MetaCentrum

Miroslav Ruda

CESNET

2. 12. 2014

Agenda

Národní gridová infrastruktura

- přehled služeb MetaCentra
- aktuální stav
- výpočetní grid
- cloudové prostředí

MapReduce výpočty

MetaCentrum Národní gridová infrastruktura

Distribuované prostředí pro sdílení výpočetních zdrojů

- motivací je optimální využití zdrojů (nejen hardwaru)
  - přenesení nárazové zátěže na volnější zdroje
  - a využití jiných zdrojů při výpadku
  - poskytnutí vlastních dočasně volných zdrojů
- aktuálně clustery vlastněné různými organizacemi
- možnost pořídit si cluster jen na standardní kapacitní požadavky
- spolupráce týmů z různých organizací, sdílení dat

MetaCentrum

- součást e-infrastruktury spravované CESNETem
  - jednotný účet, propojené služby
- pracuje v koordinaci s dalšími projekty IT4i a CERIT-SC
- přidáváme podporu zpracování rozsáhlých dat, nových typů výpočtů

MetaCentrum

- clustery a výkonné servery, úložné kapacity
  - Plzeň, Praha, Brno, Ostrava, ČB
- aktuálně 10 000 jader, (4 000, 4 800 CERIT-SC)
  - clustery CEITEC, ZČU, JČU, UK, MU, AV ČR, ČVUT
- přes 1 PB diskových prostor na zpracování dat
- centrální správa uzlů, účtů, úloh, aplikačního softwaru
- dalších 4 000 jader a 2 PB přes FZU do EGI (LHC)
- rozsáhlý aplikační software
- "placení" formou publikací s poděkováním
  - přístup všem akademickým pracovníkům a studentům bez omezení, bez podávání projektů
  - publikace využívány pro určení priority uživatele
- výpočty gridové, cloudové, MapReduce

MetaCentrum

Novinky

CESNET

- zprovozněna třetí generace GPU clusteru
  - 30 uzlů 2× NVIDIA Kepler K20, 100 TB home
- tento rok přibude
  - 400 TB home v Brně
  - 4 SMP servery (každý 4×8 jader, 512 GB RAM)
  - 18 uzlů cloudu (každý 2×8 jader, 256 GB RAM)
  - 27 uzlů Hadoop (2×8 jader, 128 GB RAM, 12×4 TB disk)
- příští rok - obnova HTC clusterů `nympha` a `tarkil`

CERIT-SC

- druhý SGI UV v Brně
  - celkem 384 jader, 6 TB RAM

Připojeny clustery na FEL ČVUT, FZÚ AV ČR, uzel Elixiru na JČU

Výpočetní grid I

- využívá většinu zdrojů, orientace na dávkové úlohy, hromadné zpracování, dlouhé vědecké výpočty
  - cmdline rozhraní, pro vybrané obory existují webové portály (Galaxy, Mascot)
- různý typ výpočetních uzlů
  - klasické dvouprocesorové uzly orientované na HTC úlohy
  - uzly s GPU kartami (30 uzlů, každý 2× NVIDIA Kepler K20)
  - větší SMP uzly (4 procesory, 512 GB RAM)
  - 2× SGI UV2 (6 TB RAM každý)
- vlastní vývoj a výzkum v oblastech souvisejících s infrastrukturou
  - plánování úloh (lepší plánovače, distribuované torque světy)
  - virtualizace
- integrováno do evropské infrastruktury EGI
  - zajímavé pro skupiny s partnery v zahraničí

Výpočetní grid II

- podpora začlenění clusterů vlastněných jinými organizacemi
  - začlenění do centrální správy
  - vlastníci mají pro svoje zdroje prioritní přístup, určují na jakou část clusteru mohou ostatní uživatelé, jak dlouhé úlohy mohou použít ...
  - možnost vlastní správy skupin pro definování přístupu k frontám, datům
- nabízíme spolupráci při úpravách prostředí pro velké projekty
  - příklad projekt Elixir
  - prototyp národního uzlu budovaný na zdrojích CESNETu
  - instalována řada softwarových balíků, přístupná všem
  - až produkční zdroje budou požadovány z projektu Elixir
- projektové zajištění
  - národní - VI CESNET, VI Elixir
  - mezinárodní - EGI Engage, Elixir - Escensum

