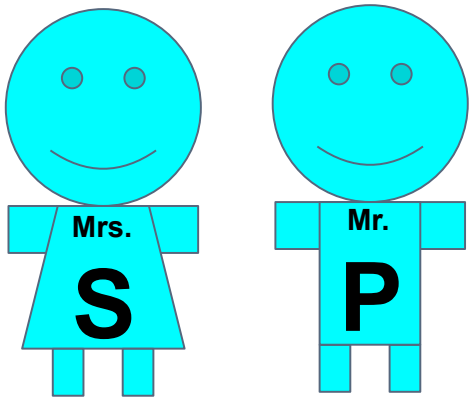

p
p
p
p
p
Mr.
B

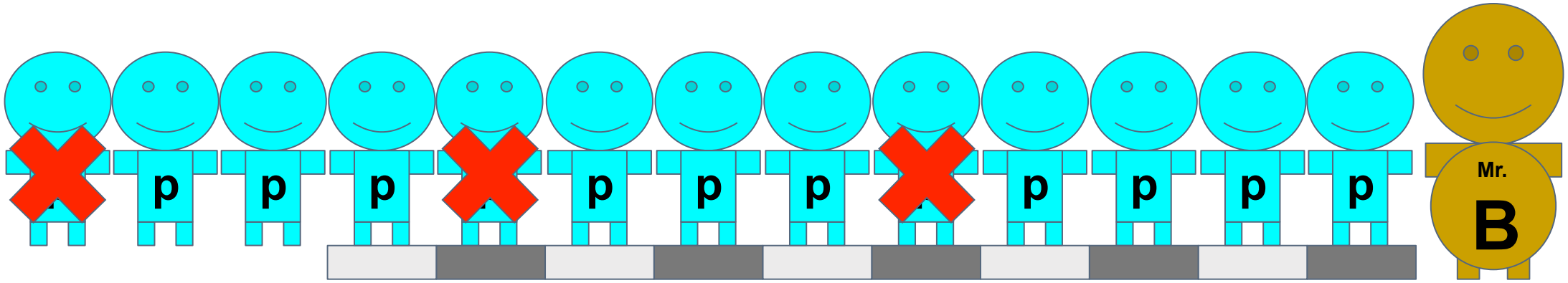

Mrs.
S
Mr.
P


p
p
p
p
p
p
p
p
p
p
Mr.
B

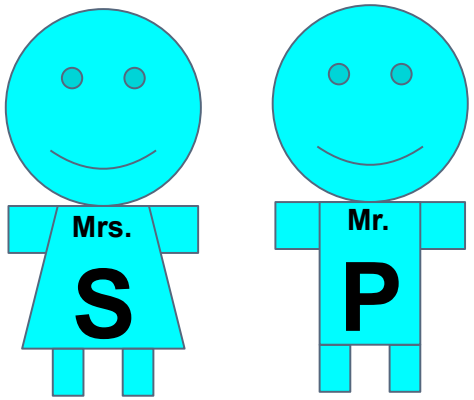

Mrs.
S
Mr.
P

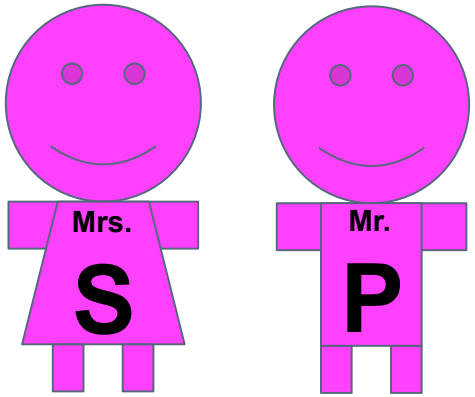

p
p
p
p
p
p
p
p
p
p
Mr.
B

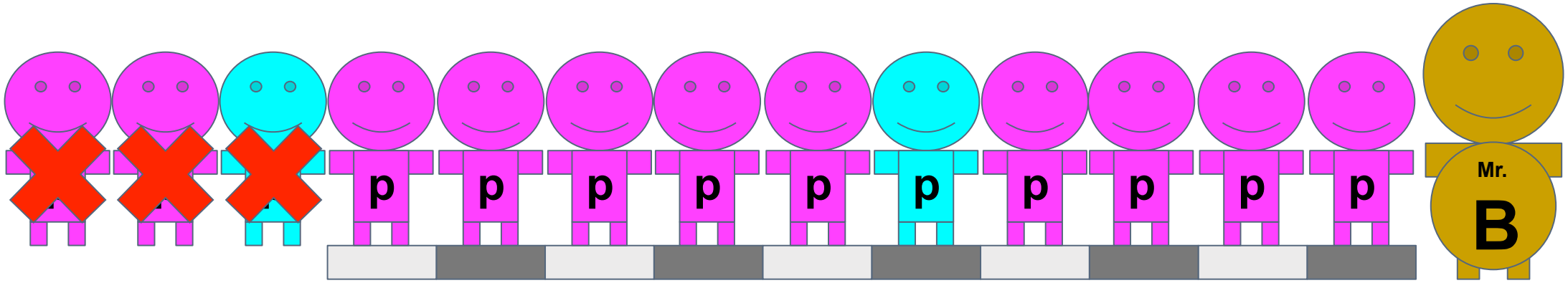

Mrs.
S
Mr.
P


p
p
p
p
p
p
p
p
p
p
p
Mr.
B

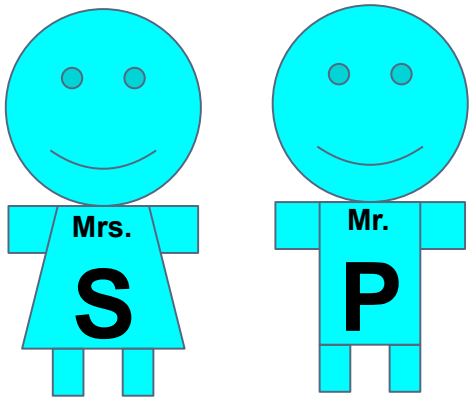
FIFO

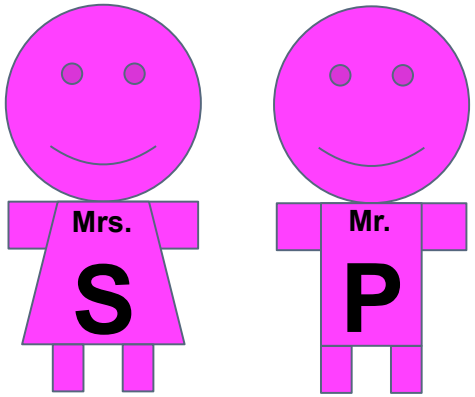


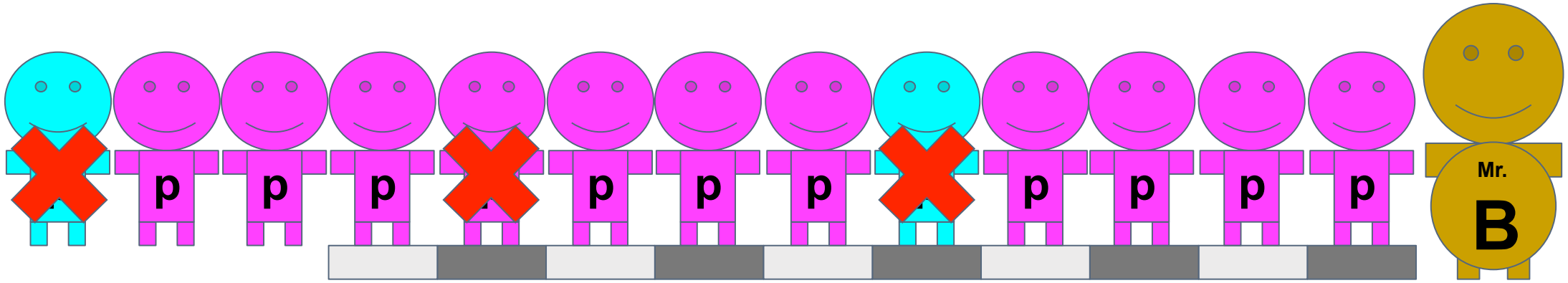




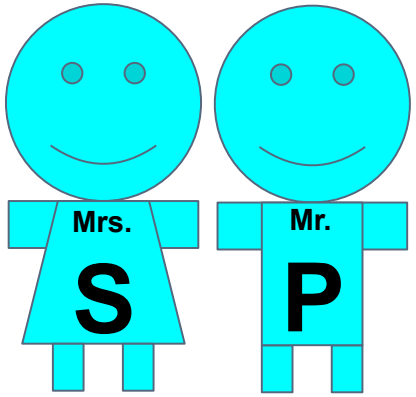
CoDel

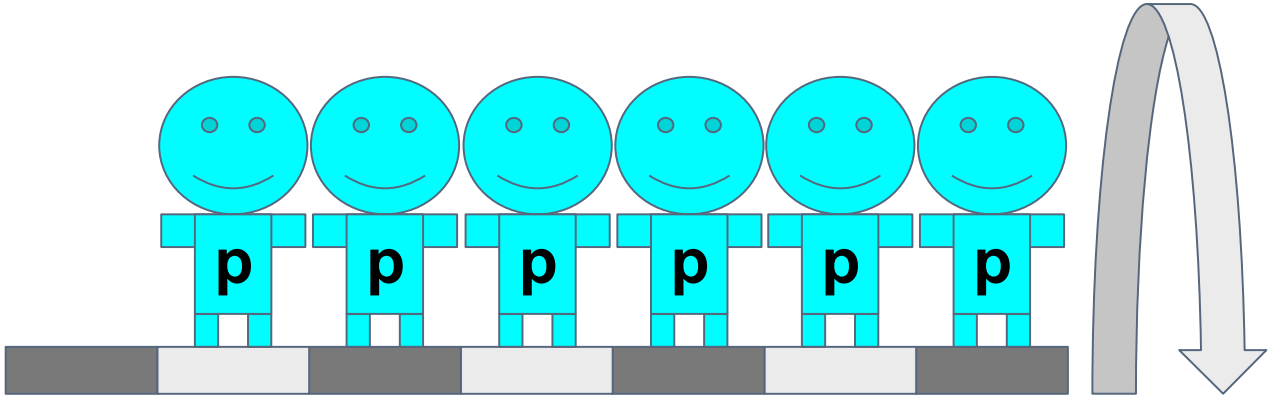


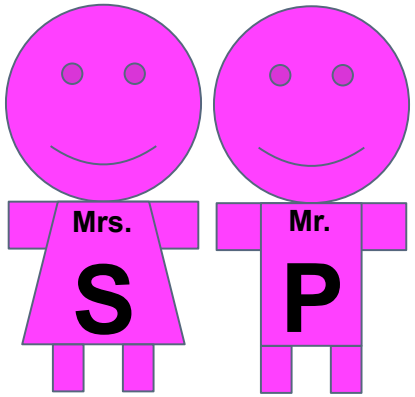


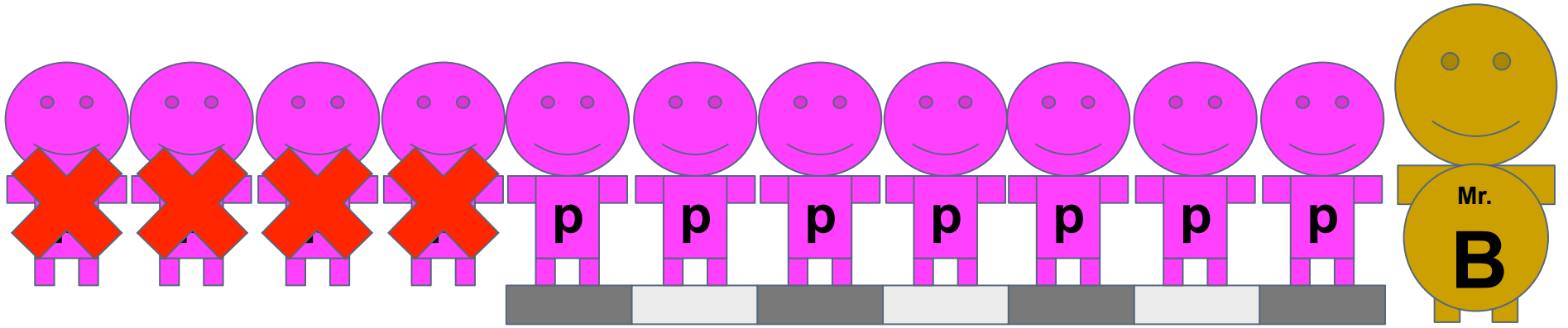


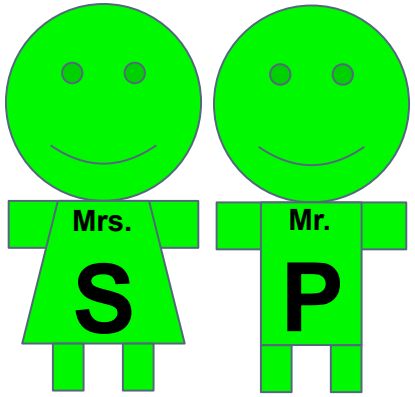
FQ, SFQ

