VÝVOJ PROGRAMOVÉHO VYBAVENÍ PRO HROMADNÉ ZPRACOVÁNÍ DAT - HADRON RUN COMPASS

Tomáš Liška
České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská
Tomas.Liska@cern.ch

ABSTRAKT:

V příspěvku popisuji principy řízení distribuovaných výpočtů pro experiment COMPASS, části hadronový run. Zabývám se konfigurací, řízením výpočtů, sledováním stavu v kvazi-reálném čase, zpracováním chyb při výpočtech a získáváním dat pro jejich distribuci na počítačový svazek, zpracování a konsolidaci výsledků.

ABSTRACT:

COMPASS is a collider experiment running on the Super Proton Synchrotron at European Organization for Nuclear Research (CERN). This paper describes the process of the massive data production system developed for hadron part of the COMPASS experiment at CERN. The state machine based solution is discussed including principles of active monitoring and fail-state decision making. The computing job life cycle is described from the developer's point of view, too.

KLÍČOVÁ SLOVA:

COMPASS, hromadné zpracování dat, distribuované výpočty

ÚVOD

Experimenty v oblasti fyziky elementárních částic patří mezi nejnáročnější oblasti současné vědy, a to i po stránce počítačového zabezpečení. Příprava experimentu trvá několik let a vlastní běh experimentu, tedy sběr dat, také. V tomto článku se zaměříme na experiment COMPASS, který v současné době běží v CERN. Ukážeme si způsob nasazení počítačů ve fyzice elementárních částic a problémy, s nimiž se přitom musí programátoři vyrovnat.

STRUČNÉ PŘEDSTAVENÍ EXPERIMENTU

Název COMPASS je zkratkou slov COmmon Muon and Proton Apparatus for Structure and Spectroscopy, tedy společné zařízení pro mionovou spektroskopii a strukturu protonů. Byl navržen v roce 1996 [1], schválen v CERN v roce 1997. Technicky byl spuštěn v roce 2000, v roce 2002 začal fyzikální běh, tedy i sběr dat (data acquisition, DAQ). Předpokládá se, že první etapa tohoto experimentu skončí v r. 2010.

Na experimentu COMPASS se podílí řada univerzit a výzkumných ústavů z více než 10 států; členy

české skupiny jsou i pracovníci a studenti z FJFI ČVUT, MFF UK a z TUL.

V současné době je připraven koncept návrhu na pokračování tohoto experimentu po dalších 5 let se třemi základními programy měření.

ORGANIZACE DAT FYZIKÁLNÍHO EXPERIMENTU COMPASS

Úložištěm dat pro experimenty probíhající v CERNu se využívá centralizovaného datového úložiště CASTOR – Cern Advanced STORage manager [2]. Toto úložiště využívají všechny experimenty prostřednictvím knihoven pro přístup k souborům zde uložených.

Data jsou v úložišti dělena do jednotlivých souborů nazývaných chunk. Určitý souhrn takových souborů – chunk – dává dohromady tzv. run. Tak označujeme ucelený soubor chunků, který obsahuje naměřená data ze všech detektorů.

Samotná naměřená data nestačí. Při zpracování je třeba znát konkrétní nastavení detektorů, provozní podmínky a mnoho dalších popisných údajů. Tato tzv. metadata jsou pro jednotlivé runy uložena v databázi. Souhrnem chunků, tvořících jeden run, spolu s popisnými metadaty a se záznamy v tzv. log-booku (deník prováděných měření) tak máme k dispozici množinu údajů, na základě které můžeme určit kdy, jak, za jakých podmínek, s jakými výsledky a případně s jakými problémy byla data konkrétního runu změřena.

Každý run trvá typicky 30–60 minut. Během této doby jsou zaznamenávána data z běžícího experimentu. Průměrná velikost souboru jednoho chunku je cca 0,82 GB. Průměrný run obsahuje cca 200 chunků. Každý run je identifikován svým číslem, které je v rámci celého experimentu jednoznačné.

