Novinky NGI & 10 způsobů, jak "sejmout" MetaCentrum

(aneb jak zacházet s výpočty a daty)

Tomáš Rebok

MetaCentrum, CESNET z.s.p.o.

CERIT-SC, Masarykova univerzita

(rebok@ics.muni.cz)

MetaCentrum NGI

Přístupná zaměstnancům a studentům VŠ/univerzit, AV ČR, výzkumným ústavům, atp.

- komerční subjekty pouze pro veřejný výzkum

nabízí:

- výpočetní zdroje
- úložné kapacity
- aplikační programy

http://metavo.metacentrum.cz

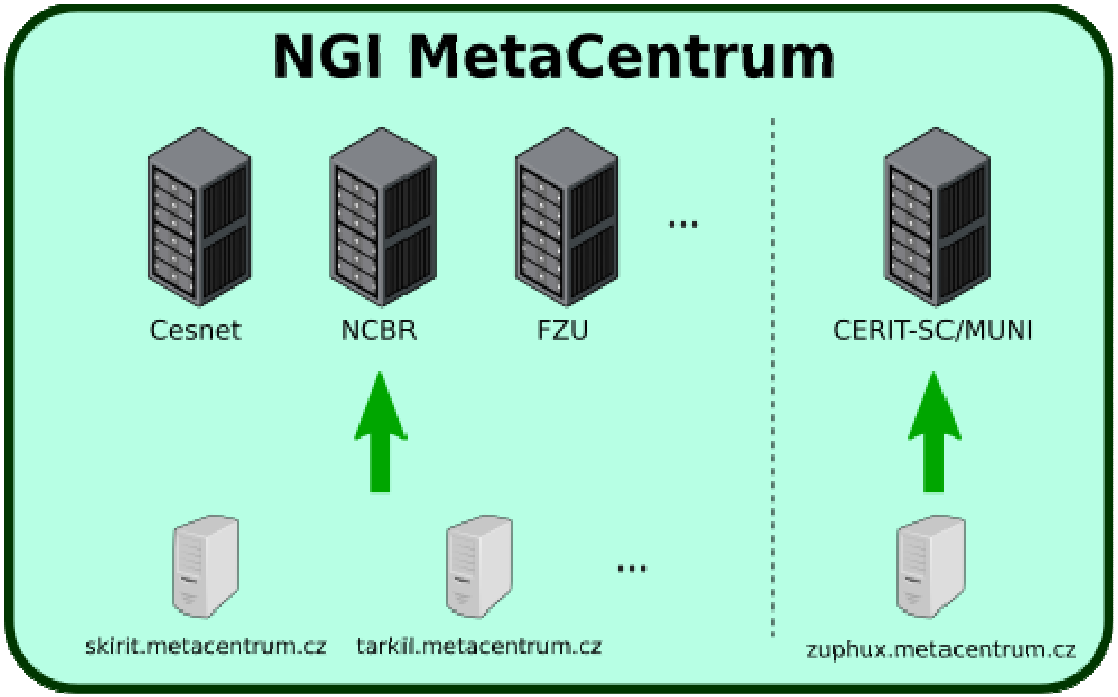
Po registraci k dispozici zcela zdarma

- „placení“ formou publikací s poděkováním

MetaCentrum & CERIT-SC

Jaký je vzájemný vztah MetaCentra a CERIT-SC?

- MetaCentrum disponuje vlastními zdroji (CESNET) a integruje zdroje externích poskytovatelů
  - CERIT-SC/MUNI je jedním z nich
  - dále CEITEC/NCBR, FZU, ČVUT, JČU, ZČU, UPOL, MU, …



+ sdílená úložiště

a sdílená SW výbava

Novinky NGI

obecná snaha infrastrukturu z pohledu uživatelů zbytečně neměnit
(pokud k tomu nemáme pádné důvody)

Hardware a software

Nový hardware

- *nové výpočetní clustery* – alfrid, zubat (GPU), krux, zefron (GPU), exmag, meduseld, tarkil, upol128 (UV2000), phi (Xeon Phi), …
- *nová úložiště*

Nový (aktualizovaný) komerční software

- *kompilátory* (Intel, PGI), *ladící nástroje* (TotalView, Allinea DDT)
- *matematický SW* (Matlab, Mathematica, gridMathematica)
- *aplikační chemie* (Gaussian + Gaussian Linda)
- *materiálové simulace* (Ansys CFD + Mechanical, 512 jader Ansys HPC)
- *bioinformatika* (CLC Genomics Workbench)

Průběžné aktualizace softwarových balíků

- cca 350 různých aplikací
  - viz http://meta.cesnet.cz/wiki/Kategorie:Aplikace
- aktualizace na žádost
- instalace nových aplikací na žádost či ve spolupráci s uživateli
  - pomůžeme Vám s kompilací