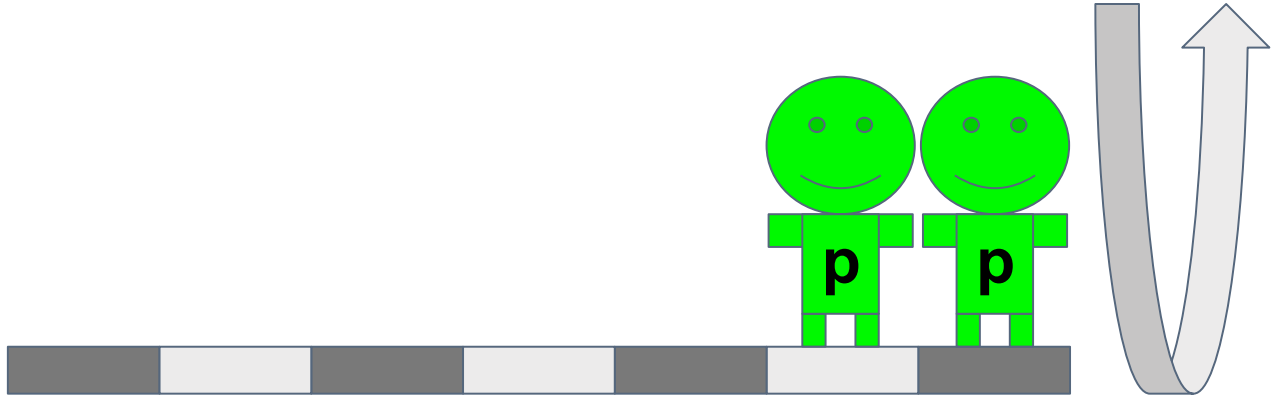




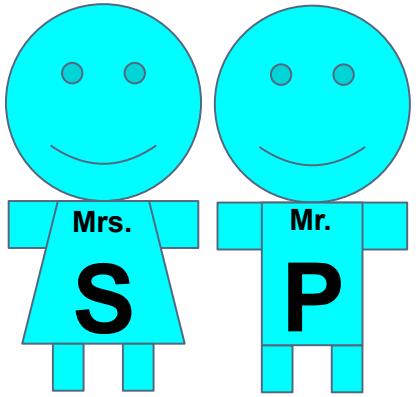


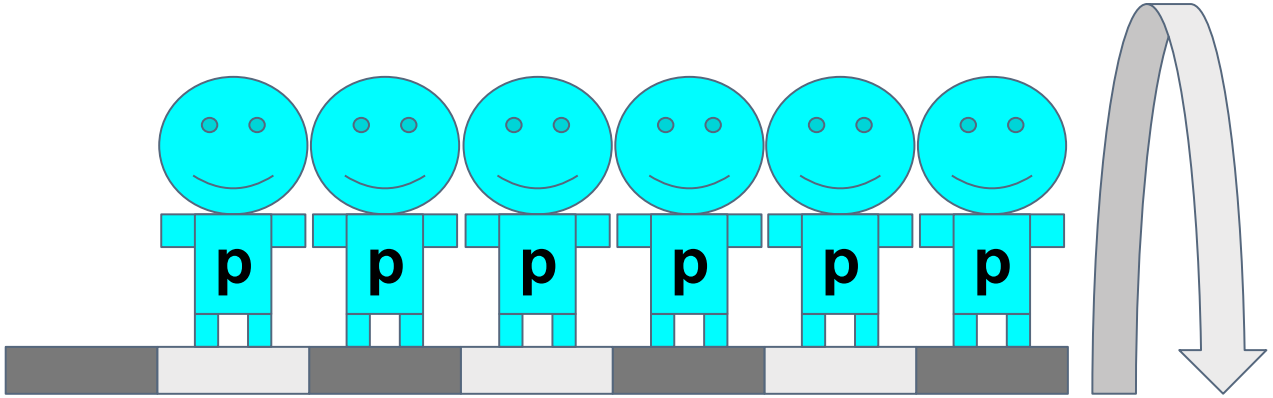


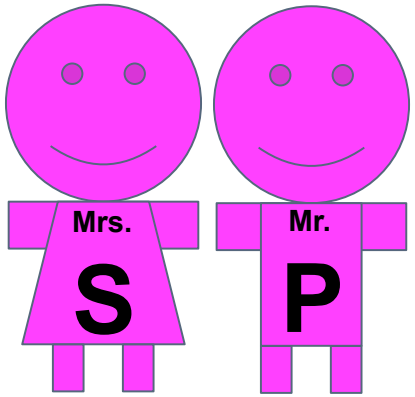


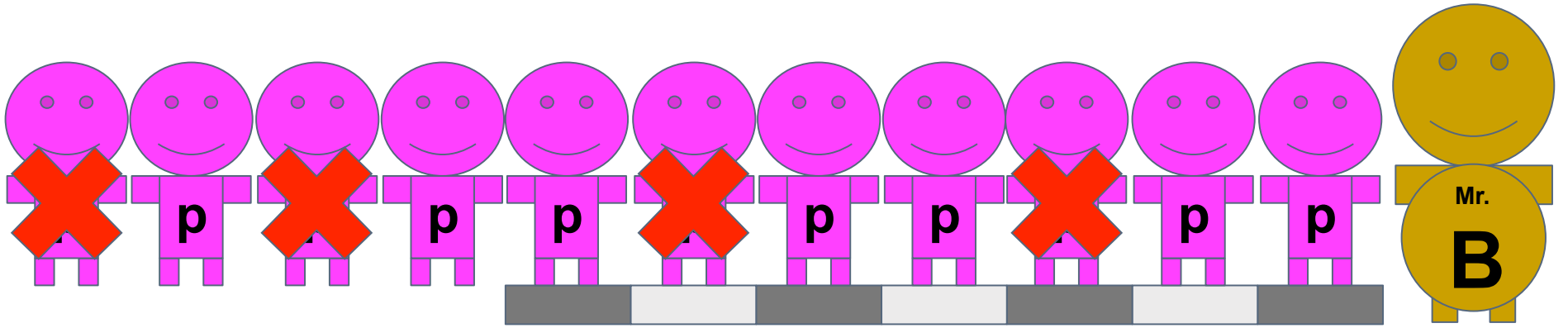


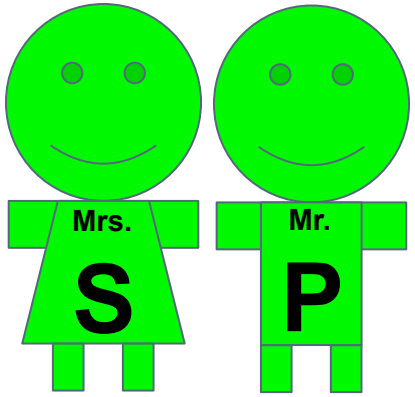
FQ_CoDel

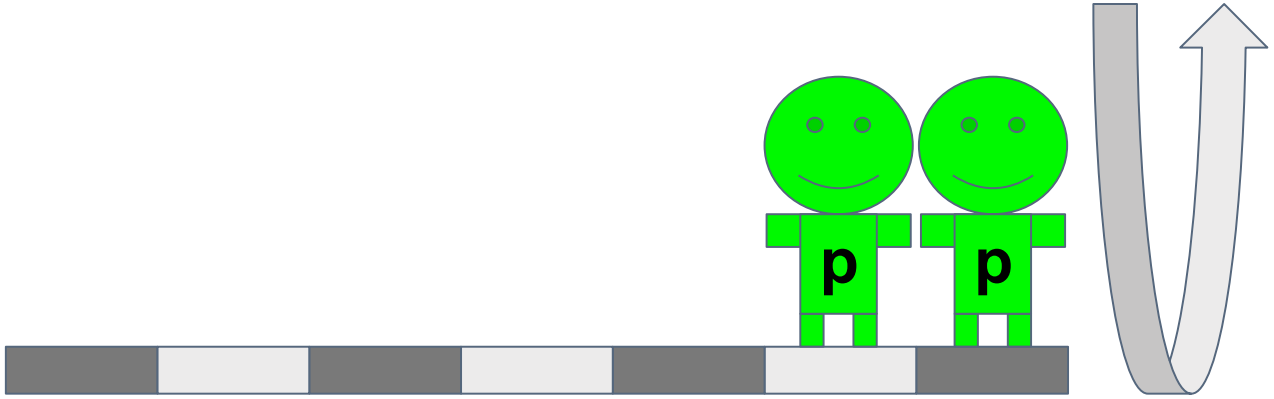




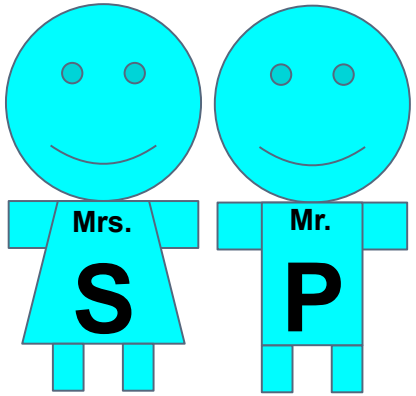


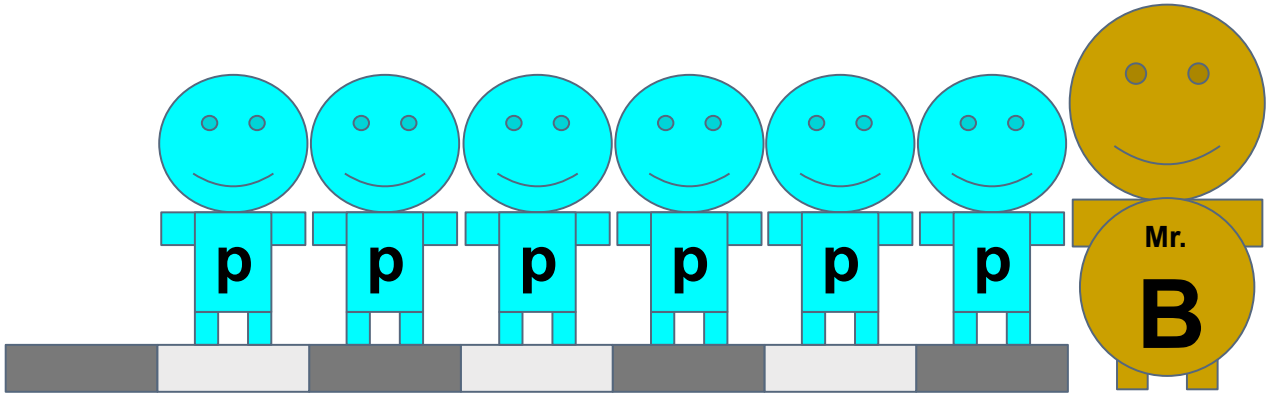


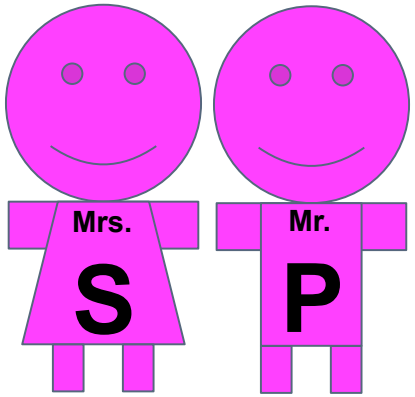


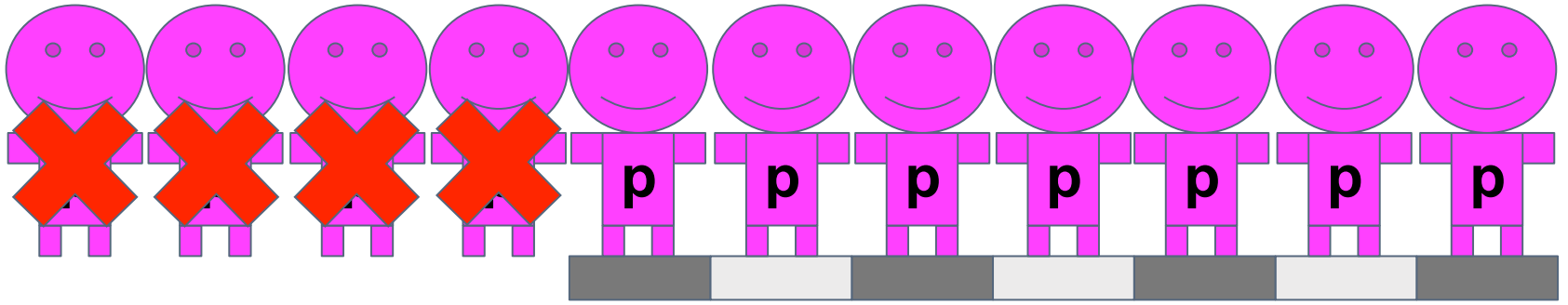


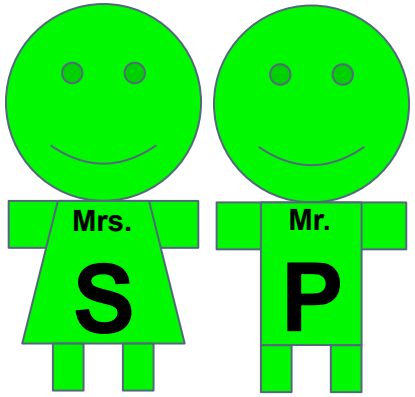
prioritizace absolutní



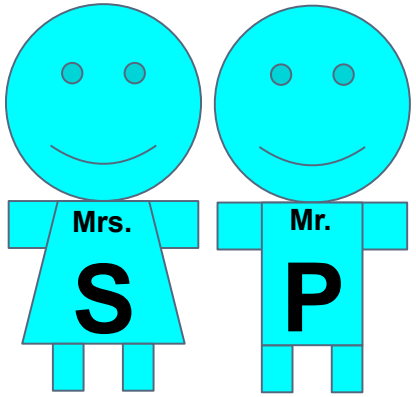


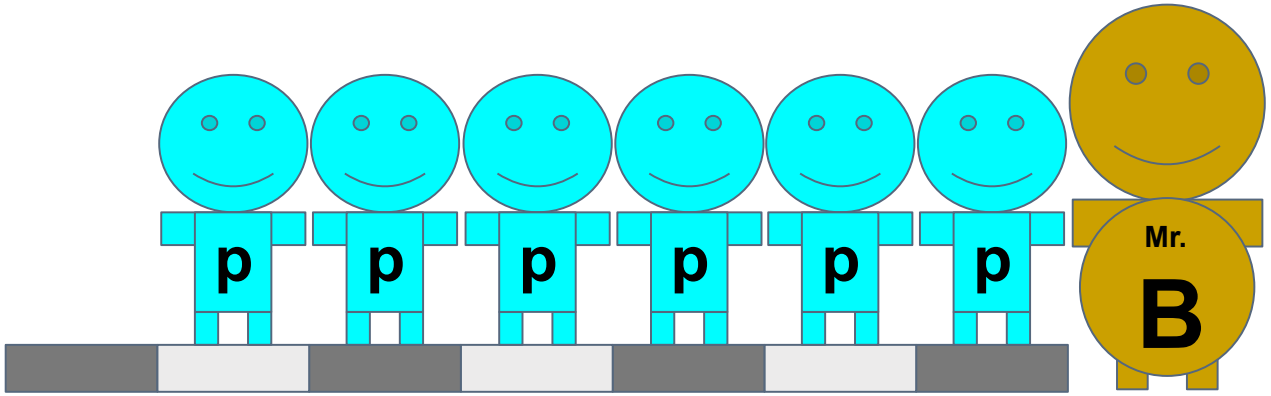


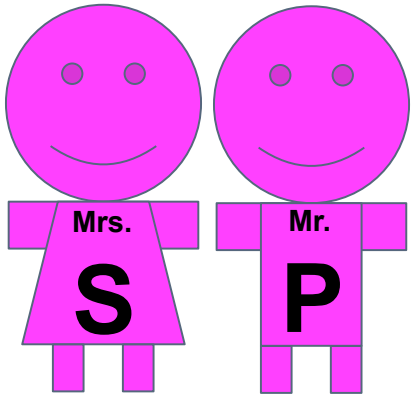


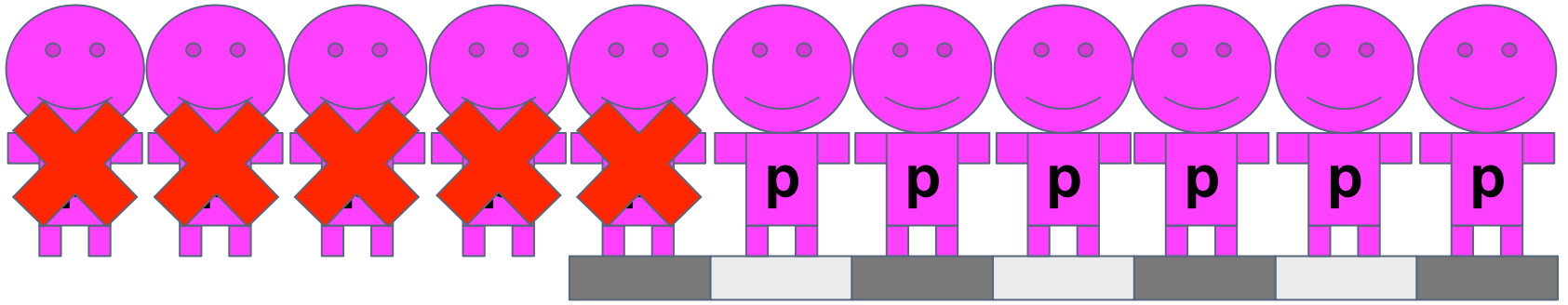


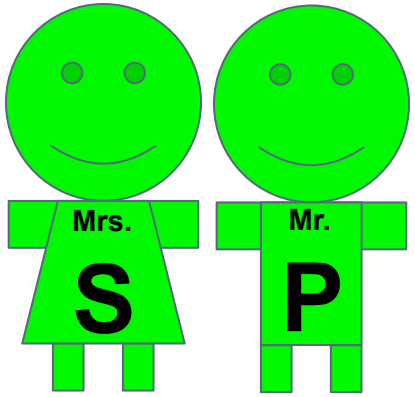