Abychom získali představu o celkové náročnosti kladené na uložení dat, je třeba ještě uvážit, že sběr dat probíhá po několik let. Dělí se na tzv. Runy označované podle let. (Stejný termín – run – je zde používán v jiném významu než předtím; abychom je alespoň trochu rozlišili, píšeme ho s velkým počátečním písmenem.) Například Run2004 obsahuje tato množství dat (uvádíme pouze fyzikální měření, nezahrnujeme v to testovací měření):

Počet runů	Počet chunků	celková velikost souborů (v GB)
3091	618274	508078

Vyjdeme-li ze skutečnosti, že 1 GB je více než 10^9 bajtů, zjistíme, že máme uloženo cca 0,48 PB (petabajtů) dat. Snadným výpočtem lze zjistit, že při kapacitě 160 GB na jeden disk bychom potřebovali téměř 3200 disků, abychom tato data byli schopni uložit v jednom PC. Protože to samozřejmě není reálné, využíváme datové úložiště v CERN.

S podobným problémem se musíme potýkat i při samotném výpočtu. Takový objem dat nelze zpracovat na jednom či několika PC, proto využíváme služeb tzv. clusteru, tedy svazku výpočetních stanic.

VÝPOČETNÍ SVAZEK LSF

Výpočetní svazek používaný v CERN, je složen z několika desítek tisíc výpočetních uzlů, kde každý uzel definuje následující konfigurace:

- procesory – 2-8 jader na 1-2 fyzické procesory
- lokální tzv. dočasné datové úložiště
- 4-16 GB paměti
- nezbytné softwarové vybavení pro výpočty

Všechny výpočetní uzly svazku jsou propojeny volně vázanou propojovací sítí na bázi gigabitové sítě Ethernet a protokolu TCP/IP a k jejich řízení je využíván systém Load Balancing Facility (LSF). LSF poskytuje služby a prostředky pro:

- řízení parametrizovaných prioritních front pro výpočetní úlohy - queues
- přípravu a odeslání úlohy k výpočtu - bsub
- sledování aktuálního stavu konkrétní fronty - bpeek
- sledování alokovaných zdrojů výpočetních (uzlů) – blist
- zastavení úlohy – bkill

PRODUKČNÍ SYSTÉM PRO ŘÍZENÍ VÝPOČTŮ ZPRACOVÁNÍ HADRON RUNU

Pro zpracování dat hadronové části experimentu COMPASS jsme vyvinuli výpočetní systém HPS – Hadron Prudction System. Systém je postaven ve spolupráci s fyzikální skupinou, která nám poskytla analytické nástroje pro zpracování dat CORAL [3] a PHAST [4]. Naším úkolem bylo řešení systému řízení výpočtů nad celým objemem dat hadronové části experimentu COMPASS.

HPS je založen na řízení výpočtů prostřednictvím zmíněného LSF s aktivním sledováním v průběhu celého zpracování dat. Data jsou nahrávána z datového úložiště CASTOR a související metadata z databázového systému ORACLE RAC. Výsledky zpracování jsou zapisovány zpět na CASTOR; jedná se o tyto výstupy:

- po zpracování nástrojem CORAL – tzv. miniDST (Data Summary Tape Files)
- po zpracování nástrojem PHAST – tzv. microDST
- příslušné log-soubory jednotlivých analytických nástrojů

Do databáze je po každém zpracovaném souboru zapsána informace o jeho zpracování včetně informací o výsledcích zpracování. Tato informace o zpracování nazývána off-line data production information se připojuje k již dříve pořízeným metadatům a doplňuje tak komplexní informaci o každém souboru dat experimentu.