Výpočetní grid - software

- aplikační software jako další zdroj pro sdílení, centrální správu
- část aplikačního software dostupná pro celou ČR
- Matlab, gridMathematica, Maple
  - CERIT-SC - Mathematica, Matlab DCS
- vývojové prostředí Intel, PGI, Totalview, Allinea
- Gaussian + Gaussian-Linda, Amber
  - CERIT-SC - Turbomole + Molpro
  - Mascot
- Ansys CFD (Fluent, CFX, HPC)
  - CERIT-SC - Ansys Mechanical
- volně dostupný software (500, 190 upgrade)
  - nově zejména bioinformatika

Výpočetní grid - data

Souborové systémy rozdělené podle využití uložených dat

- lokální scratch disky pro dočasná data (HTC úlohy, <1 TB)
- 3× sdílený scratch filesystém, viditelný přes jeden cluster – paralelní zpracování větších dat (10 TB)
- home v každém městě – semipermanentní, zálohovaná data (100 TB)
- projektové adresáře – semipermanentní, pro sdílení v rámci projektu
- zpřístupněná data ze souborových serverů vlastníka clusteru
- přímo přístupné souborové systémy DU – zálohování, odložení aktuálně nepoužívaných dat

Výpočetní cloud - IaaS I

- určeno pro vědecké výpočty nebo služby související s výpočty
  - ne pro obecné hostování webových služeb
- integrované s MetaCentrem, ale s přidanou přihláškou (ssh)
- přístup přes webové rozhraní OpenNebuly (heslo) nebo cmdline (certifikáty) pro správu virtuálních strojů
  - ssh nebo remote desktop pro přístup do systému
- výpočetní uzly s možností využití vlastního systému
  - uzly totožné s MetaCentrem – pro potřeby ladění softwaru
  - uzly dodané mezinárodní VO ve které jsou uživatelé zapojeni – dodaný obraz, předinstalovaný software
  - předinstalované obrazy se základním systémem
    - Debian 6/7, Centos 6/7, Scientific Linux 6, MS Windows 2012 Server/2008
  - uzly s uživatelem připravenou instalací systému a vlastního softwaru
    - kopie desktopu, serveru, stažený obraz

Výpočetní cloud - IaaS II

- webové služby pro zadání úloh, permanentní část výpočetního prostředí
  - Galaxy portál nabízený MetaCentrem (autentizace prostředky MetaCentra, úlohy běží v gridu)
  - platformy připravené MetaCentrem nebo jinými skupinami
    - Hadoop
- přístup k persistentním datům
  - přístup k filesystémům MetaCentra (NFS, GPFS)
  - vlastní persistentní obraz disku, montovaný jako druhý disk
  - objektový storage server (S3)
- sítě
  - uzly mohou být uzavřené do VLAN, mít pouze privátní IP adresy
  - VLAN umožňuje i L2 sítě, propojení do domácí sítě, zahrnutí gridových uzlů
  - příklad: projekt Kypo

Výpočetní cloud - IaaS III

- evropská infrastruktura FedCloud jako součást EGI
  - OCCI rozhraní jako sjednocující prvek
    - OpenNebula, OpenStack, Oceanos
  - CESNET vyvíjí rOCCI
    - OpenNebula, Amazon, výhledově Azure, VMWare
- příklady využití
  - BioVeL, WeNMR, DIRAC, Kypo
  - MS Windows – Peachnote, Physiome
  - `https://wiki.egi.eu/wiki/Federated_Cloud_Communities`
- GUI: `https://cloud.metacentrum.cz/`
- dokumentace `http://meta.cesnet.cz/wiki/Kategorie:Cloudy`

