cesnet
metacentrum

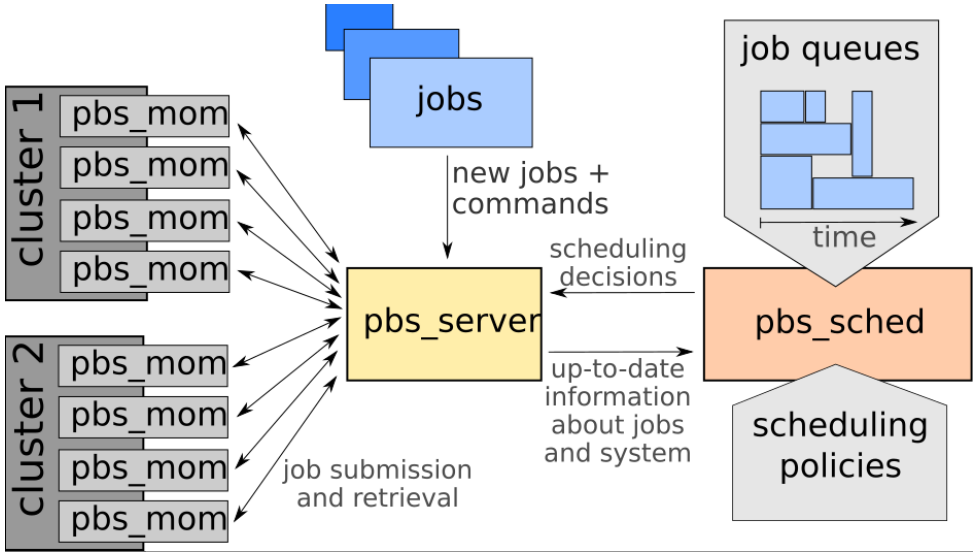
PBS PROFESSIONAL
ZKUŠENOSTI A TIPY

Dalibor Klusáček
Václav Chlumský
CESNET

květen 2018
Praha

Co je vlastně PBS Professional

- **Systém pro správu a přidělování zdrojů úlohám**
  - Řídí „životní cyklus“ úlohy
    - Submit do fronty, výpočet, dokončení, „úklid“…
  - Řídí souběžný běh úloh na sdíleném HW podle daných kritérií
  - Poskytuje rozhraní pro uživatele a administrátory
    - Je doplněn dalšími službami (webový portál, síťové disky, statistiky, …)