prioritizace poměrná



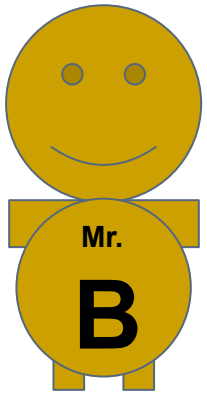








Co dělá pan Buffer?



maká do roztrhání → **work conserving**

fifo, sfq, fq, codel, fq_codel

jede na normu → **work non-conserving**

htb, hfsc

QoS - Quality of Service

- řízení datových toků v sítích s předáváním paketů
- postupy, které zaručují, že daná internetová přípojka bude mít co nejvyšší kvalitu za daných podmínek

Kdy QoS použít ? - Úzká místa = špatně

- nedostatečná kapacita transportní sítě
- nedostatečná kapacita poslední míle
- efekt zahlcení přenosového média
- efekt nerovnoměrného čerpání přidělené kapacity v rámci jednotlivých datových toků
- nekvalitní médium

Co s tím ?

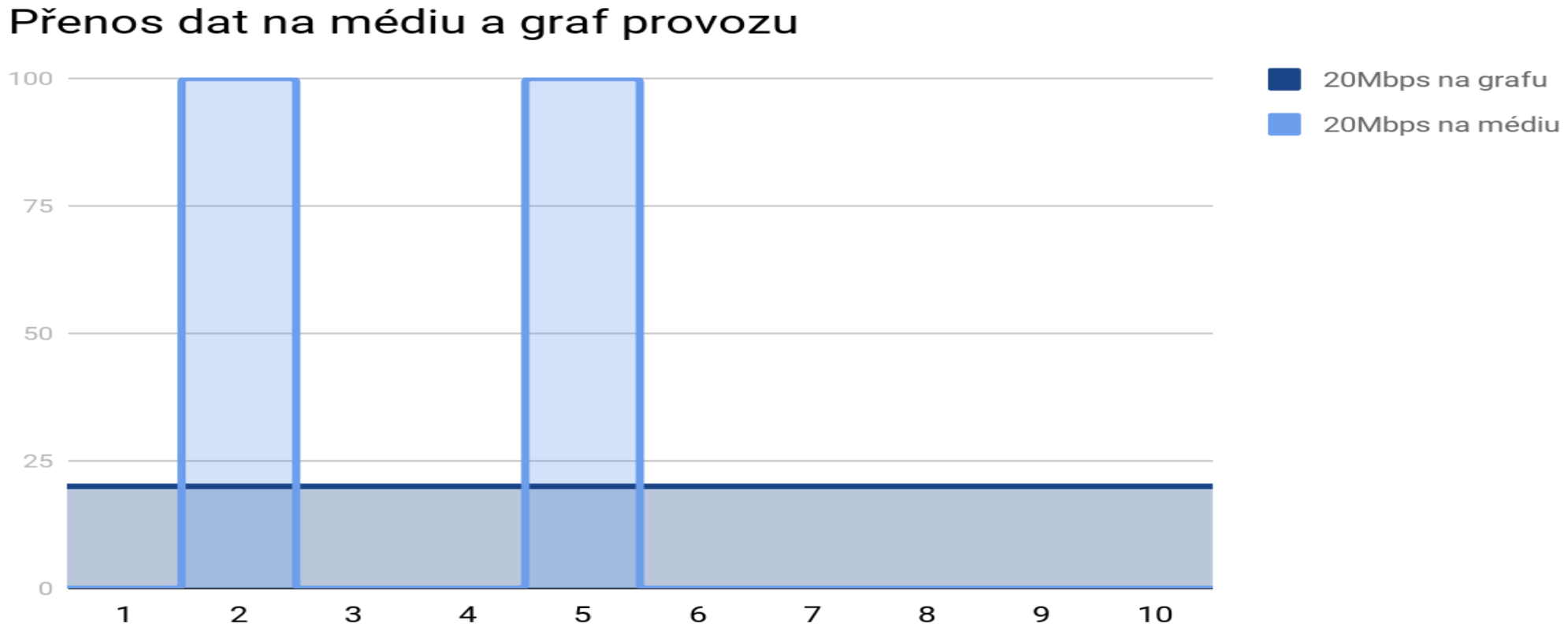
- Lze rozšířit úzké místo ? Rozšiřme ho!
- Nasaďme QoS

Metody QoS

- alokace a řízení šířky a sdílení pásma - shaping
- prioritizace provozu
- řízení provozu

Jak to vypadá na přenosovém médiu

- sítě s předáváním rámců a paketů - data cestují v blocích
- na většině běžně používaných médií jsou data předávána maximální fyzickou rychlostí média



Příklad okruhu

- 1Gbps ethernet, 300Mbps rádio
- při provozu 100Mbps dochází k výpadkům paketů a nevíme proč?

Zahlcení bufferu

- Rádio má malý buffer, data do něj cestují rychlostí 1Gbps a nestačí se rychlostí 300Mbps vysílat

Řešení - Flow control

- přijímací strana hlásí vysílací, že má plný buffer, a žádá o pozastavení vysílání
- POZOR: přenos problému dál

Buffer - co tam s pakety ?

