Databáze ve fyzice vysokých energií

Vladimír Jarý[1]

[1]Fakulta jaderná a fyzikálně inženýrská
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
mailto:Vladimir.Jary@cern.ch

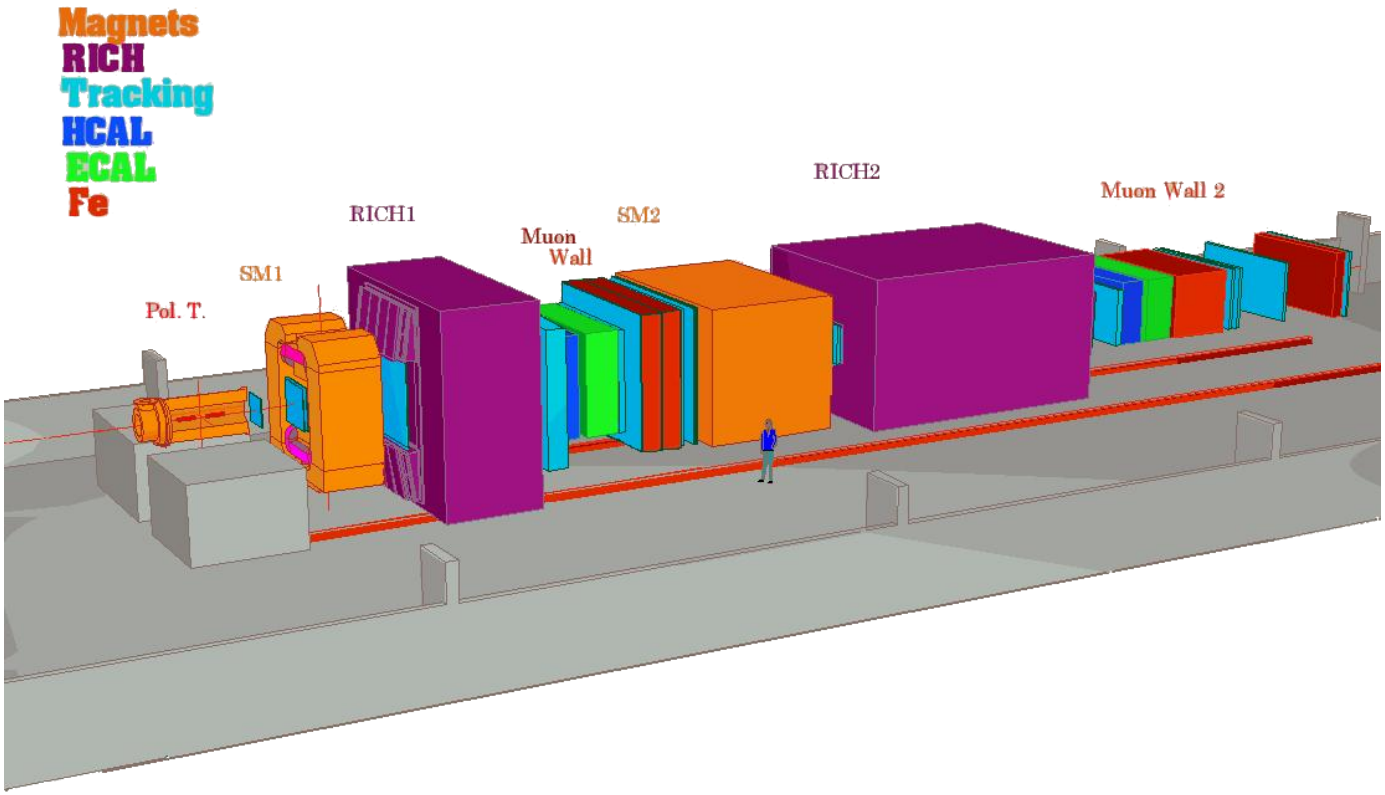
InstallFest 2011
Školicí centrum Silicon Hill, Praha
5. března 2011

Přehled

Představení experimentu COMPASS

- COMPASS: COmmon Muon and Proton Apparatus for Structure and Spectroscopy
- experiment s pevným terčem umístěný na urychlovači SPS (Super Proton Synchrotron) v laboratoři CERN [1]
- vědecký program: studium struktury a spektroskopie hadronů
  - experimenty s hadronovým a s mionovým svazkem
  - program schválen vědeckou radou CERN v roce 1997
  - instalace v letech 1999-2000, sběr dat od roku 2002
  - momentálně se čeká na schválení 2. fáze experimentu [3] (program na dalších cca 5 let)
- 240 vědců z 11 zemí
- česká účast: vývoj fotonásobičů pro detektor RICH, kryogenika polarizovaného terče, sběr dat