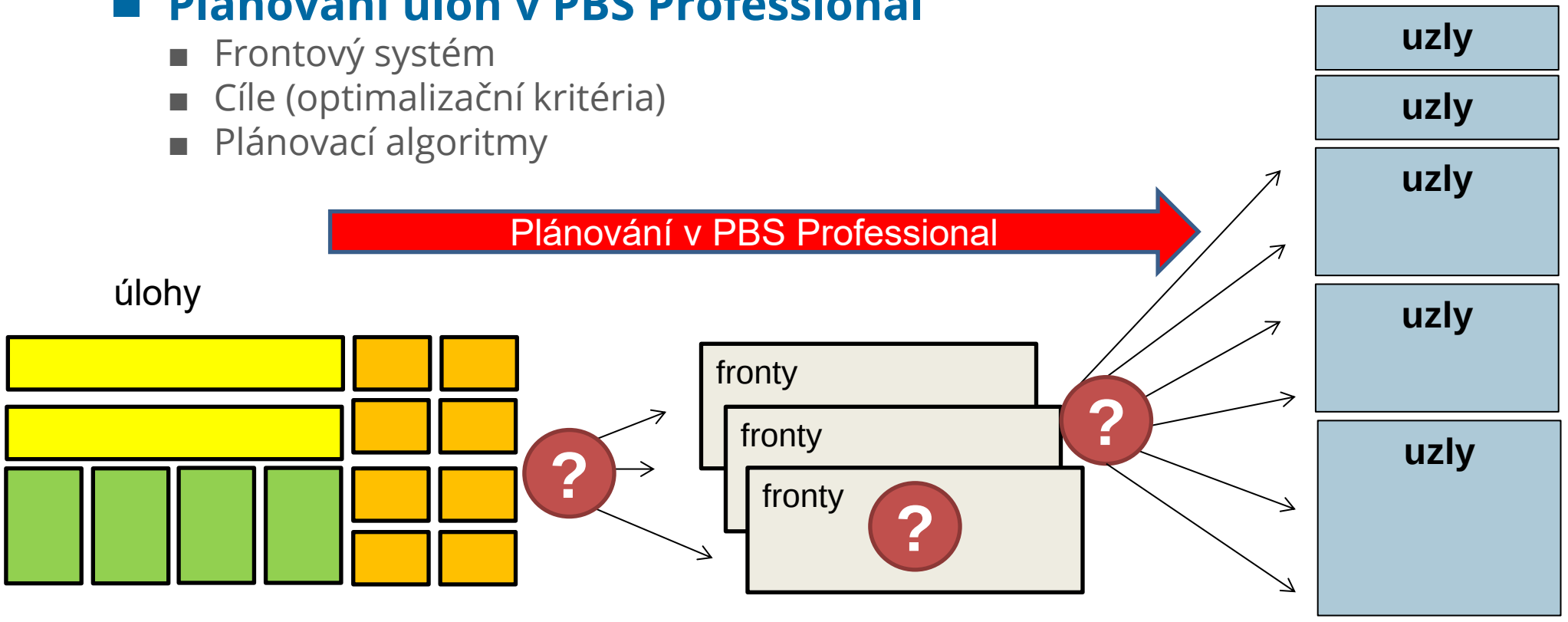
Osnova

- Jak používat systém PBS Professional bylo popsáno v přednášce na **Semináři gridového počítání 2017**
  - `https://metavo.metacentrum.cz/cs/seminars/seminar2017`
- Principy plánování úloh v MetaCentru
- Tipy pro efektivní plánování úloh

Principy plánování úloh v MetaCentru

Plánování úloh v PBS Professional

- Frontový systém
- Cíle (optimalizační kritéria)
- Plánovací algoritmy



Cíle plánování

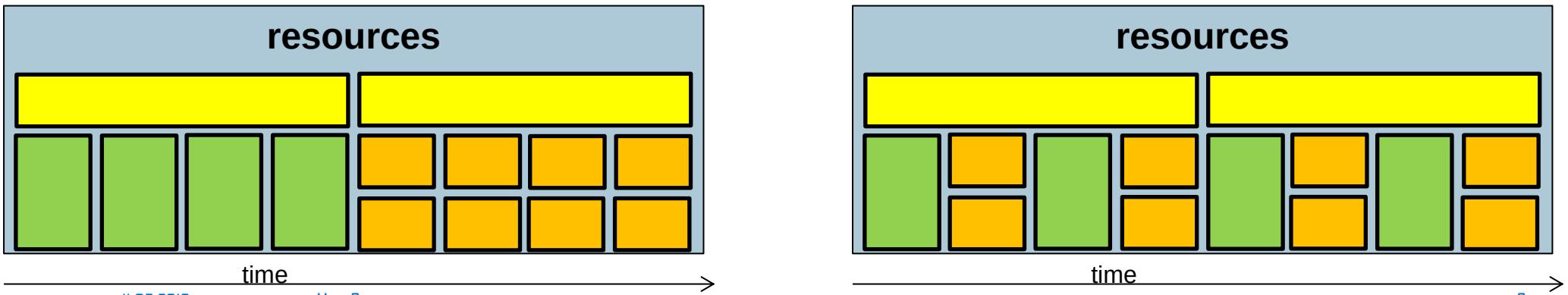
- Férovost
- Důraz na rozumný „mix“ běžících úloh
- Maximalizace vytížení a snižování doby čekání

Cíle: férovost

- **Férovost**
  - Dominantní politika, dynamické priority pro uživatele
  - Vyšší priorita pro uživatele s dosud nízkou spotřebou a/nebo vysokým počtem poděkování MetaCentru