Map Reduce v MetaCentru

- nové prostředí, nabízené dočasně v cloudovém prostředí
  - PaaS v cloudové terminologii
- Apache Hadoop prostředí
  - MapReduce, YARN, Hive, Pig, HBase, ...
- na konci tohoto roku bude provozováno na vlastním dedikovaném hardwaru
  - 3 servery sloužící jako front-end, metadata, management
  - 24 výpočetních serverů
  - dual-socket, 128 GB RAM, 12×4 TB v každém uzlu
  - celková kapacita HDFS filesystému 1 PB
- opět integrováno s MetaCentem
  - přihláška, autentizace, accounting

Map Reduce – motivace I

- MapReduce přístup publikován firmou Google (2003/2004)
- typ výpočtů rozpoznán jako klíčový i dalšími firmami s potřebou prohledávat velká nestrukturovaná data
  - logy (webových) serverů
  - textové soubory (viz náš příklad počítání slov)
  - XML/JSON soubory, dump databáze
  - bioinformatická data
- volně dostupná implementace HDFS a MapReduce pod hlavičkou organizace Apache
  - přispívají i firmy typy Facebook, Amazon, Twitter, IBM
- vznikají i formy poskytující komerční podporu
  - Cloudera, Hortonworks
- postupně vznikají další nadstavby (Pig, HBase, Hive, YARN)

Map Reduce – motivace II

- motivace – tradiční servery nejsou schopné zpracovat data v potřebném rozsahu
  - posilování serverů naráží na finanční limity
  - nedostatek paměti
  - rychlost I/O na serveru
- nabízí se možnost data distribuovat přes sadu uzlů clusteru
  - levnější, zaručí se škálovatelnost nákladů
  - data se zpracovávají paralelně po částech
- potřeba koordinace, zaručení stability a konzistence při výpadku uzlu clusteru
- potřeba výpočetního prostředí ve kterém lze jednoduše takové úlohy zpracovávat

Map Reduce výpočty – HDFS

- speciální filesystém pro uložení dat
  - data jsou při uložení do Hadoop clusteru rozdělena na menší kousky, ty jsou rozhozené přes jednotlivé uzly clusteru
  - soubory nelze jednoduše přepisovat, pouze prodlužovat
- při samotném výpočtu se data již nestěhují
  - úlohy se stěhují za daty
- systém zaručuje, že je uloženo několik kopií každého kusu dat
  - při výpadku uzlu sám zajistí dodělání chybějících kopií
- uzly mezi sebou komunikují minimálně, uživatel programuje v prostředí, kde neřeší výpadky, síťovou komunikaci . . .
- pro nalití dat do HDFS existují i specializované nástroje
  - (Flume – logy, Sqoop – data z databází)

Map Reduce výpočty

- typická úloha se skládá z Map a Reduce fází, jazyk Java
- komunikace probíhá přes dvojice (klíč, hodnota)
- úloha pracuje nad jedním blokem dat
  - Master úlohy koordinuje, plánuje tam kde leží zpracovávaná data, ve vlastní režii řeší výpadky apod.
- Mapper zpravidla čte (jméno souboru,data), produkuje seznam (klíč, hodnota)
- Reduce proces dostane dvojice s jemu přiřazeným klíčem (setříděné podle klíče), produkuje výsledný (klíč, hodnota) do HDFS

```
Map(input-key,input-value)
  {foreach word w in input value; emit(w,1)}
Reduce(key, list of value)
  {foreach value sum=sum+value; emit(key,value)}
```

Hadoop ecosystem

- Hive – SQL-like interface pro dotazy nad Hadoop daty
- HBase – sloupcová databáze nad HDFS (vybírá jen řádky, omezený jazyk pro dotazy, jeden klíč, vstup pro MapReduce)
- Flume – nástroj pro průběžné produkování dat
- YARN – Haddop 2.0, workflow z MapReduce úloh
- Pig – skriptovací jazyk pro tranformaci velkých dat v Hadoopu

```
A = load './input.txt';
B = foreach A
    generate flatten(TOKENIZE((chararray)$0)) as word;
C = group B by word;
D = foreach C generate COUNT(B), group;
store D into './wordcount';
```

MetaCentrum

Děkuji za pozornost

http://www.metacentrum.cz