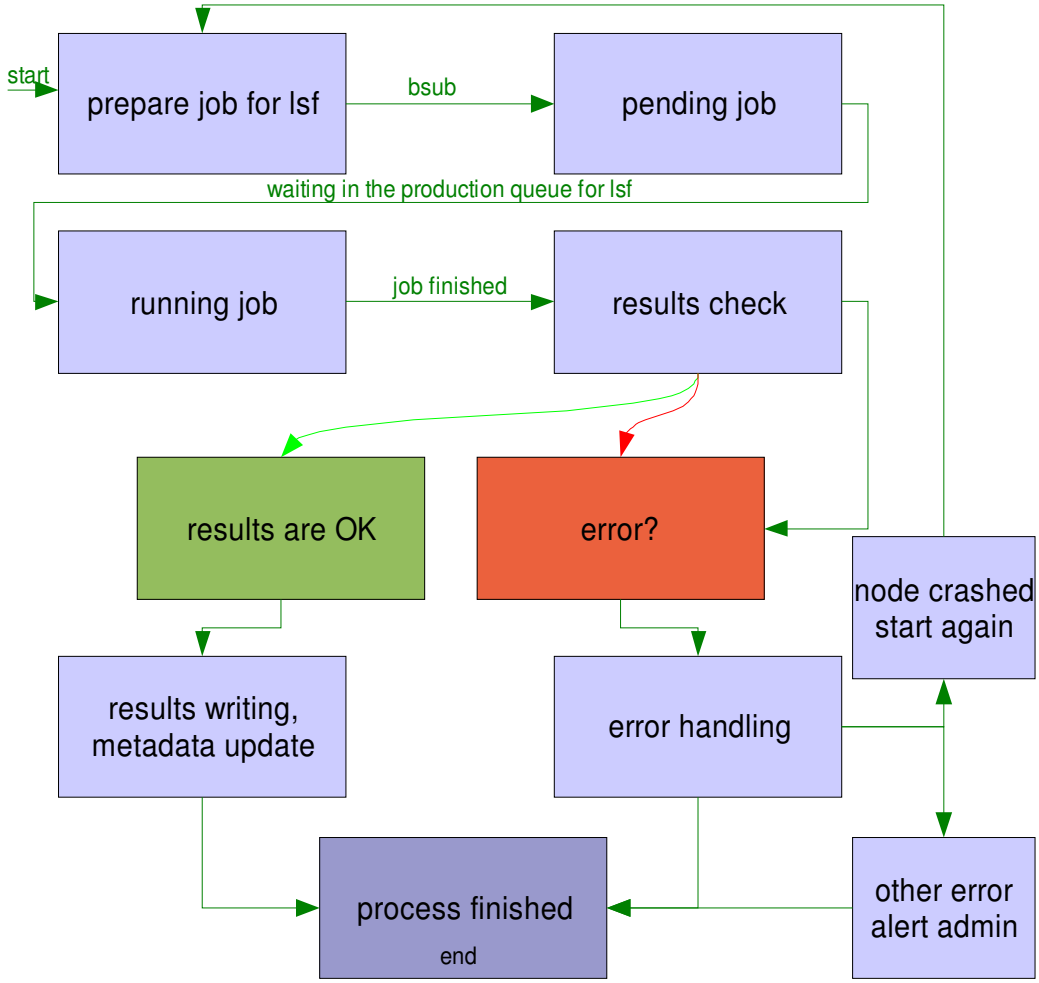
PRINCIPY ŘÍZENÍ VÝPOČTU

Celý proces zpracování dat je řízen stavovým automatem. Během zpracování každého souboru udržujeme v systému veškeré informace nezbytné pro odeslání, výpočet a jeho řízení, konsolidaci výsledků, uložení výsledků a tzv. úklidu po výpočtu. Systém detekuje chybové stavy a na základě jejich typu se pokouší o automatické korekce či informuje o nemožnosti výpočet pro konkrétní soubor dokončit s příslušnými výstupy chybových stavů a hlášení.

Životní cyklus zpracování jednoho souboru sestává z těchto sekvenčních kroků:

- příprava a odeslání úlohy na výpočetní svazek pomocí LSF
- sledování stavu úlohy během výpočtu
- detekce případných chybových stavů a jejich náprava
- konsolidace výsledků, uložení výsledků, zápis na permanentí úložiště a aktualizace metadat v databázi

Schéma řešení stavového automatu popisuje následující diagram.



Hadron Data Production system – stavový automat řízení

Popišme si podrobněji celý proces:

1. Ze seznamu nezpracovaných souborů vybereme jeden soubor. K tomuto souboru získáme z databáze metadata. Součástí těchto metadat je nastavení detektorů a další množství parametrů. Tato nastavení a parametry použijeme k přípravě konfiguračních souborů pro analytické nástroje CORAL a PHAST.

2. Všechny připravené údaje ve formě souborů odešleme do výpočetního svazku pomocí příkazu bsub rozhraní LSF. Soubor údajů definuje výpočetní úlohu a úloha je zařazena do jedné z front svazku jako čekající – pending job.

3. Automatizovaný systém řízení svazku LSF periodicky kontroluje dostupné výpočetní zdroje a postupně úlohy čekající odesílá na konkrétní vhodný vybraný výpočetní uzel. V momentě, kdy je úloha odeslána na konkrétní uzel, přechází do stavu běžící – running job. Systém HDP stavy jednotlivých úloh zjišťuje pomocí seznamu úloh běžících na svazku a pomocí získávání detailů o jednotlivých úlohách – blist, bpeek.