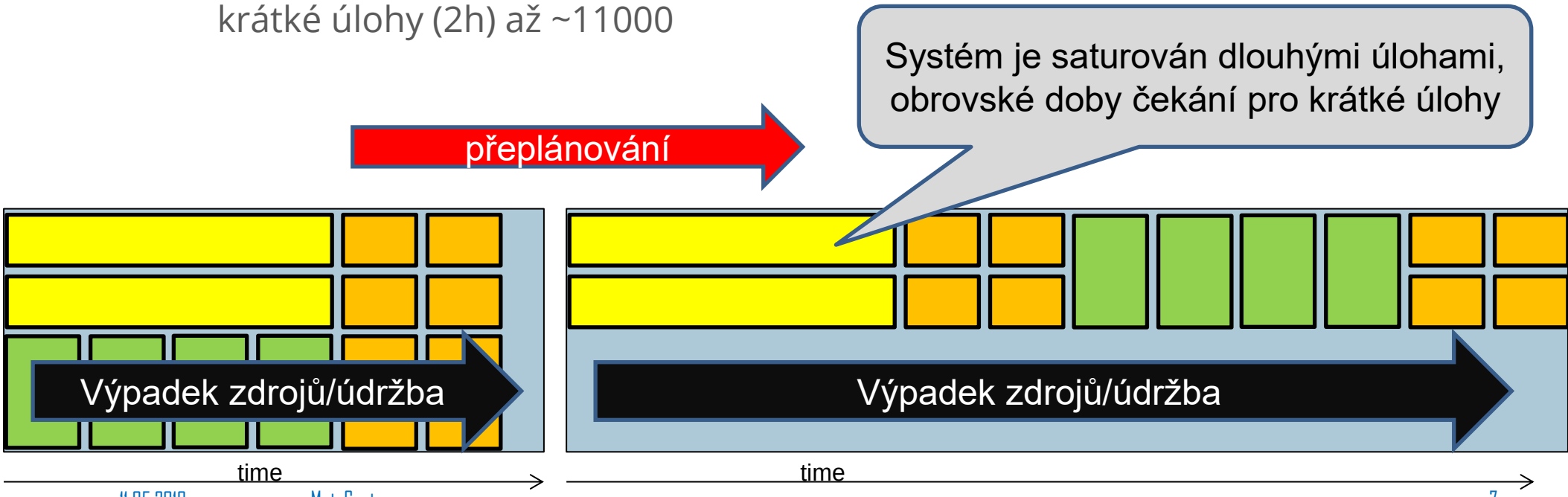
Neférový rozvrh

Fair-sharing



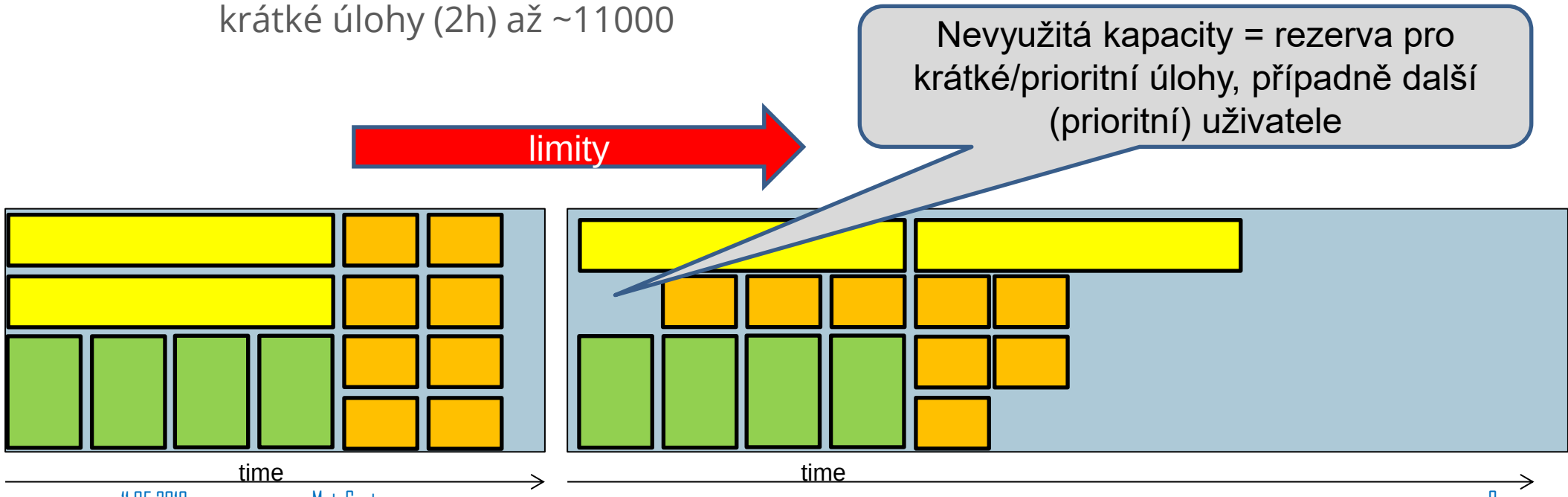
Cíle: skladba úloh

- **Důraz na rozumný „mix“ běžících úloh**
  - Omezení na počet běžících úloh dané třídy
  - Např. měsíční úlohy (2w plus) mohou zabrat max. ~6000 CPU zatímco krátké úlohy (2h) až ~11000