- když prázdný, tak nic :D
- pozdržet vysílání - SHAPING
- přeřazovat pořadí paketů - PRIORITIZACE
- zahodit pakety - POLICING

Policing

- dropování paketů na vstupu
- vysokorychlostní okruhy, nízké nároky na HW
- nebudeme se zabývat :)

Prioritizace

- přeřazování paketů ve vysílacím bufferu
- standardně buffery FIFO (pfifo), žádná prioritizace, best effort
- prioritizujeme provoz nejvíce citlivý na zpoždění a výpadky
- POZOR: při prioritizaci nesmíme přehazovat pakety v jednotlivých tocích

Explicitní prioritizace

- pakety označeny už při vysílání (nebo na cestě)
- většinou bohužel absolutní (nedobře)

TOS

- 3 bity precedence (priority), 5 bitů type of service

DSCP

- Náhrada za TOS (1998), 6 bitů, z toho první tři zpětně kompatibilní s precedence v TOS

ECN

- rozšíření DSCP o dva bity (2001)
- pokud uzel na konci konexe dostane paket s nastaveným ECN, zpomalí vysílání v daném flow
- dnes podporují všechny běžně dostupné operační systémy
- kromě shapingu nejzásadnější věc pro QoS, protože v úzkém místě nastavíte dle potřeby ECN a snížení rychlosti vysílání paketů provádějí komunikující uzly, tj. nejedná se o dropování paketů či zvýšení latencí
- FQ_CoDel má v sobě !!!

Prioritizace - work conserving disciplíny

- pakety rozděleny do toků, přidělování vysílací kapacity

PRIO

- defaultně tři třídy, klasifikace paketů pomocí filtrů, v každé třídě defaultně FIFO fronta

SFQ

- rozdělení provozu do jednotlivých FIFO front dle hashe hlavičky paketu, z front pak vysíláno round-robin
- problém se změnou hash funkce - přeřazení paketů v toku

CoDel

- Controlled Delay
- začíná proporcionálně dropovat pakety, jakmile délka fronty způsobuje zpoždění větší jak 5 ms

FQ_CoDel

- rozdělení provozu do toků (def. 1024), v každém pak CoDel
- aktuálně asi nejlepší, co je možné nasadit
- při zaplňování fronty nejdříve nastavuje ECN, až pak dropuje

Shaping - work non-conserving disciplíny

- pozdržují vysílání paketů, i když je linka volná
- hierarichické disciplíny - klasifikace paketů do hierarchických tříd

HTB

- timer based, nepřesná na rychlých linkách
- nevhodné rozdělování zbývající šířky pásma u velmi rozdílných tříd

HFSC

- tři křivky, každá znamená úplně něco jiného - matoucí
- Realtime - matoucí název, lze definovat všude, i když má smysl pouze u leaf tříd, chybí kontrola na overbooking
- Link Sharing - je udávána v bps, i když vlastně vyjadřuje poměr sdílení přebytečné kapacity
- Upper Limit - omezení jednotlivých tříd na maximální rychlost

Výhody HFSC:

- velice přesná ve srovnání s HTB
- každá křivka může mít dvě části, burst-delay-backlogged, možnost implementace burstů
- prioritizace pomocí link sharing není absolutní
- realtime křivka pro zaručenou kapacitu

Praktická implementace shapingu na Linuxu

Klasifikace paketů

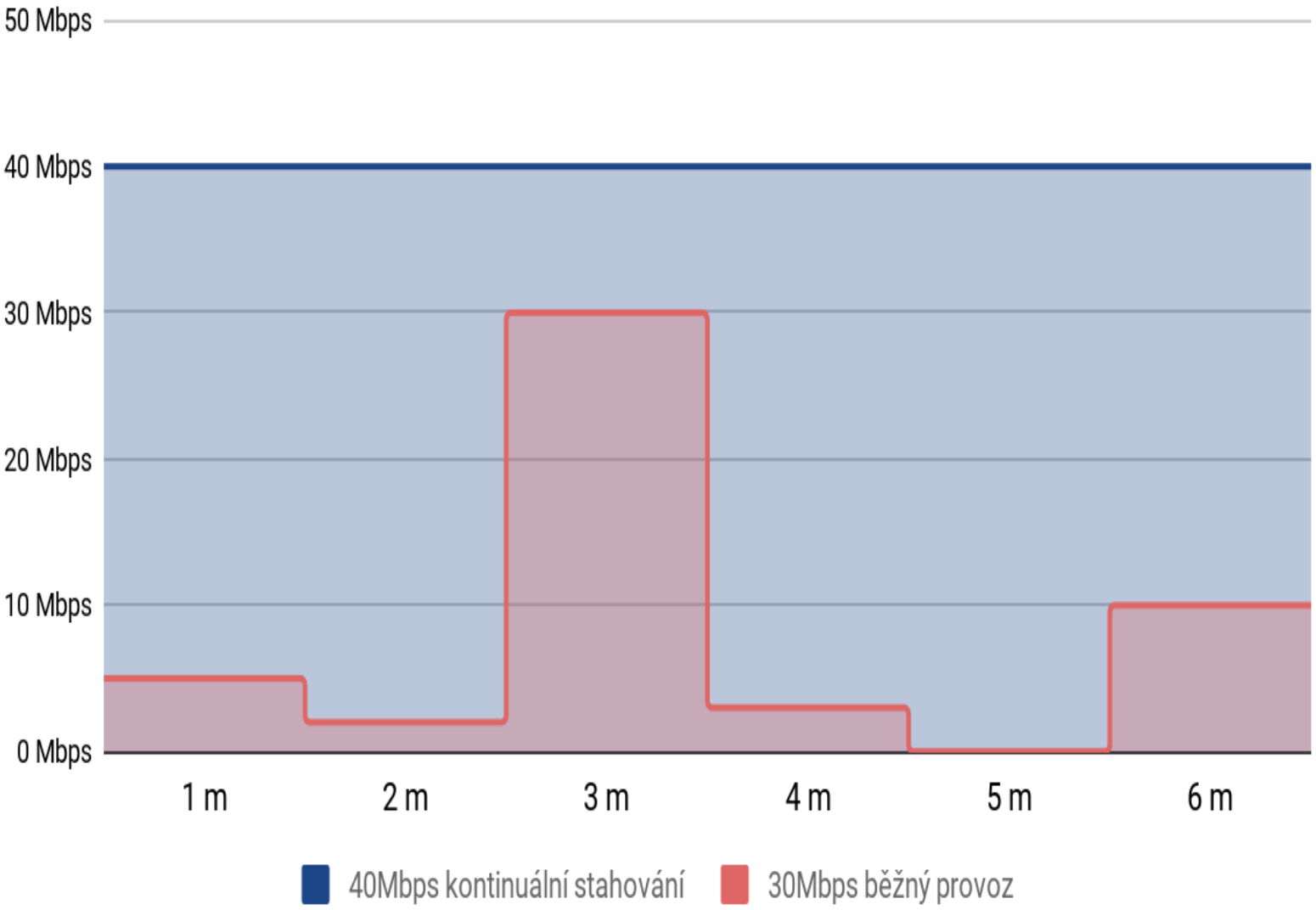
- iptables - aktulizace kopírováním do userspace, lagy, není thread safe
- nftables - náhrada iptables, řeší nedostatky, datové struktury

Na co dát pozor

- omezení handles na 16bit, tj. max 65k tříd
- nutná affinita NIC front na vyhrazená CPU jádra, jinak jitter a případně dropy (Mikrotik ?)
- pro vyšší rychlosti vícenásobné NIC fronty, použít offloading