Popis experimentu



Systém detektorů, svazek částic dopadá na terč zleva, délka spektrometru přibližně 60 m; obrázek převzat z [4]

Detekce částice

- interakcí částic svazku s polarizovaným terčem vznikají sekundární částice
- průlet částic detekován systémem detektorů:
  1. měření energie částic (elektromagnetický, hadronový kalorimetr)
  2. identifikace částic (RICH detektor)
  3. určení trajektorie částice (různé druhy drátových komor)
- událost (event): data sesbíraná z různých detektorů popisující průlet částice
- cyklus urychlovače SPS: svazek (beam) není spojitý, skládá se z úseků (spills, bursts)
  - systému pro sběr dat používá vyrovnávací paměti pro rozložení zátěže na celý cyklus urychlovače

Struktura systému pro sběr dat

- systém pro sběr dat (DAQ, data acquisition) se skládá z několika vrstev:
  1. frontend elektronika detektorů ($\sim$ 250000 kanálů)
     - provádí načtení (readout) a digitalizaci dat
     - načtení dat vyvoláno trigger systémem, který zároveň šíří identifikátor události a časovou značkou
     - data z několika kanálů shromažďována v modulech CATCH, GeSiCA, kde je doplněna subevent hlavička
  2. ROB servery (readout buffers): klasické servery doplněné o *spillbuffer* PCI kartu
     - slouží jako vyrovnávací paměť (využití pauzy mezi úseky)
  3. EVB servery (event builders): klasické servery, spojené s ROB vrstvou prostřednictvím Gb Ethernetu
     - využívají subevent hlaviček pro sestavení událostí
     - metainformace o událostech uloženy do Oracle databáze
     - vlastní soubory s událostmi odeslány na páskové permanentní úložiště CASTOR

Software pro sběr dat

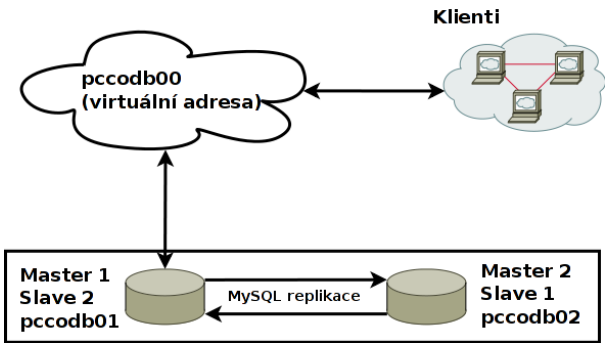
- balík DATE (Data acquisition and test environment) [3]
  - navržen pro experiment ALICE
  - multiprocesorové distribuované prostředí
  - důraz na škálovatelnost
  - funkcionalita:
    - readout a event building
    - řízení toku dat (Event distribution manager EDM)
    - řízení
    - interaktivní konfigurace (nastavení v MySQL)
    - dohled nad kvalitou dat, částečná online analýza (COOOL)
    - ukládání deníku (MySQL databáze)
    - filtrování událostí (Cinderella)
- operační systém: Scientific Linux CERN 4
  - distribuce založená na RHEL4 (vlastní repozitáře s CERN software, podpora pro AFS systém, . . .)
  - v současnosti probíhá migrace na SLC5
  - 32b verze OS (s výjimkou databázových a NFS serverů)

Online a offline databáze

1. online databáze
   - spravuje informace přímo související se sběrem dat
   - využívána především operátory experimentu pro dohled a nastavení systému DATE
   - servery umístěny v experimentální hale
   - software: MySQL
   - více na dalších stránkách prezentace . . .
2. offline databáze
   - spravuje metasoubory s informacemi o událostech
   - využívána při offline analýze dat
   - servery umístěny ve výpočetním středisku (cca 5 km od experimentální haly)
   - software: Oracle

Původní architektura online databáze

- dva fyzické databázové servery
- master–master replikace (každý server je zároveň master a zároveň slave)
- klienti přistupují k serverům přes přes virtuální adresu, která ukazuje přímo na jeden z fyzických serverů
- při detekci výpadku serveru je virtuální adresa dočasně převedena na zbývající server