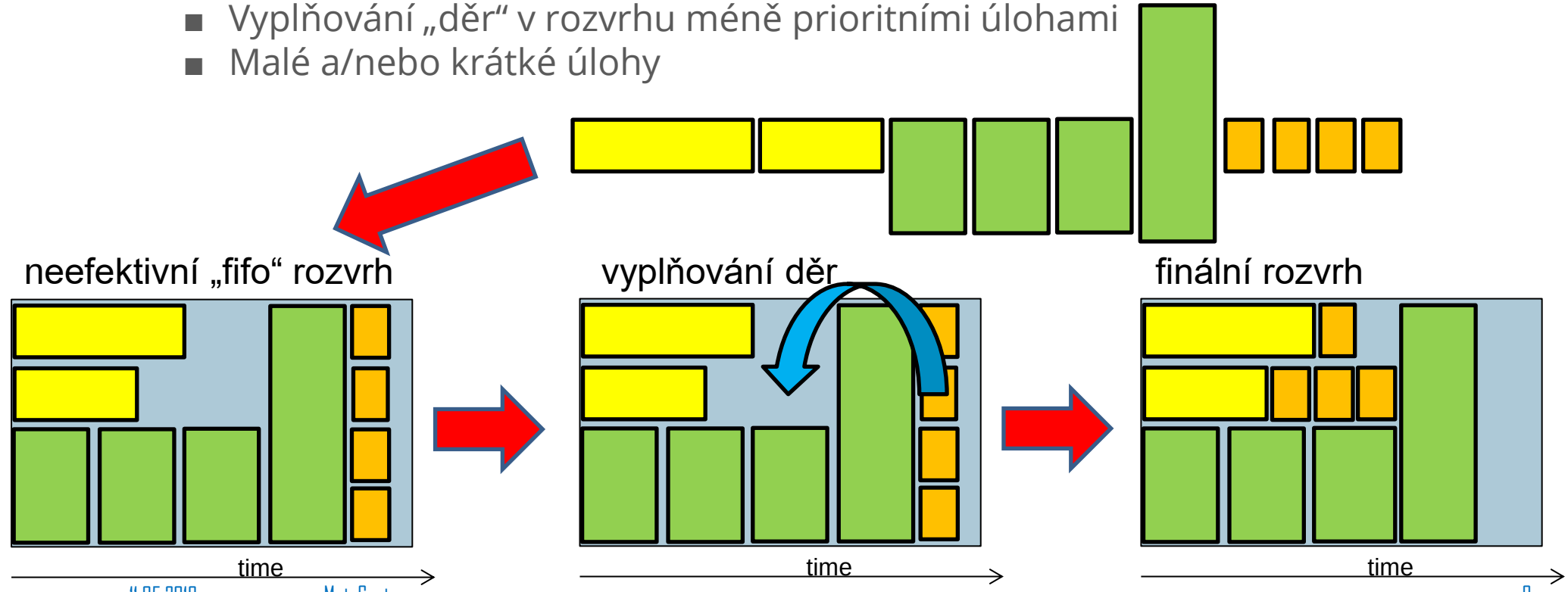
Cíle: skladba úloh

- **Důraz na rozumný „mix“ běžících úloh**
  - Omezení na počet běžících úloh dané třídy
  - Např. měsíční úlohy (2w plus) mohou zabrat max. ~6000 CPU zatímco krátké úlohy (2h) až ~11000



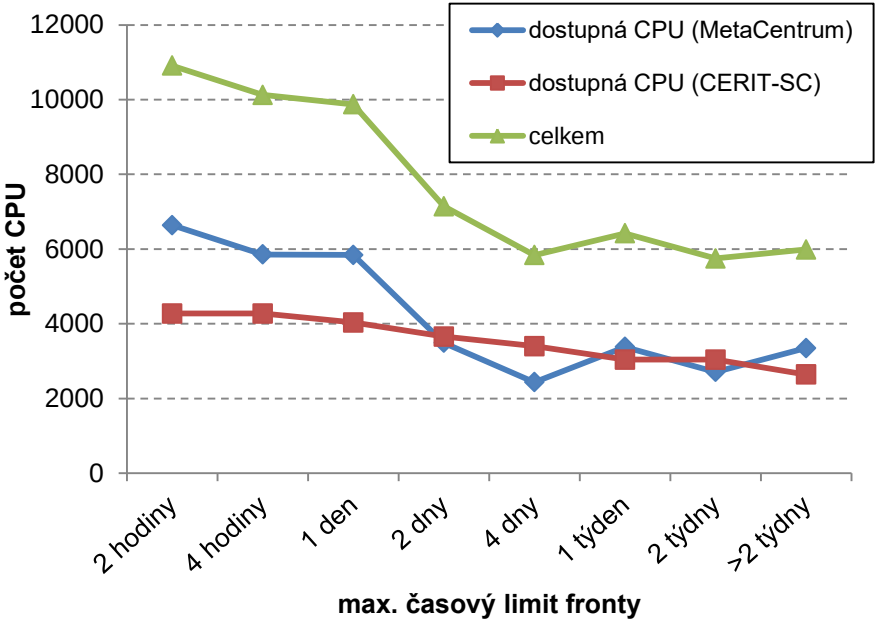
Cíle a algoritmy

- **Maximalizace vytížení a snižování doby čekání**
  - Vyplňování „děr“ v rozvrhu méně prioritními úlohami
  - Malé a/nebo krátké úlohy



Důležité poznámky

- **Uživateli specifikovaný walltime je rozhodující pro zařazení do fronty**
  - Fronta ovlivňuje počet dostupných CPU
  - **Překročení walltime => zabití úlohy**
  - Pokud nejsme dostatečně brzo varováni
- **Férovost je nadřazena frontám**
  - Neexistuje „férovější“ nebo „lepší fronta“
  - Výjimkou jsou prioritní fronty vlastníků clusterů