Shaping bez burstu, priorit a s dostatečnou kapacitou linky

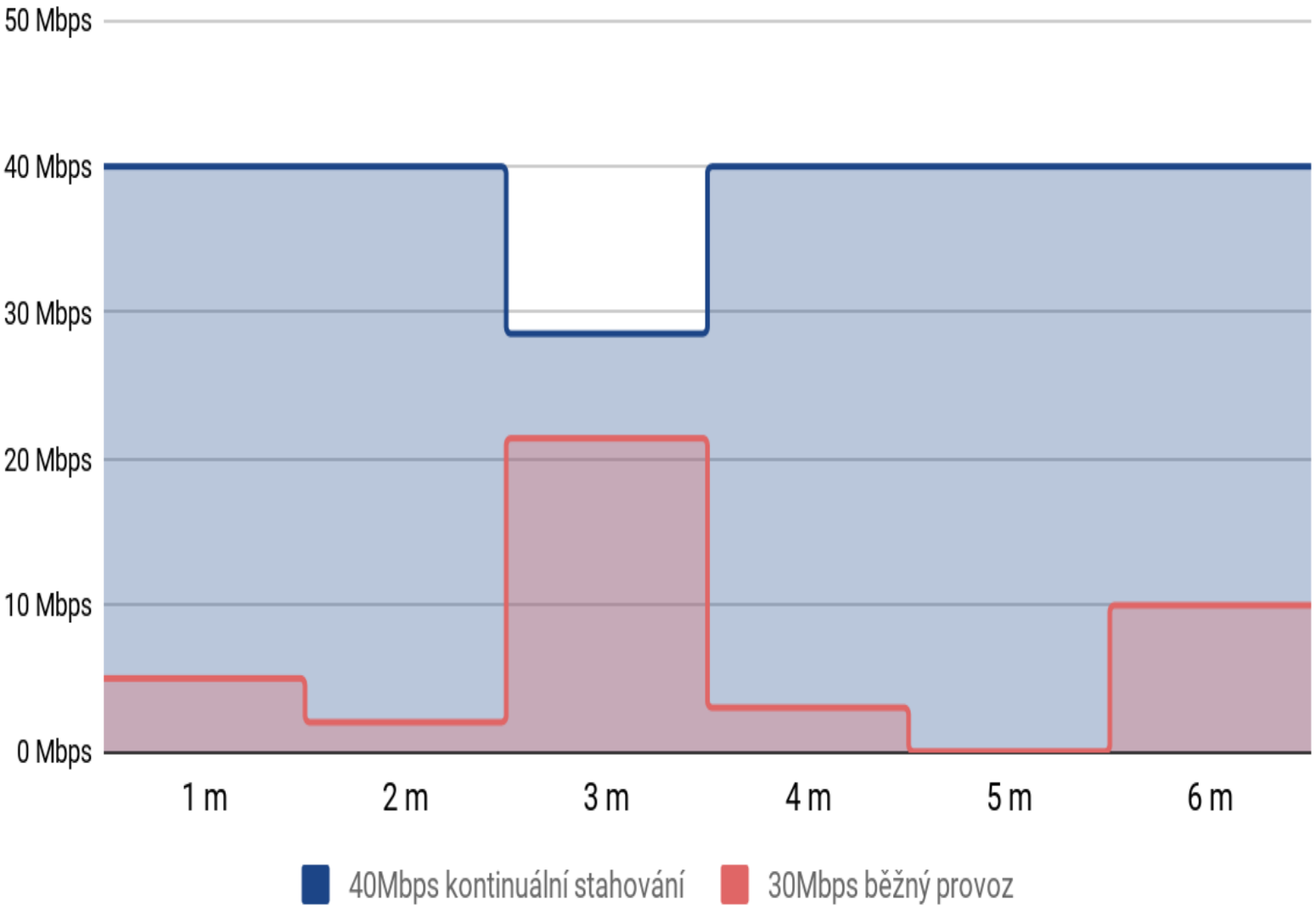
Dva zákazníci, 40Mbps kontinuálně stahuje, 30Mbps běžně využívá linku



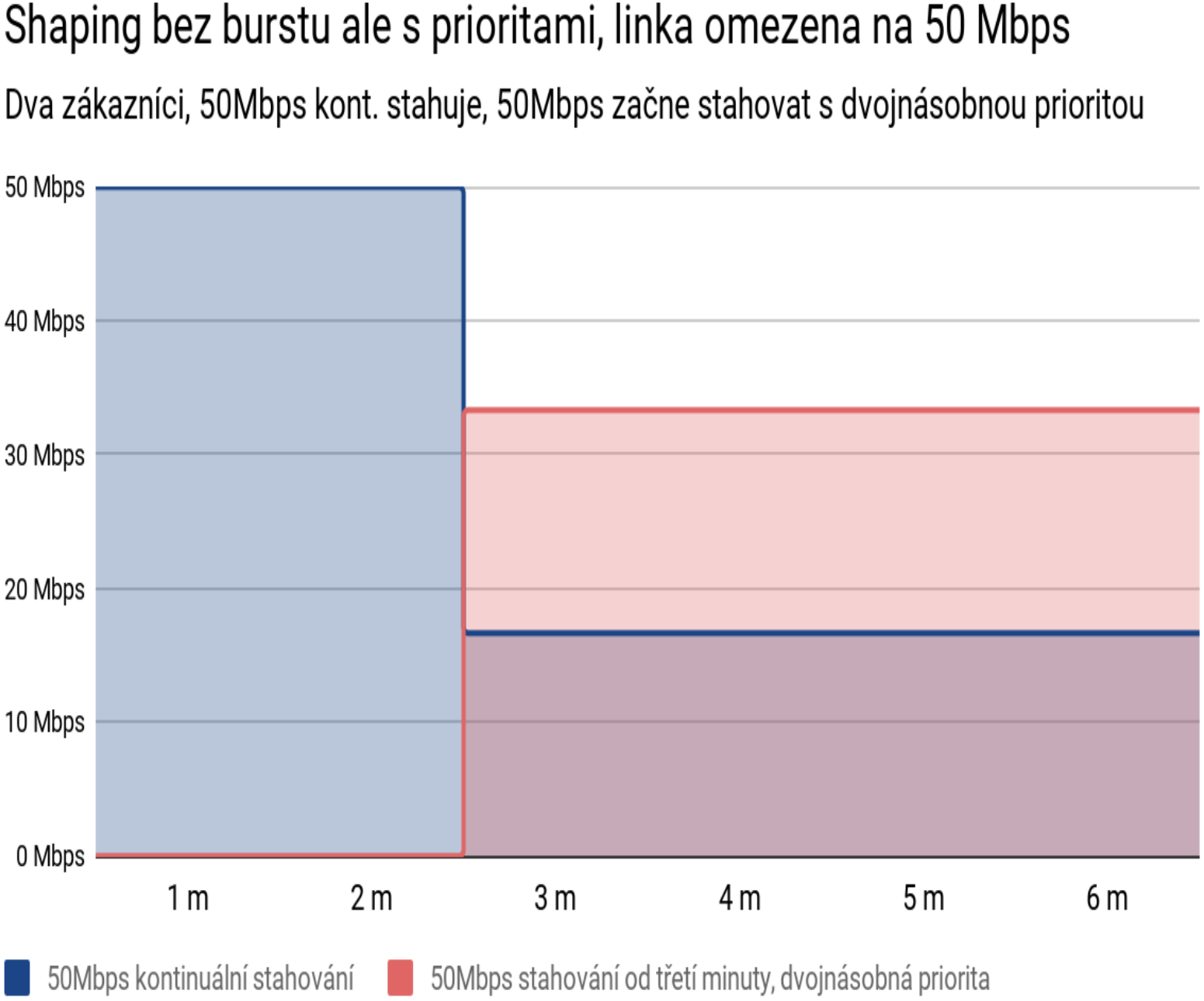
- běžný shaping, kde se pouze specifikuje maximální rychlost, které může zákazník dosáhnout
- graf ukazuje pouze jeden směr (například download), druhý se chová obdobně
- pokud je celková kapacita linky větší, než součet maximální rychlosti obou zákazníků, tak to není problém

Shaping bez burstu a priorit, linka omezena na 50 Mbps

Dva zákazníci, 40Mbps kontinuálně stahuje, 30Mbps běžně využívá linku



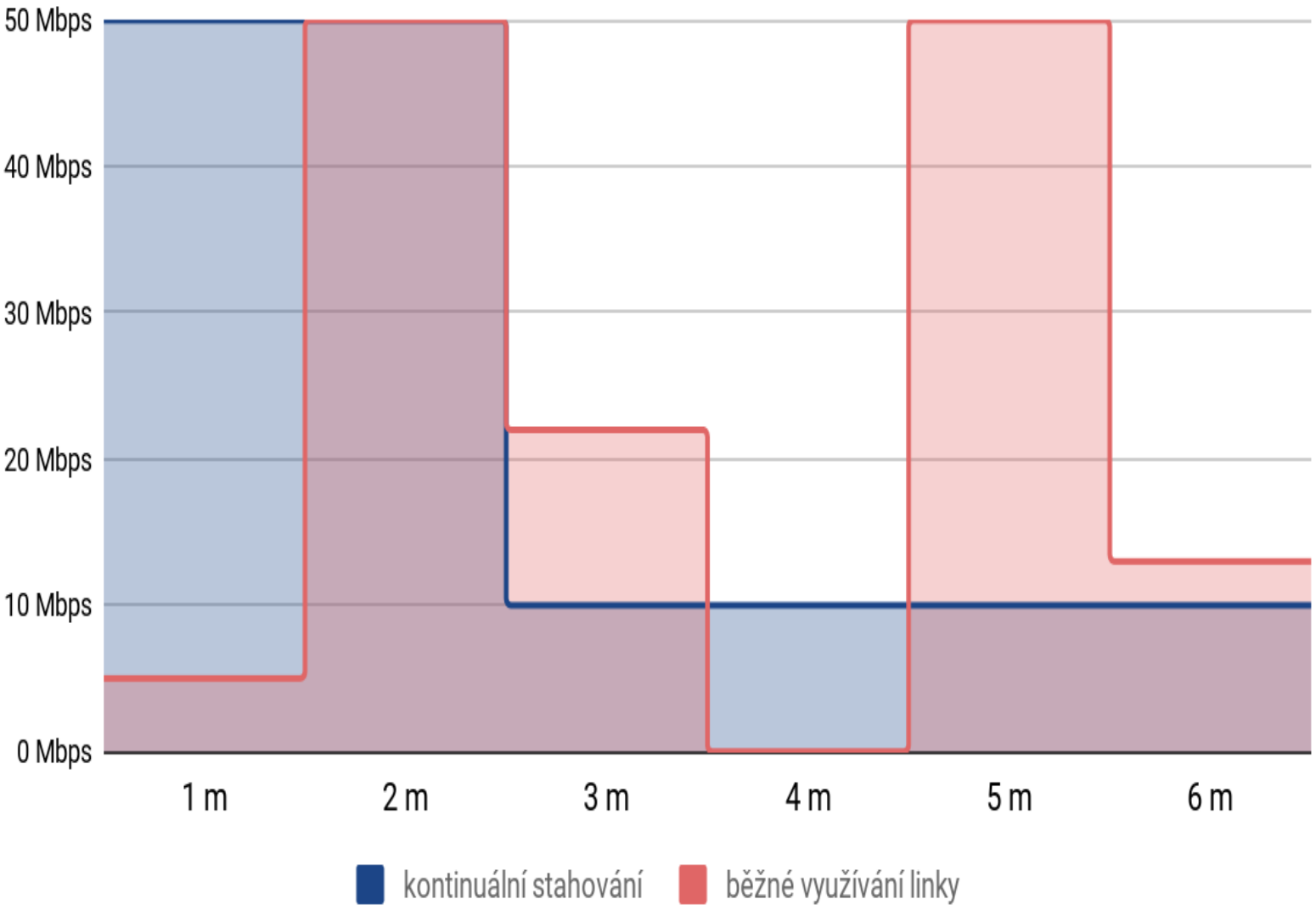
- Linka je omezena na 50 Mbps, tj. pokud požadavky obou zákazníků na šířku pásma překročí v součtu tuto mez, bude se reálná rychlost u obou poměrně snižovat.
- Chybou je že zákazníkovi, který kontinuálně stahuje, tento systém prakticky vůbec neuškodí, zatím co zákazník, který jen občas potřebuje maximální rychlost, ji nikdy nedostane.
- Pokud je maximální kapacita linky omezena shapingem, tak reálná situace odpovídá grafu. Pokud je ale omezena technologií (propustnost WiFi, ethernetu …), pak dojde k “přirozenému” shapingu, kdy bude situace často mnohem horší v neprospěch druhého zákazníka (přirozený shaping neořezává jednotlivé zákazníky, ale jen celou linku, tj. zákazník, který vysílá data “intenzivněji”, dostane podstatně větší šířku pásma, než zákazník s pár tcp konexemi).



