MetaCentrum

Novinky z vývoje v MetaCentru

Miroslav Ruda
miroslav.ruda@cesnet.cz

CESNET

Brno, 2011

Novinky z vývoje MetaCentra

- otázky a odpovědi, čím více otázek, tím lépe
- přechod na plánovací systém Torque
- úpravy priorit a fairshare
- virtualizace v PBS a cloudové rozhraní
- diskové prostory
- začlenění CUDA clusteru
- FAQ

Plánovací systém Torque I.

Přešli jsme z plánovacího systému PBSPro na Torque

- volně dostupná implementace, rozumná kompatibilita
- je možno instalovat na všechny připojené clustery
- reimplementovány nejzajímavější vlastnosti
  - plánování, stabilita
- přenesena všechna naše rozšíření
  - plánovač, virtualizace
- další vývoj viditelný i v distribuci Torque
  - více spolupracujících serverů

Vývoj

- virtualizace
- více samostatných serverů = odolnost proti výpadkům
- průběžné odstraňování chyb
- nové vlastnosti

Plánovací systém Torque II.

Více serverů

- samostatné instalace Torque spravující jeden velký cluster/město
- odstranění problémů
  - s výpadky sítě mezi městy
  - škálovatelností
  - rychlostí odezvy
- plánovače vidí i další servery, mohou úlohy přesouvat podle potřeb
- bude nasazeno pro cluster CERITu, postupně v každém městě

Plánovací systém Torque III.

Nové vlastnosti

- požadavek na kompletní obsazení uzlu
  qsub -l nodes=2:nodecpus2#excl
- negativní vlastnosti
  qsub -l nodes=1:cl_skirit:^i386
- v přípravě node packing
  qsub -l nodes=12:ppn=1#pack

Priorita uživatelů

- fairshare = třídění podle propočítaného času za poslední období
- třídění ve frontě, priorita front je silnější
  - prioritní fronty vlastníků clusteru
- aktuální plán
  - promítnout počty publikací do fairshare
  - každá zaregistrovaná publikace snižuje propočítaný čas o X procent/zvětšuje využitelný podíl zdrojů MetaCentra
  - publikace platné jen rok/dva
  - publikace impactované, v RIVu apod.
  - následně zrušit frontu `privileged`
- připravujeme nový registrační formulář
  - snadnější zadávání, import z ISu
  - možnost definovat rozpočítání mezi autory

Virtualizace

Pokračuje vývoj v podpoře v Torque

- pomocí virtualizace provozujeme obrazy debian5 a debian6 současně
  - přepínání podle aktuálních požadavků uživatelů
- v plánu je i obraz SL5 (hlavně pro EGI)
- pomocí přepínání virtuálních strojů řešíme i priority pro vlastníky clusterů
  - pozastavení `backfill` úlohy
- umíme i obraz s MS Windows, pilotní provoz pro Laboratoř bezpečnostních technologií MU
  - poskytujeme čistý obraz s MS Windows
  - uživatelé doinstalují vlastní aplikace
  - pak je možné pustit více kopií
  - uzavřená sít, DHCP, VPN
- testujeme i využití pro výuku
- v přípravě jednorázové postavení uzlu pro výpočet

Virtualizace – cloud rozhraní

Společně s CERIT-SC pracujeme i na cloud rozhraní

- ještě tento rok zpřístupníme pilotní instalaci
- kompatibilita s Amazon EC2
- další rozhraní (OCCI) a GUI pro snadnější použití
- obrazy virtuálních strojů
  - uživatelské
  - originální MetaCentrové, možnost si je změnit
  - v další fázi speciálně zaměřené na některé aplikace (map-reduce)
- možnost vyladit si obraz a přenést ho zpět do MetaCentra
- podpora pro aplikace, kterým gridový přístup nevyhovuje
- další vývoj na integraci/překrytí obou přístupů

Diskové prostory

- NFSv4 jako základ, postupně v každém městě
  - plus na dalších připojených clusterech (např. CERIT)
  - rozumná rychlost pro vzdálený přístup, standard, Kerberos
- všechny svazky viditelné ve /storage/MĚSTO
- nejbližší svazek použitý jako /home
  - stejné pro celý cluster
- hierarchie `/afs`, `/home` a `/scratch` zůstává
  - v plánu je i sdílený Lustre scratch
- spolupráce s datovými úložišti CESNETu, plán:
  - opět automaticky viditelné NFSv4 svazky
  - zálohování domovských adresářů
  - migrace dat do archivu
  - klientské programy pro další služby

Cluster s CUDA kartami

- vlastník Katedra matematiky a Katedra kybernetiky FAV ZČU
- podrobnosti o vybavení v další přednášce
- priorita přístupu pro vlastníka,
  - ale cluster je k dispozici i dalším uživatelům
- plánovací systém Torque plánuje i grafické karty
  - uživatel si pomocí -lcuda=X říká o karty
  - systém spustí úlohu jen na uzlu s volnou kartou
  - karta je zpřístupněna jen vlastníkovi úlohy
- pracujeme i na virtualizovaném řešení
  - HVM virtualizace

FAQ - proč úloha neběží

- úlohy z fronty s vyšší prioritou (vlastníci, privileged)
- v jedné frontě se úlohy řadí podle fairshare
- "strádající úlohy" si rezervují zdroje
- stroj je rezervován nebo je ve frontě "maintenance"
- nemám na stroji účet
- špatná kombinace vlastností
- zdroje není jen CPU, ale i paměť, scratch, software
- na některé stroje nemohou multi-node úlohy
- webové rozhraní v "osobním pohledu" pomůže

FAQ - proč úloha běží pomalu

- příliš krátké úlohy = režie převažuje
  - zabalit do skriptu obsahujícího více úloh najednou
  - přístup přes "pilotní úlohy" - Diane
- úloha z fronty backfill může být pozastavena
- příliš vzdálené uzly pro paralelní úlohu
- špatně zadané požadavky na paměť
- špatné použití filesystémů (`scratch` a `home`)

FAQ MPI I

Torque

- MPI nesmí spouštět pod-úlohy přes ssh, musí použít knihovny Torque
  - aby systém uměl zastavit/hlídat všechny pod-úlohy
- liší se protokol PBSPro a Torque
  - je nutno programy překompilovat
- moduly openmpi, lam, mpich2 jsou předělané
- pozor na vlastní kompilace, MPI obsazené v aplikacích

Testovací příklad pro openmpi

- pro test je dobré použít `cpi.c`
- automaticky se používá Infiniband

FAQ MPI II

```
~$ qsub -I -lnodes=2:debian50:x86_64:plzen:infiniband
qsub: waiting for job 424352.arien.ics.muni.cz to star
qsub: job 424352.arien.ics.muni.cz ready

nympha1-1$ module add openmpi
nympha1-1$ mpicc /software/mpich-1.2.7/amd64_linux26/
           ch_p4/examples/cpi.c -o cpi.openmpi

nympha1$ mpirun ./cpi.openmpi
Process 0 on nympha1-1.zcu.cz
Process 1 on nympha4-1.zcu.cz
pi is approximately 3.1416009869231241, Error is 0.000
nympha1-1$

nympha1-1$ mpirun --mca btl tcp,self ./cpi.openmpi
```

Díky za pozornost!

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

EVROPSKÁ UNIE
EVROPSKÝ FOND PRO REGIONÁLNÍ ROZVOJ
INVESTICE DO VAŠÍ BUDOUCNOSTI

2007–13
OP Výzkum a vývoj pro inovace