cesnet
metacentrum

TIPY PRO LEPŠÍ PLÁNOVÁNÍ ÚLOH

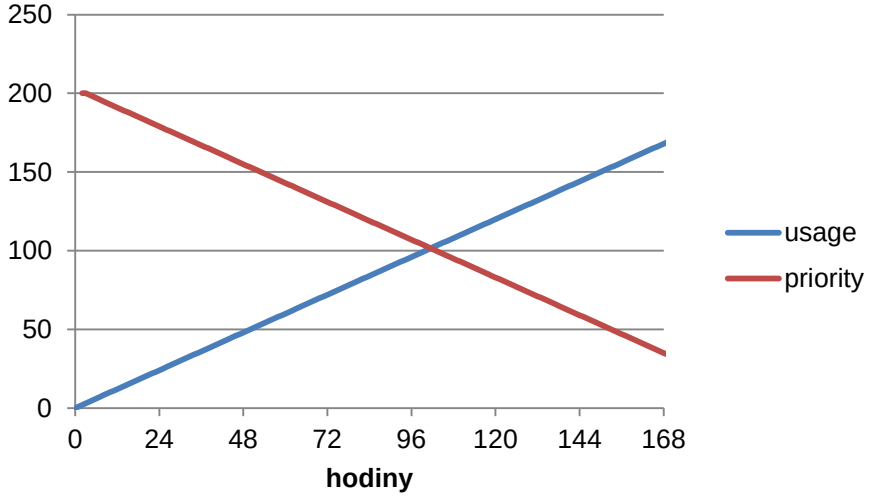
Anebo jak pomoci sobě i plánovači...

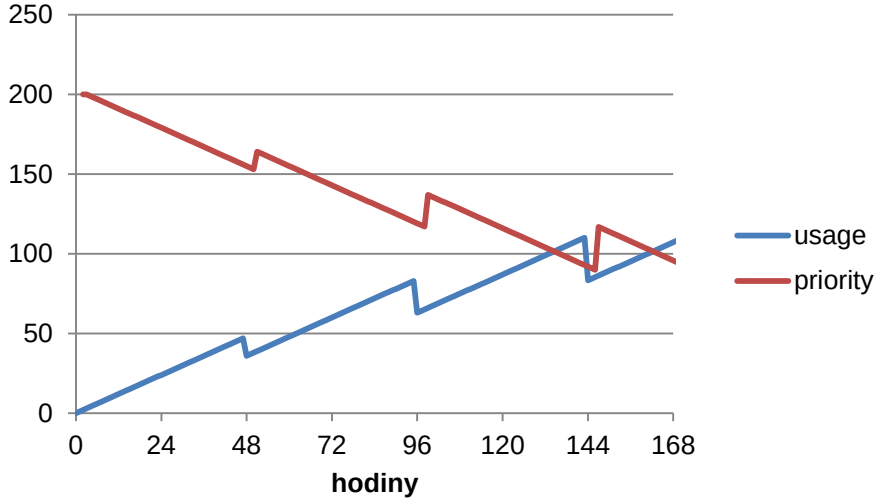
Tipy pro efektivní plánování úloh

- **Brzké spuštění Vašich úloh závisí zejména na 2 kritériích**
  - Jak vysokou máte prioritu (fair-share)
  - Jak „snadno“ lze úlohu pustit
- **Jak tyto dvě věci ovlivnit?**

Triky s fair-share

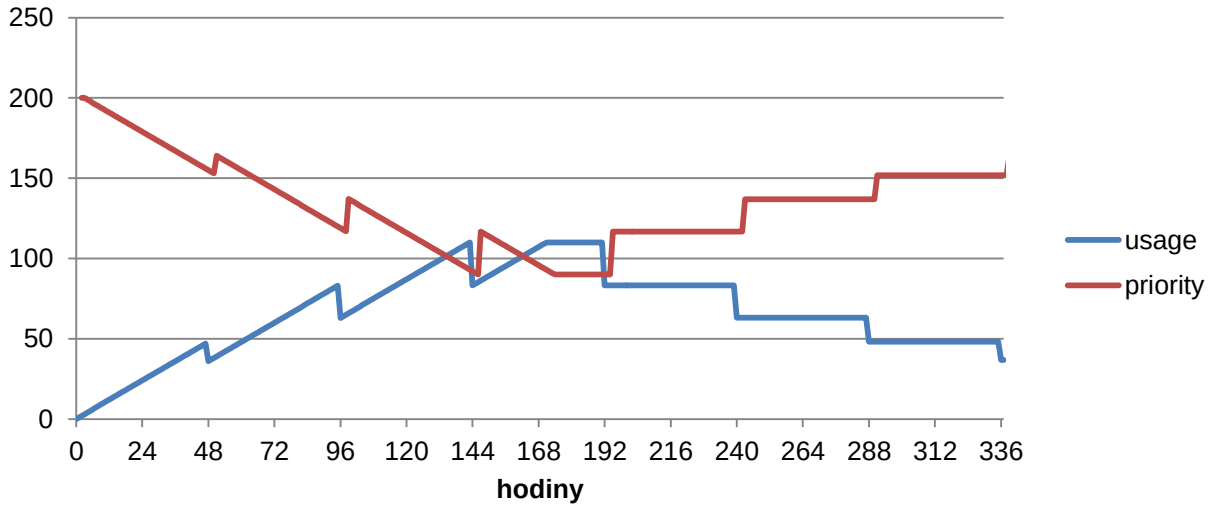
- **Vaše priorita klesá s množstvím výpočtů**
- **Zároveň se ale s časem postupně obnovuje**
  - Algoritmus „stárnutí“
  - Každých 48 hodin se dosavadní „spotřeba“ přenásobí 0,75