- Pokud by nebyla využita priorita, tak oběma dvěma zákazníkům se rozdělí kapacita linky ve stejném poměru, tj. V tomto případě by od třetí minuty měli oba dva zákazníci alokováno 25Mbps.
- Protože ale má druhý zákazník nastavenu dvojnásobnou prioritu sdílení než první zákazník, tak v případě překročení kapacity linky se tato bude dělit v poměru 2:1 ve prospěch zákazníka s vyšší prioritou.
- Velmi zjednodušeně lze říci, že pokud oba zákazníci mají stejnou maximální rychlost a součet těchto rychlostí je vyšší než kapacita linky, tak v případě rozdílných priorit si linku rozdělí v poměru těchto priorit. Například pokud má první zákazník prioritu 2 a druhý prioritu 5, rozdělí se linka na sedm “dílů” (2 + 5) a první zákazník dostane 2 díly a druhý 5 dílů.
- Priorita zde tedy neznamená absolutní prioritu (jako například ve frontě), ale poměr sdílení. Tj. i zákazník s nízkou prioritou vždy nějakou linku dostane.

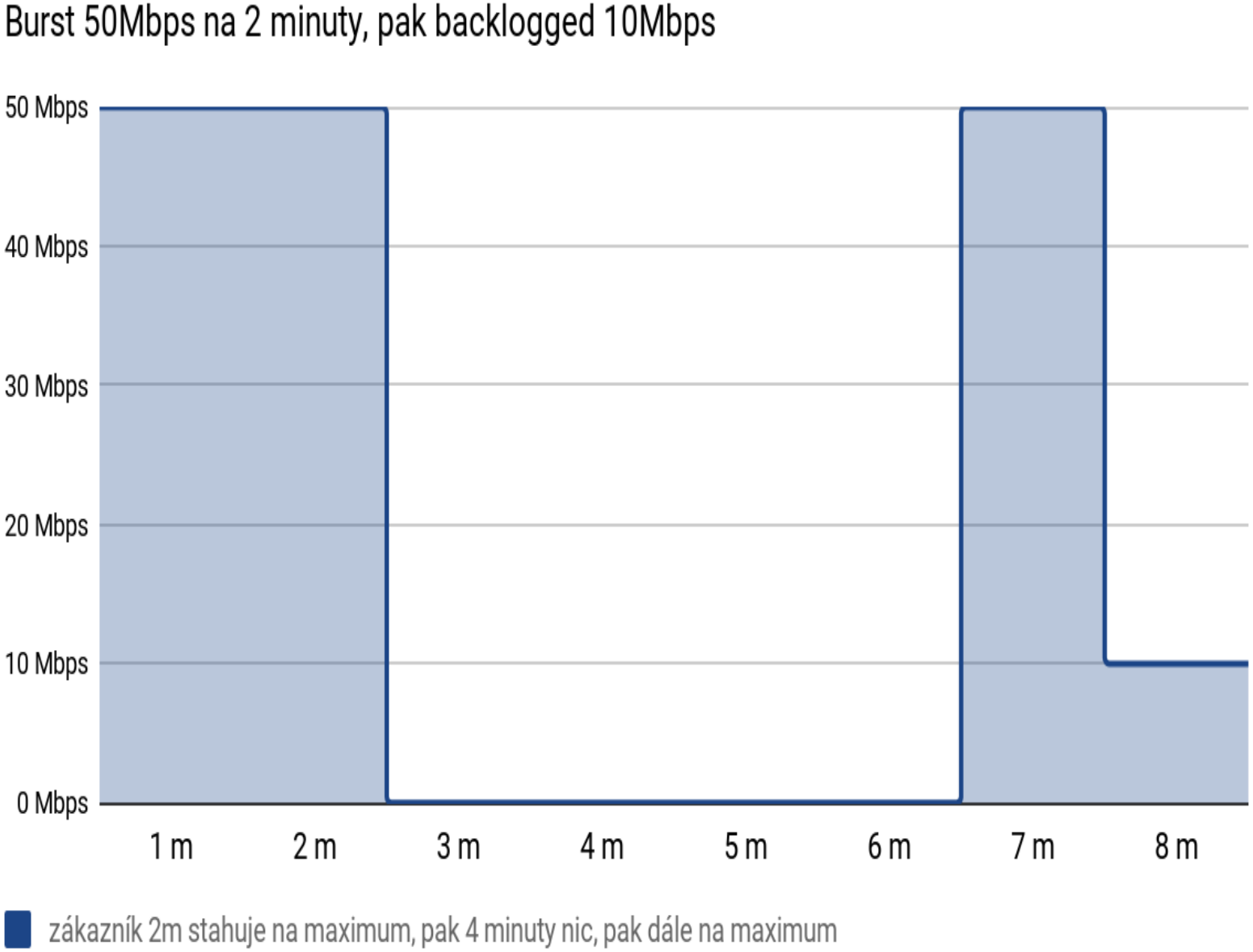
Shaping s burstem bez priorit s dostatečnou kapacitou linky