Plánovací systém PBS Professional

Nový plánovač (PBS Pro)

- náhrada doposud využívaného plánovače Torque
- podrobnější informace, možnosti a způsoby použití – viz samostatná přednáška (D. Klusáček)
- postupný přechod infrastruktury na nový plánovač
  - současné využívaní obou plánovačů
    dočasné nepohodlí při migraci ☹

MetaCentrum & PBS Pro

- přechod (téměř) dokončen
  - čelní a výpočetní uzly přemigrovány
  - zbývají izolované součásti (NCBR)
- nespočítané úlohy přemigrovány do Torque plánovače CERIT-SC
  - nepřevoditelné úlohy komunikovány s uživateli

Plánovací systém PBS Proffesional

CERIT-SC & PBS Pro

- přemigrováno několik clusterů
  - **PBS Pro**: phi, ungu, urga, zebra (jen část)
  - ostatní stále v Torque
- všechny zdroje dostupné přes čelní uzel
  zuphux.cerit-sc.cz
  - *standardní plánovač:* Torque
  - *volitelný plánovač:* PBS Pro
    dostupný po přidání modulu „pbspro-client“
    - (`module add pbspro-client`)
    - při výhradním využívání lze umístit do inicializačních skriptů Bash (`bash_profile`)

cílový stav: plný přechod na PBS Pro

Akcelerátory Xeon Phi KNL

phi[1-6].cerit-sc.cz

- nový cluster pořízený infrastrukturou CERIT-SC
  - dostupný skrze frontu "`phi`" (PBS Pro) na čelním uzlu zuphux.cerit-sc.cz

    ```
    $ qsub -q phi -l select=…
    ```
- **nejnovější generace Xeon Phi** (7210 Knights Landing)
  - aktuálně jediná taková instalace v ČR
- bližší informace k výhodám a použití – viz samostatná přednáška (J. Filipovič)

<u>Specialita instalace:</u> centrální úložiště dostupná přes SCP

Centrální úložiště clusteru phi.cerit-sc.cz

Opuštění centrálního sdílení uložišť skrze NFS

tj. konceptu /storage/XXX/home/<username>

- technické důvody
- přístup k datům skrze SCP
  - ve většině případů pouze minimální změna ve skriptech

NFS sdílení (aktuální stav)	SCP sdílení (phi[1-6].cerit-sc.cz)
DATADIR="**/storage/brno3-cerit/home/<username>/example**" **cp** –R $DATADIR/mydata $SCRATCHDIR	DATADIR="**storage-brno3-cerit.metacentrum.cz:~/example**" **scp** –R $DATADIR/mydata $SCRATCHDIR

Množství datových úložišť přerůstá únosnou mez

- hledáme způsoby, jak práci s úložišti zpřehlednit

  např. jedno velké úložiště? („object storage“)
- předmět intenzivního rozvoje v letošním a příštích letech

Novinky v cloudovém prostředí

Nová verze systému OpenNebula

- nové uživatelské rozhraní

  přehlednější, větší možnosti konfigurace VM

- **podpora „*security groups*“**
  - nově standardně uzavřena většina síťových portů stroje
    větší zabezpečení VM
  - pro každou VM lze zvolit požadovaná kombinace „security group“
    otevírají požadované síťové porty
    - „default-permissive“ – (téměř) vše otevřeno (původní chování)
    - nové „security groups“ vytvářeny na žádost
- nová systémová funkcionalita

  podpora VXLAN (moderní varianta VLAN), atp.

Další dílčí novinky

Nová uživatelská dokumentace

- snaha o zpřehlednění, zaktualizování a vyčištění původní dokumentace
- viz https://wiki.metacentrum.cz/wiki
  - původní stále dostupná na https://wiki.metacentrum.cz/wikiold
- **zachováno právo editace vás uživatelů**

  uvítáme Vaši pomoc s rozšiřováním dokumentace
  - např. tipy a návody k aplikacím, ukázkové příklady, atp.

Systémové změny

- plný přechod na Debian 8
- mnoho „neviditelných“ změn

Hadoop v MetaCentru

Apache Hadoop

- open-source framework pro distribuované zpracování velkých objemů dat (BigData)
  - za použití programovacího paradigma MapReduce
    výpočetní funkce jsou posílány k datům
    - (standardně data putují k výpočetním funkcím)

Hadoop v MetaCentru

- doplněno obvyklými nadstavbami

  Pig, Hive, Hbase, YARN, …
- **není novinkou** (dostupnost od 2015)
  - zatím však stále relativně sporadické využití

Jak „sejmout“ MetaCentrum?