Triky s fair-share

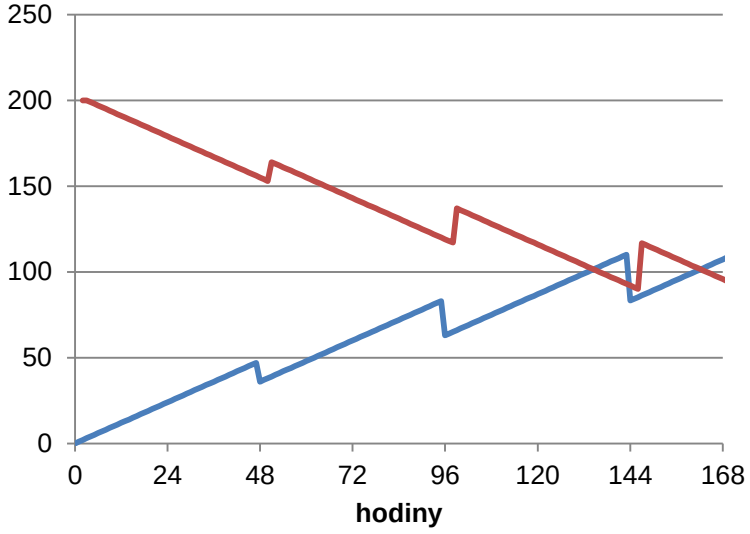
- **Vaše priorita klesá s množstvím výpočtů**
- **Zároveň se ale s časem postupně obnovuje**
  - Algoritmus „stárnutí“
  - Každých 48 hodin se dosavadní „spotřeba“ přenásobí 0,75
  - Po čase se pak opět „vynuluje“

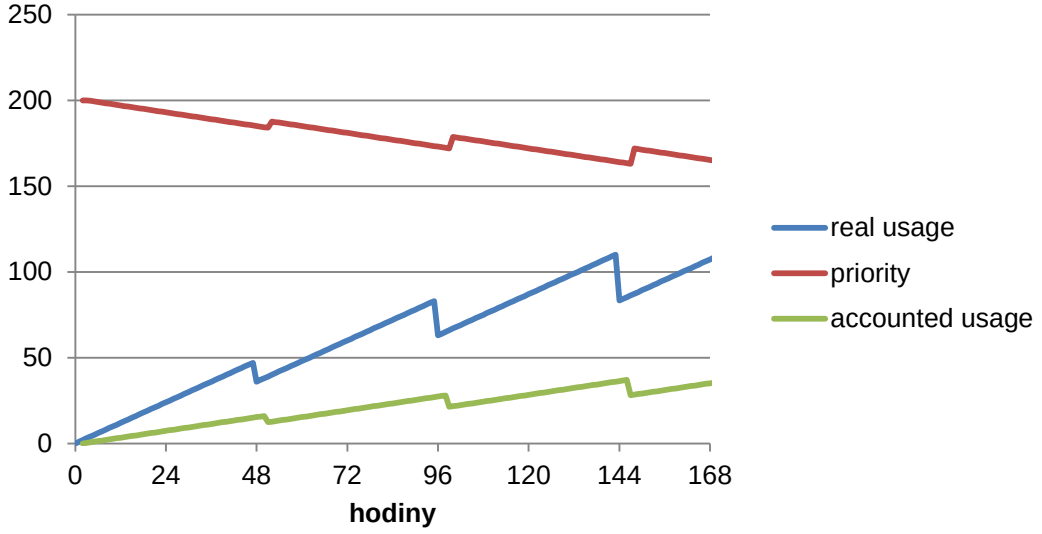


Triky s fair-share

Vaše priorita klesá s množstvím výpočtů

- Poděkováním MetaCentru v (kvalitní) publikaci se Vám navyšuje tzv. „share“
- Projeví se to tak, že Vaše reálná spotřeba se započítává jako zlomek
- Čím více (aktuálních) publikací máte, tím větší bonus





Triky s fair-share

- **Vaše priorita klesá s množstvím alokovaných zdrojů**
  - Uživatelé často plýtvají alokacemi
  - Reálné využití je malé
  - **Ovšem fair-share počítá s alokacemi!**
- **Zlepšení odhadu nutných zdrojů**
  - Zmenšení usage
  - Zlepšení fair-share

Úloha využila méně než 3/4 vyhrazených CPU!

využití přidělených prostředků		
RAM	9%	7gb / 80gb
CPU	68%	1081:35:31 / (16 * 99:07:27)
walltime	29%	99:07:27 / 337:00:00

Základní informace

úloha	CPU	vyhraz. paměť	použitá paměť
[redacted]c.cz	16	80gb	7gb

Jak usnadnit spuštění úlohy?