Burst 50Mbps na 2 minuty, pak backlogged 10Mbps



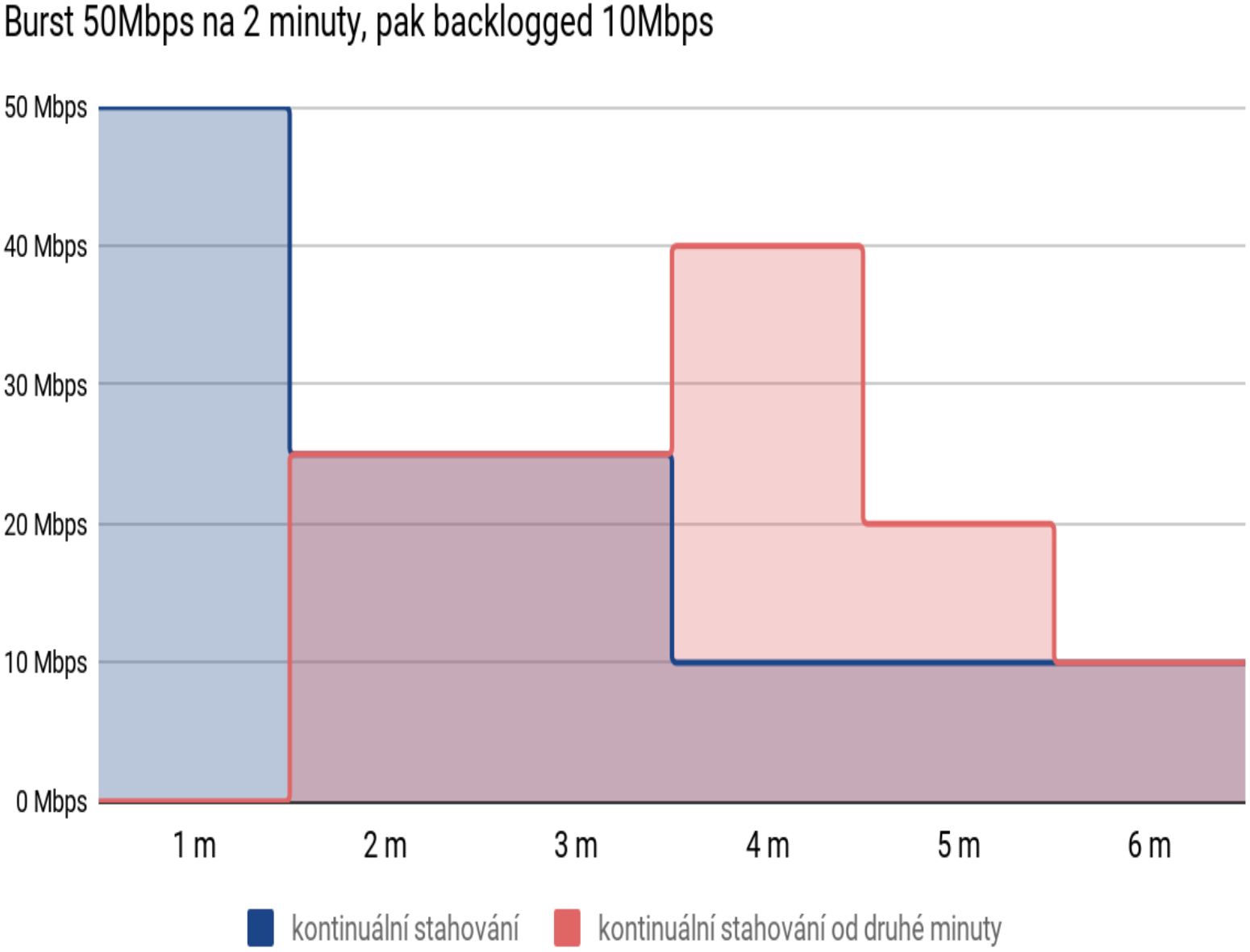
- Zákazník, který kontinuálně stahuje, bude po dvou minutách omezen na 10Mbps, na kterých pak zůstane do té doby, dokud na dostatečně dlouhý čas nepřestane využívat linku.
- Zákazník, který linku běžně používá, bude v případě potřeby dosahovat maximální rychlosti. Nesmí ji ale dosahovat příliš často za sebou, jinak se mu bude délka burstu poměrně krátit.
- Lze říci, že na burst si zákazník musí “našetřit” tím, že linku využívá méně, než je rychlost backlogged. V grafu je objem dat dán plochou pod křivkami a plocha burstu musí odpovídat nebo být menší než plocha předchozího “našetření”.
- Nepovinný úkol: V tomto grafu je numerická chyba. Kdo ji první najde a vysvětlí, má u mne pochvalu a pivo.

Příklad "našetření" na burst

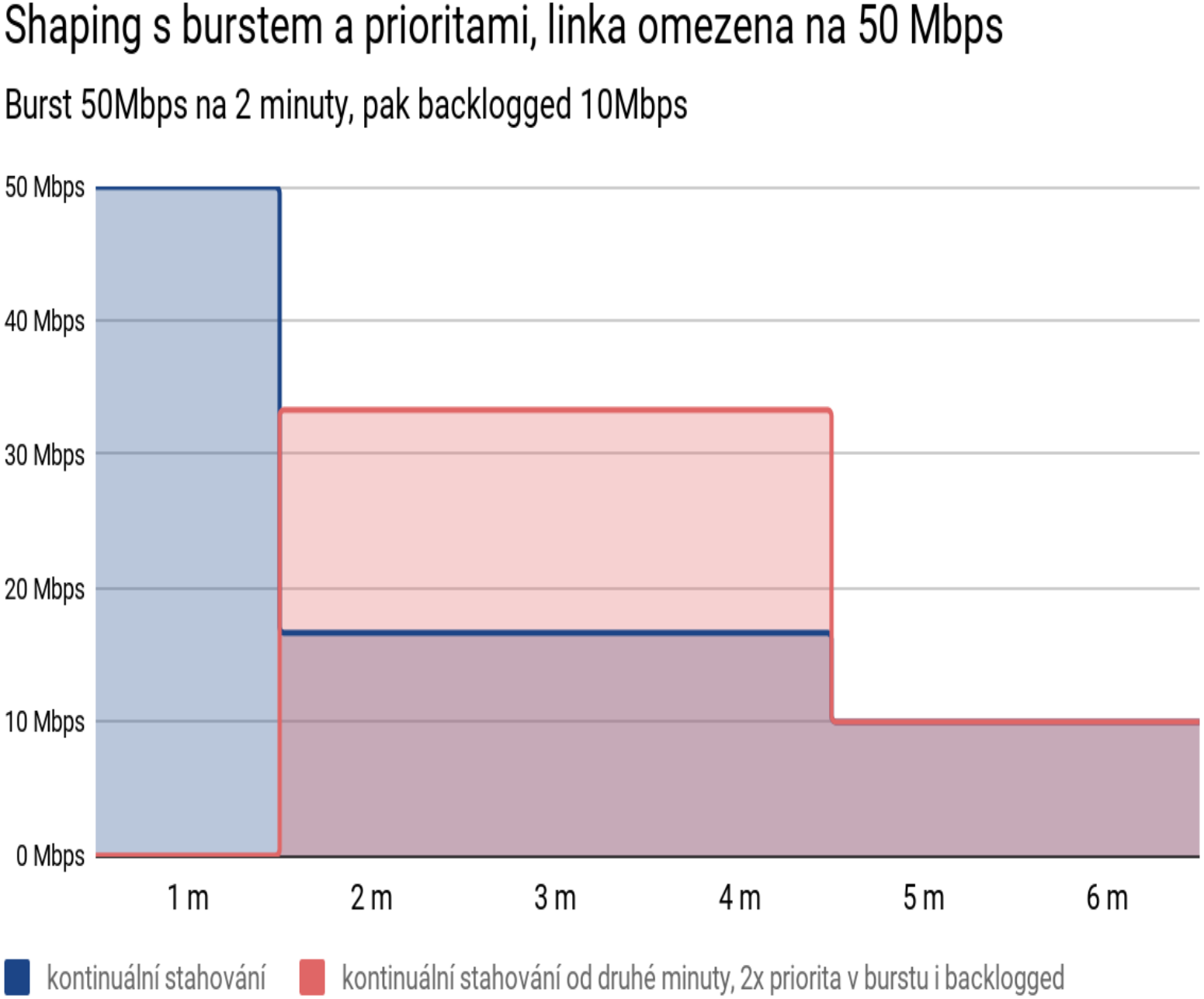


- Zákazník první dvě minuty vyčerpal burst, další čtyři minuty ale šetřil 10Mbps za minutu. Burst je ale o 40Mbps vyšší než backlogged, takže našetřil jen na jednu minutu burstu a pak mu linka spadne na backlogged.
- Otázka: Jak dlouho bude trvat druhý burst, pokud bude pauza ve stahování dlouhá šest minut?
- Otázka: Jak dlouho bude trvat druhý burst, pokud bude pauza ve stahování trvat deset minut?

Shaping s burstem bez priorit, linka omezena na 50 Mbps



- První zákazník začne stahovat burstem, protože je na lince sám. Na začátku druhé minuty chce začít stahovat burstem druhý zákazník, ale protože první má také burst a mají stejnou prioritu, linka se rozdělí v poměru 1:1. Tj. první oba zákazníci pojedou od druhé minuty 25 Mbps. Prvnímu zákazníkovi se tak prodlouží druhá polovina burstu na dvě minuty, protože ji čerpá poloviční rychlostí. Stejně tak druhý zákazník vyčerpá první polovinu burstu za dvě minuty, opět poloviční rychlostí. Na začátku čtvrté minuty bude mít první zákazník vyčerpaný burst a spadne na backlogged 10 Mbps. Druhý zákazník musí “dobrat” druhou polovinu burstu, což se mu povede rychlostí 40 Mbps za minutu a čtvrt (proto je u něj v grafu páté minuty 20 Mbps jako průměr), a pak spadne na 10 Mbps.



- První zákazník začne stahovat burstem, protože je na lince sám. Na začátku druhé minuty chce začít stahovat burstem druhý zákazník, ale protože první má také burst, linka se rozdělí v poměru jejich priorit 2:1. Tj. první zákazník dostane 16,66 Mbps a druhý 33,33 Mbps. Prvnímu zákazníkovi se tak prodlouží druhá polovina burstu na tři minuty, protože ji čerpá třetinovou rychlostí, a druhému zákazníkovi se prodlouží celý burst ze dvou na tři minuty, protože ho čerpá dvou třetinovou rychlostí. Oba tedy na začátku páté minuty budou mít vyčerpaný burst a spadnou do backlogged.
- Samozřejmě i když má druhý zákazník dvojnásobnou prioritu v backlogged, jedou od páté minuty v backlogged oba 10 Mbps, protože celkový součet jejich rychlostí není vyšší než propustnost linky.

Děkuji za pozornost :)

Podívejte se na ISP Gateway na www.u-n.cz
případně večer pivo