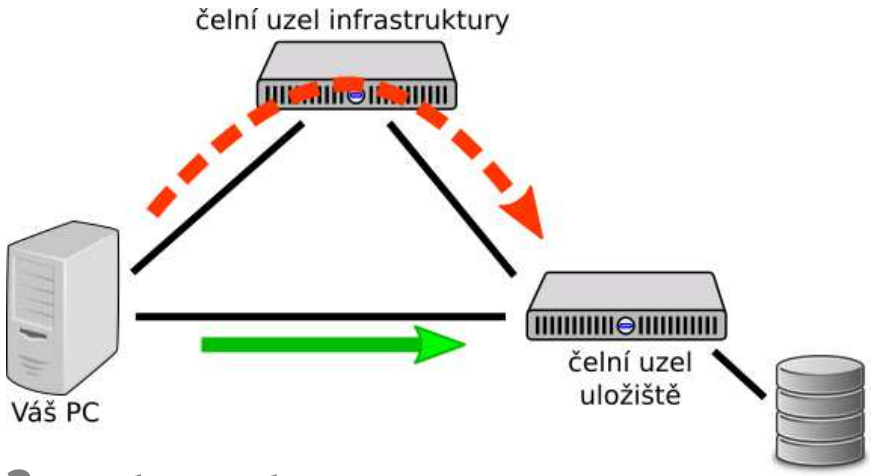
(aneb Jak správně zacházet s výpočty a daty)

Nebojte se infrastrukturu používat – pokud něco „sejmete“, je to naše chyba. ☺

Kopírování objemných dat

Nekopírujte objemnější data přes čelní uzly

- pomalejší kopírování
- zatížení čelních uzlů



Data lze kopírovat přímo skrze přístupové uzly úložišť

- SCP, WinSCP

  /storage/brno2 -> storage-brno2.metacentrum.cz

  /storage/brno3-cerit -> storage-brno3-cerit.metacentrum.cz

  …

  - https://wiki.metacentrum.cz/wiki/Working_with_data/Direct_access_to_data_storages

Výpočty nad centrálními úložišti

Nespouštějte výpočty nad daty v centrálním úložišti

- zejména s intenzivnějšími I/O operacemi
  - vede k ochromení úložiště a prodloužení doby běhu úlohy

Kopírujte pracovní data do scratche

- *pozitivní vlivy:*
  - zrychlení běhu úlohy
  - odstranění závislosti na dostupnosti centrálního úložiště
- postup:
  - $ qsub –l select=1:ncpus=4:**scratch_local=1gb** …
    cp /storage/…/home/<username>/mydata $SCRATCHDIR/mydata
    cd $SCRATCHDIR
    <compute>
    cp $SCRATCHDIR/results /storage/…/home/<username>/results
  - …:scratch_shared=Xgb … *sdílený scratch (distribuované úlohy)*
  - …:scratch_ssd=Xgb … *lokální scratch – SSD disky*

Data ve scratchích

Promazávejte data po ukončených úlohách

- pracovní data ve scratchích obdobou pracovních dat v paměti
  - po korektním ukončení úlohy by měla být odmazána
- scratche automatizovaně promazávány
  - avšak většinou až 2 týdny po ukončení úlohy

Promazávání scratchů v úlohách

- utilita „`clean_scratch`”
- postup:

```
trap ‘clean_scratch’ TERM EXIT
...
cp results /storage/...  || export CLEAN_SCRATCH=false
```

  - při nedostupnosti centrálního úložiště (selhání vykopírování výsledků) data ponechá ve scratchi
    - informuje o korektním promazání scratche či ponechání dat
    - informuje o nepromazaných scratchích z jiných úloh (na daném uzlu)

Nadužívání místa na úložištích

Centrální (pracovní) úložiště nejsou nekonečná ☹

/storage/<MĚSTO>

Promazávejte/odsunujte nepotřebná data

- *možnosti:*
  - odmazání nepotřebných dat
  - odsun aktuálně nepotřebných dat do archivních úložišť
    viz https://wiki.metacentrum.cz/wiki/Archival_data_handling

Velké výstupy úloh a zápisy do `/tmp`

Výpočetní uzly mají omezené kvóty (1GB) pro zápis na lokální disky (mimo scratche)

- vliv na zápisy aplikací do `/tmp` (dočasné pracovní soubory)
- vliv na objemné výstupy úloh (stdout, stderr)

Přesměrovávejte objemnější výstupy do scratche

- přesměrování dočasného úložiště pracovních souborů

  mnoho aplikací reflektuje systémovou proměnnou TMPDIR
  - nastavení: `export TMPDIR=$SCRATCHDIR`
- přesměrování standardního a chybového výstupu
  - `myapp … 1>$SCRATCHDIR/stdout 2>$SCRATCHDIR/stderr`
- zjištění stavu lokální uživatelské kvóty a zabírajících souborů (prvotní informace emailem)
  - utilita `$ check-local-quota`
    spuštěno na předmětném uzlu