4. V momentě, kdy je úloha na svazku označena jako ukončená, je třeba zkontrolovat její stav a návratové hodnoty ukončení. Rozlišujeme dvě ukončení pomocí stavu testování výsledků – results check.

5. Ukončení je korektní – results check OK – úloha dokončila svůj výpočet a to bez chyb. V tomto případě se pokračuje bodem 7.

6. Ukončení není korektní – results check false – úloha byla ukončena z různých důvodů nestandardně. V tomto případě se rozhoduje dle návratových a stavových informací ukončené úlohy o automatizovaném znovuspuštění výpočtu na jiném vhodném uzlu výpočetního svazku (tedy přecházíme zpět do bodu 3) nebo se úloha označuje jako nedokončená a pokračuje se bodem 7.

7. Po dokončení výpočtu se zapisují na permanentní úložiště všechny získané výsledky, aktualizují se metadata a úloha je označena jako zpracovaná.

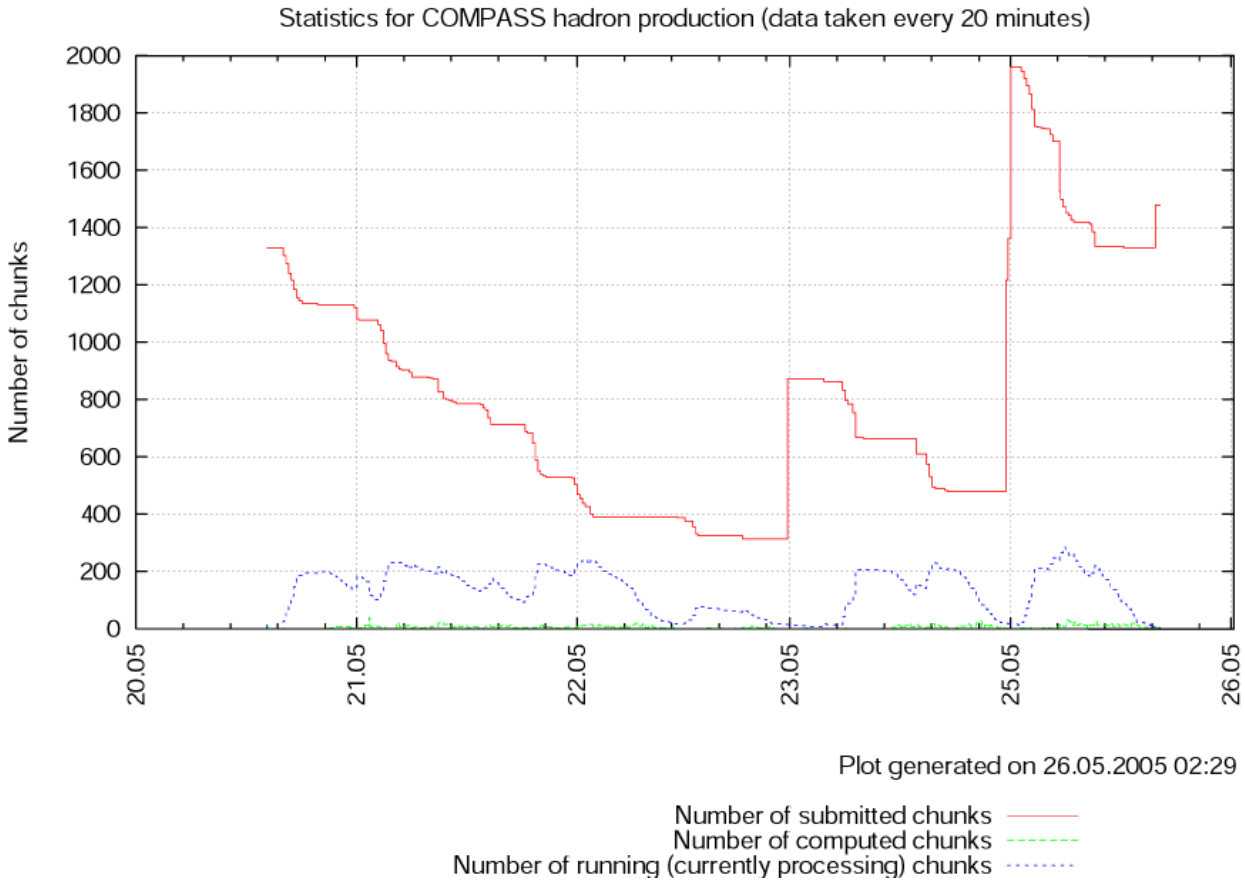
8. Životní cyklus výpočetní úlohy končí.