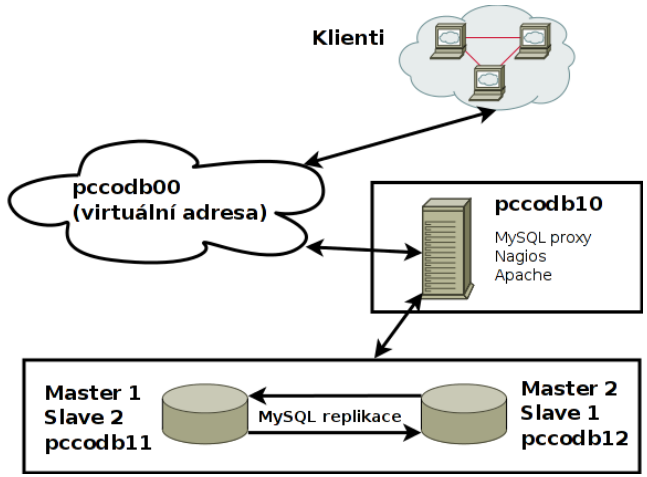
Původní architektura

Problémy s původní architekturou

- během sběru dat v roce 2009 časté výpadky, pro rok 2010 očekávaný nárůst zátěže
- fyzický server spravuje 20 logických databází:
  - *beamdb20??*: informace o svazku, detektorech
  - *DATE20??_log*: chybová hlášení systému DATE
  - *runlb*: deník
- další služby: web server *Apache*, dohledový systém *Ganglia*, server *infoLogger*
- problém: zastaralost serverů (HW i SW):
  - SLC4 (linux 2.6.9, i386), MySQL 4.1.22 (úložiště MyISAM)
  - 2× Intel Xeon @3 GHz, 4 GB RAM (resp. 1× Intel Xeon @3 GHz, 3 GB RAM)
- návrh řešení: rozdělit logické databáze na dva páry fyzických serverů, pátý server použít pro další služby

Popis nové architektury

- dodány pouze 3 nové servery
- dva servery použity pro MySQL, opět synchronizovány master–master replikací
- třetí server pro pomocné úlohy:
  - proxy server (*MySQL Proxy*)
  - dohled (*Nagios*)
  - web server (*phpMyAdmin*, deník)



Implementovaná architektura

- proxy server předává všechny dotazy na jeden ze serverů
- stejná virtuální adresa nyní ukazuje na proxy server ⇒ přechod na novou architekturu je pro klienty transparentní

Migrace na novou architekturu

- proběhla výměna hardware i software:
  - 2× Intel Xeon E5420 @2,5 GHz (2 × 4 jader), 16 GB RAM
  - 64-bit SLC5 (linux 2.6.18), MySQL 5.1.45
- instalace MySQL ze zdrojových kódů (povolena podpora pro velmi velké tabulky, pro dělené tabulky)
- přesun struktury a dat ze starých na nové servery pomocí klientských programů *mysqldump* a *mysql*
- východiskem pro nastavení serverů šablona *my.cnf.huge*: nastavení replikace, zaznamenávání pomalých dotazů, zvětšení limitu pro otevřené soubory
- test integrity dat:
  - porovnání dumpů pomocí nástrojů *md5sum* a *diff*
  - dumpy některých tabulek se lišily: příčinou byla změna definice type *DECIMAL(m, n)*
- nastavení MySQL Proxy a připojení klientů

Dohled a zálohování

1. Dohledový systém
   - systém Nagios pro dohled na databázové servery (na ostatní DAQ servery dohlíží Ganglia)
   - agent NRPE pro komunikaci mezi Nagiosem a monitorovanými servery
   - sledování zatížení CPU, stav plánovače, volné místo na discích, stav MySQL serveru, stav replikace, teplota jader
   - při detekci incidentu zaslán e-mail správcovi a případně vykonána nějaká akce (přeprogramování MySQL Proxy)
2. Typy záloh
   - denní, hodinová: vytvářeny nástroji *mysqldump*, *gzip*
   - inkrementální: binární log vytvářený při replikaci
   - geografická: databáze je replikována do výpočetního centra

Shrnutí

- migrace dokončena před začátkem sběru dat
- během testování se vyskytl problém s tabulkou s monitorovacími daty elektromagnetického kalorimetru
  - několikrát došlo k uzamčení tabulky
  - příčinou byla chyba v aplikaci *Cinderella* (online filtr)
  - největší tabulka, obs. více než miliardu záznamů
- během sběru dat nezaznamenán žádný větší problém
- nové databázové aplikace (*daqmon*)
- nárůst objemu dat z 54 GB (duben) na 138 GB (listopad)
- prostor pro navýšení výkonu:
  - využití MySQL Proxy pro distribuci zátěže
  - dělené (partitioned) tabulky?

Literatura

- P. Abbon et al. (the COMPASS collaboration): *The COMPASS experiment at CERN*, In: Nucl. Instrum. Methods Phys. Res., A 577, 3 (2007) pp. 455–518
- Ch. Adolph et al. (the COMPASS collaboration): *COMPASS-II proposal*, CERN-SPSC-2010-014; SPSC-P-340 (May 2010)
- T. Anticic et al. (the ALICE collaboration): *ALICE DAQ and ECS User's Guide*. CERN, ALICE internal note, ALICE-INT-2005-015, 2005.
- *COMPASS page* [online]. 2010. Available at: `http://wwwcompass.cern.ch`