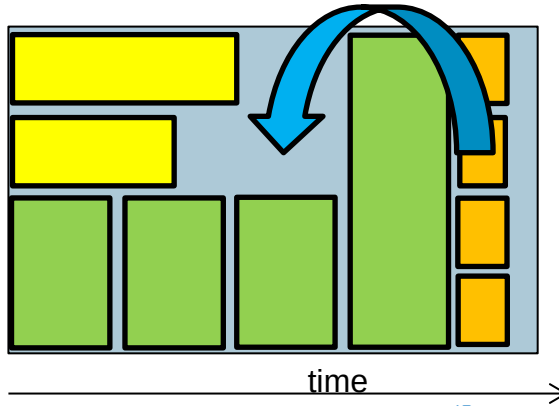
- **„Šířka“ úlohy a šířka chunků**
  - Čím „užší“ úloha, respektive chunk úlohy tím snáze se hledá stroj

- **Pro připomenutí: chunk je množina zdrojů přidělená „jako nedělitelná jednotka“ úloze**
  - Všechny zdroje daného chunku budou přiděleny na jednom uzlu (nedělitelnost)
  - Úloha může žádat o více (různých) chunků
  - Příkaz: `-l select=[chunk_1][+chunk_2]...[+chunk_n]`
  - Příklad úlohy s 2 množinami různých chunků:

```
qsub -l select=6:ncpus=2:mem=4gb+3:ncpus=8:mem=4gb ...
```

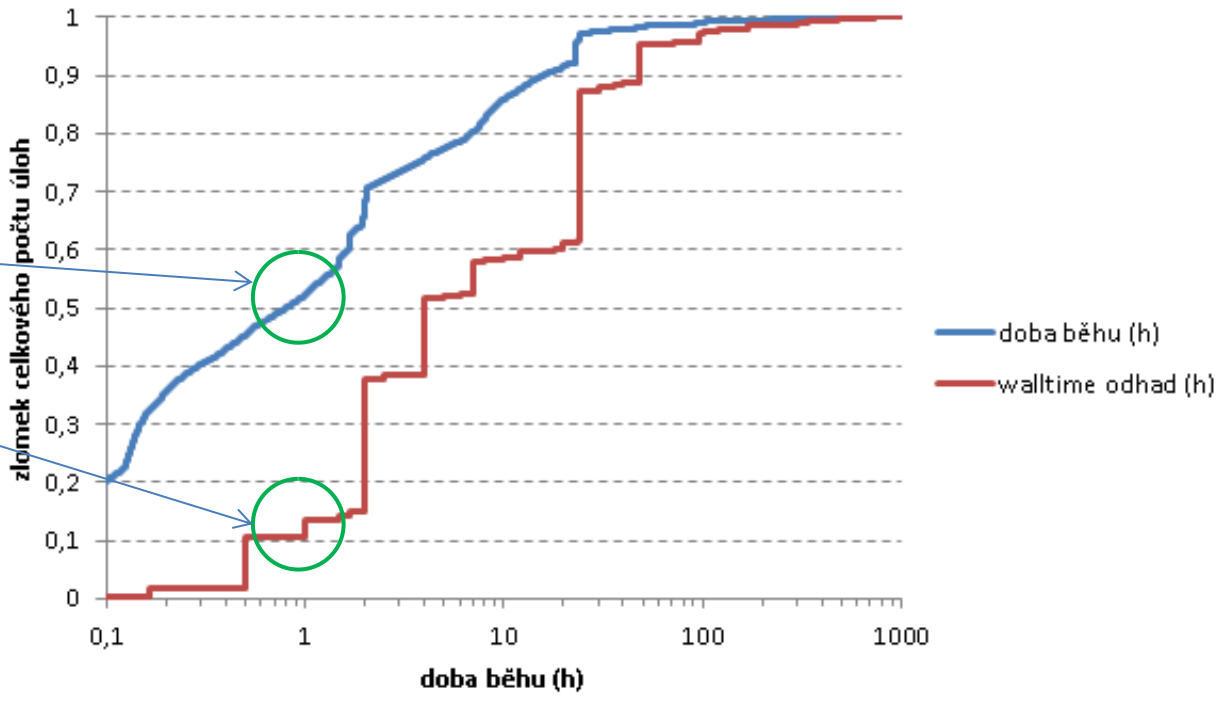
Jak usnadnit spuštění úlohy?