Tento stavový automat provádí celý proces pro každou jednotlivou úlohu opakovaně stejným způsobem. Odlišnosti jsou pouze ve výběru vhodného uzlu pro výpočet, který se ponechává v kompetenci řízení LSF, tzv. Load Balance Scheduleru.

SLEDOVÁNÍ VÝPOČTU A ŽIVOTNÍHO CYKLU ÚLOH

Celý proces řízení výpočtu systému HDP jsme aktivně sledovali prostřednictvím pravidelně spouštěných skriptů. Základní sledované parametry výpočtu jsou:

- seznam všech úloh ve výpočetním svazku a jejich aktuální stavy
- vytíženost přidělených systémových zdrojů – alokovaných výpočetních uzlů
- aktuální stav konkrétní běžící úlohy včetně pohledu do aktuálního stavu jejích výstupů zapisovaných na standardní výstup



Z grafu je zřejmé kolik úloh je aktuálně odeslaných na výpočetní svazek (červená čára), vidíme kolik úloh je aktuálně běžících (modrá přerušovaná čára) a u kolika úloh byl výpočet v daný časový okamžik dokončen (zeleně). Graf poskytuje globální pohled o vytíženosti alokovaných výpočetních zdrojů. Dalšími skripty jsme sledovali detaily pro jednotlivé úlohy v případech potřeby, tedy zejména při výskytu libovolné chyby, která znemožňuje další pokračování výpočtu konkrétní úlohy.

SOUHRNNÉ VÝSLEDKY ZPRACOVÁNÍ HADRONOVÝCH DAT EXPERIMENTU COMPASS

Celkový objem dat ke zpracování hadronové části experimentu činí 496 TB. Zpracovali jsme více než 98% všech souborů. Zbylá dvě procenta byla po analýze označena jako poškozená data, a proto nevhodná pro zpracování. Toto rozhodnutí bylo učiněno po konzultaci s fyzikální skupinou.

Další detaily o výsledcích a návaznostech je možné najít v Massive Data Production on Distributed Computation Environment - principles, algorithms, COMPASS results [5], Massive Data Production on Distributed Computation Environment [6], Experiment COMPASS a počítače [7] a Počítače ve světě fyzikálních experimentů [8]. Na výsledky zpracování dat hadronové části experimentu COMPASS navazuje má disertační práce The Cluster and GRID Computing on Mass Data Production Systems for High Energy Physics Experiments [9].

PODĚKOVÁNÍ

Práce na tomto příspěvku byla podporována z grantu MŠMT LA08015 a SGS 10/094.

LITERATURA

[1] F. Bradamante: *The Common Muon and Proton Apparatus for Structure and Spectroscopy. Proposal.* CERN/SLPC 96-14, SPSC/P297, March 1996

[2] *CASTOR* [online]. 2010 [cit. 20. března 2010]. Dostupné na: http://castor.web.cern.ch/castor

[3] *CORAL* [online]. 2010 [cit. 20. března 2010]. Dostupné na: http://pool.cern.ch/coral/

[4] *PHAST* [online]. 2010 [cit. 20. března 2010]. Dostupné na: http://ges.web.cern.ch/ges/phast/

[5] Liska, T.: *Massive Data Production on Distributed Computation Environment - principles, algorithms*, COMPASS results, RIKEN, Wako-shi, Japan, 08/2005

[6] Kral, A., Liska, T.: *Massive Data Production on Distributed Computation Environment*, Nikko, RIKEN Spin Fest Workshop, Japan, 08/2005. In proceedings RIKEN Spin Fest Nikko Workshop. 10 pages

[7] Král, A., Liška, T., Virius, M.: *Experiment COMPASS a počítače*, Prague, Czech Republic, 05/2005. In Československý časopis pro fyziku. 5 pages

[8] Král, A., Liška, T., Virius, M.: *Počítače ve světě fyzikálních experimentů*, Prague, Czech Republic, 02/2005. In Tvorba software conference. 8 pages

[9] Liška, T.: *The Cluster and GRID Computing on Mass Data Production Systems for High Energy Physics Experiments*, Prague, Czech Republic, 06/2009. Disertation thesis. 95 pages