Neefektivní výpočty

Zajímejte se o efektivitu Vašich úloh

- požadavek na více jader nepromění jednoprocesorový/ sériový výpočet na paralelní (= nedojde ke zrychlení)
  - využíváno bude stále jediné CPU
- mnoho aplikací mění počet využívaných jader v průběhu výpočtu
  - větší počet jader může být využíván jen po krátkou dobu běhu aplikace

Sledování využití (nejen) CPU úlohou:

- *v průběhu běhu úlohy:*
  - na výpočetním uzlu (SSH) s využitím standardních nástrojů (`top`, `htop`, …)
- *po ukončení úlohy:*
  - na portále v přehledu úloh
    červené podbarvení neefektivních úloh

Infiniband

Distribuované úlohy mohou být neefektivní kvůli pomalému komunikačnímu kanálu

- komunikace skrze standardní síťové propojení (Ethernet) je pomalá
- **Infiniband** – specializované nízkolatenční propojení pro podporu rychlé komunikace distribuovaných úloh

Mnohé clustery NGI jsou vybaveny Infinibandem

- výrazně urychluje běh distribuovaných (MPI) úloh
  - dostupnost detekována automaticky
    vždy identické spuštění: `mpirun myapp`
  - v případě nedostupnosti je využit Ethernet
- *požadavek:*
  - $ qsub –l select=… **–l place=group=infiniband** script.sh

Mnoho krátkých úloh

Seskupujte příliš krátké úlohy

- např. v délce do jednotek minut
  - režie spuštění tvoří netriviální podíl doby běhu
    neefektivní využití zdrojů

V jedné úloze spusťte více instancí výpočtu

- *možnosti realizace:*
  - sériové spuštění instancí v rámci běhu jedné úlohy
    process data1
    process data2
    ...
  - paralelní spuštění instancí v rámci běhu jedné úlohy (nezbytná alokace dostatku CPU)
    - `pbsdsh`
    - `parallel`

Výpočty na čelních uzlech

Nepočítejte na čelních uzlech

- ať už pro výpočty nebo složitější analýzu výsledků
  - výrazné omezení přístupového uzlu (mnohdy vedoucí až k pádu)
- primárním posláním je příprava úloh a jednoduché/ krátkodobé operace

Využívejte interaktivní úlohy

- *požadavek:*
  - `$ qsub -I -l select=…`
- *možnosti práce:*
  - textový režim
  - grafický režim – VNC přístup
    $ module add gui
    $ gui start
    - viz https://wiki.metacentrum.cz/wiki/Remote_desktop

Interaktivní úlohy

Minimalizujte prodlevy v interaktivních úlohách

- zejména čas mezi spuštěním úlohy a počátkem Vaší práce (spuštěním výpočtu)
  - -> neefektivní využívání zdrojů

Nechte se informovat o spuštění úlohy

- *požadavek:*
  - `$ qsub -m ab -I -l select=…`
    zašle Vám email při spuštění úlohy
    - („`-m abe`” i při jejím ukončení)
- přepínač lze využít i při dávkových úlohách

  pozor při spouštění většího množství úloh!
  - zahlcení Vaší schránky
  - blacklist mailového serveru

Cloudové stroje

Udržujte si přehled o Vámi spuštěných virtuálních strojích

- i Vámi nevyužívaný stroj využívá zdroje infrastruktury
  - -> plýtvání zdroji, které může efektivněji využít někdo jiný

Ukončujte/suspendujte nepoužívané VM

- připravujeme nasazení systému, který Vám bude běžící VM pravidelně (cca každé 3 měsíce) připomínat

  a v případě nereakce (= aktivního prodloužení) tyto suspenduje

Závěrem

Nebojte se zeptat, rádi Vám poradíme!

- v případě problémů prvně **zkuste vlastní analýzu**

  náhled do dokumentace programu

  náhled do dokumentace MetaCentra
  - může mj. obsahovat i tipy na efektivní spouštění aplikací (např. distribuované výpočty v Gaussian, Matlab, Mathematica, aj.), požadavky na licence, …

  náhled do aktualit/novinek MetaCentra

- pokud potřebujete poradit (s problémem či využitím infrastruktury), pište na meta@cesnet.cz nebo support@cerit-sc.cz

  Váš požadavek založí „ticket" v ticketovacím systému
  - rozeslán všem relevantním lidem, umožňuje sledování historie vyřízení požadavku

  doporučení:
  - vždy odpovídejte na daný ticket (přes uvedenou hromadnou adresu) nepište konkrétním osobám
  - pro nové problémy nepoužívejte staré tickety vždy založte nový (tj. pište nový email)

Tomáš Rebok

rebok@ics.muni.cz