- **Usnadňovat alokaci chunků na uzly**
- **`-l place=[arrangement][:sharing]…`**
  - „aranžovací“ volba **`place=free`** dává větší volnost v obsazování uzlů
  - „sdílecí“ volba **`place=excl/exclhost`** zmenšuje množinu vhodných uzlů
  - „drobení chunků“, menší chunky snáze naleznou volné zdroje
- **Délka úlohy (parametr walltime)**
  - Kratší úloha má k dispozici typicky větší množství uzlů
  - Kratší úloha pak snáze „předbíhá“
  - Vleze se do díry v rozvrhu
  - „Výplňová“ úloha předbíhá „větší“ úlohy s vyšší prioritou



Současný stav (ne)přesnosti walltime

- **Walltime specifikován uživateli**
- **Odhady jsou často velmi nepřesné**
  - Např. 53% úloh skončí během 1 hodiny
  - Pouze 13% úloh požaduje max. 1 hodinu



Diagnostika problémů

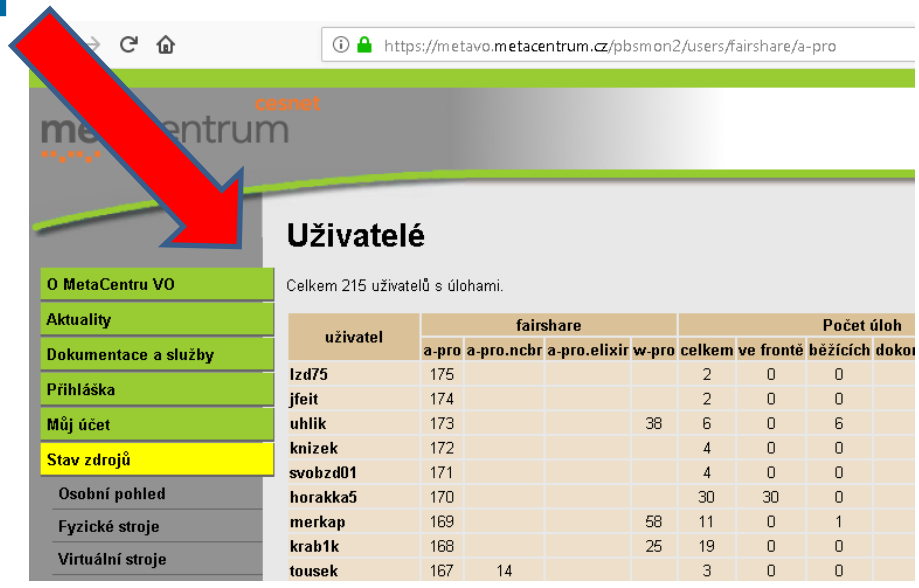
- Výpis informací úloh v PBSMon: `https://metavo.metacentrum.cz/`
- Případně pomocí příkazu: `qstat -f <jobid>`
- Věnovat pozornost **komentáři u čekající úlohy**
  - „**Can Never Run**“ – úloha nikdy nepoběží (chyba v zadání, oprava pomocí `qalter`)
  - „**Not Running: Insufficient amount of resource: XY**“ – úloha hlásí, že chybí konkrétní zdroj. PBS ale hlásí pouze první chybějící, může jich být víc…
  - Pokud má úloha nastaven `estimated.start_time` a `estimated.exec_vnode`, je to „důkaz“, že je správně zadána

požadované stroje	1:ncpus=2:infiniband=brno:scratch_local=10gb:mpiprocs=2
vytvořena	Pondělí, 9. října 2017 12:31:53
způsobilá k běhu	Pondělí, 9. října 2017 12:31:53
poslední změna stavu	Úterý, 31. října 2017 8:16:42
komentář	Can Never Run: Insufficient amount of resource: infiniband (brno != luna,minos,manegrot,alfrid,tarkil,le

Další tipy

Ověřte si svoji fair-share prioritu

- Priorita není absolutní!
- Priorita je dynamická
- Ověřte si, že máte zadané publikace



uživatel	fairshare				Počet úloh			
	a-pro	a-pro.ncbr	a-pro.elixir	w-pro	celkem	ve frontě	běžících	dokor
lzd75	175				2	0	0	
jfeit	174				2	0	0	
uhlik	173			38	6	0	6	
knizek	172				4	0	0	
svobzd01	171				4	0	0	
horakka5	170				30	30	0	
merkap	169			58	11	0	1	
krab1k	168			25	19	0	0	
tousek	167	14			3	0	0	

Chcete počítat náročnou úlohu?

- Pak se vyvarujte souběžného spouštění „snadných“ úloh
  - Předběhnou náročnou úlohu
  - Tím Vám zhorší fair-share
  - A odsunou náročnou úlohu z čela fronty

Další tipy

- **Řazení vlastních úloh uživatele**
  - Úlohy se řadí podle priority, kterou si **uživatel může sám definovat**
  - Při submitu úlohy: `qsub -p <priorita>`
  - Pro svoji již submitnutou úlohu: `qalter <jobid> -p <priorita>`
  - Kde `<priorita>` může být v intervalu [-1024, +1023] a výchozí hodnota je 0
- **Následně je úloha řazena podle „eligible time“**
  - Doba jak dlouho je úloha dostupná ke spuštění
  - Typicky doba od submitnutí úlohy
- **Dopředné rezervace uzlů (pouze z dobrých důvodů)**
  - Vytváření rezervací podléha ACL a není standardně přístupné
  - Rezervace je spojena se speciální frontou
  - Tu použije uživatel pro submitování úloh

cesnet
metacentrum

DĚKUJI ZA POZORNOST

klusacek@cesnet.cz