Prof. dr hab. inż. Tadeusz Czachórski
Instytut Informatyki, Wydz. AEiI Politechniki Śląskiej

Gliwice, 20.02.2014 r.

Recenzja pracy doktorskiej mgr inż. Michała Simona
„Fault Tolerant Data Acquisition through Dynamic Load Scheduling"
(„Odporna na błędy akwizycja danych poprzez dynamiczne szeregowanie obciążenia").

Kontekst pacy.

Praca dotyczy bardzo istotnego i obecnie intensywnie rozwijanego działu informatyki – przetwarzania bardzo dużej liczby danych, tzw. Big Data, co ma miejsce w różnych zastosowaniach, m.in. w meteorologii, biologii molekularnej (analiza genomu), tworzeniu map połączeń neuronalnych, ekologii, kryminologii, badaniach środowiska, astronomii (analiza zbieranych przez teleskopy danych), a także, jak w przypadku omawianej pracy, w analizie danych rejestrowanych w trakcie złożonych doświadczeń fizycznych, w szczególności w dziedzinie fizyki cząsteczkowej.

Praca proponuje ulepszenia bardzo rozbudowanego i wyspecjalizowanego systemu akwizycji danych wykorzystywanego przez detektor Compact Muon Solenoid (CMS) przy Wielkim Zderzaczu Hadronów (ang. Large Hadron Collider – LHC) w ośrodku badawczym CERN w Genewie. LHC jest obecnie największym akceleratorem kołowym na świecie (ma długość 27 km) skonstruowanym, aby przyśpieszać, a następnie zderzać dwie przeciwległe wiązki hadronów (protony lub ciężkie jony) z energią 14 TeV.

Celem CMS jest badanie kolizji, które zachodzą w akceleratorze LHC. Detektor składa się z wielu warstw detektorów odpowiedzialnych za równoczesne pomiary różnego rodzaju zjawisk oraz z olbrzymiego solenoidu, który dostarcza pola magnetycznego różnicującego tory powstałych w kolizji cząstek. Ponieważ prawdopodobieństwo, że w czasie kolizji nastąpią interesujące z punktu widzenia doświadczenia zjawiska (np. świadczące o istnieniu bozonu Higgsa) jest bardzo małe, trzeba wykonać ogromną liczbę kolizji i obserwować skutki każdej z nich. Wykonuje się to z częstotliwością kilkudziesięciu milionów kolizji na sekundę. Ponieważ brak wystarczającej pamięci do zapisu wszystkich zarejestrowanych w eksperymencie danych, konieczna jest ich filtracja, do czego służą dedykowane układy FPGA, tak że rejestruje się dane z częstotliwością ok. 1000 razy mniejszą. Dane przesyłane są siecią do farm komputerów filtrujących i rozdzielane pomiędzy nie za pomocą algorytmu round-robin; każda farma jest zorganizowana wokół pojedynczego switcha, a zrównoleglenie w farmie osiągnięte jest za pomocą techniki Single Process, Multiple Data. Odbiór danych jest wykonywany przez wyspecjalizowane węzły odbiorcze, z których każdy jest odpowiedzialny za odbiór danych opisujących kilka fragmentów kolizji oraz za wstępne złożenie tych fragmentów w większe elementy (super-fragmenty), które węzeł nadzorujący farmę przydziela proce-

som rekonstruującym kolizję. Następnie dane trafiają do jednego z procesów filtrujących, który odpowiada za selekcję interesujących kolizji na podstawie pełnej informacji. Procesy rekonstrukcyjny i filtrujący znajdują się na tym samym węźle obliczeniowym

Układ pracy

Praca obejmuje siedem rozdziałów oraz wykaz cytowanej literatury.

W pierwszym rozdziale autor krótko scharakteryzował problem, którego dotyczy praca – podje informacje o Wielkim Zderzaczu Hadronów i eksperymencie CMS, na potrzeby którego powstał projekt opisany w rozprawie doktorskiej. W kontekście tego eksperymentu zdefiniowano istotne dla projektu pojęcia szeregowania i równoważenia obciążenia, a następnie sformułowano cele badawcze oraz tezy pracy. Tezy pracy można przetłumaczyć następująco:

- *Dynamiczne szeregowanie obciążenia wpływa pozytywnie na niezawodność rozproszonego systemu akwizycji danych*
- *Asynchroniczne, rozproszone szeregowanie obciążenia może zostać przeprowadzone na horyzontalnie podzielnym, rozproszonym strumieniu danych pod warunkiem, że każdy z podstrumieni dostarcza dane w tej samej kolejności.*

W rozdziale drugim przedstawiono przegląd algorytmów szeregowania i równoważenia obciążenia; w szczególności dokonano przeglądu algorytmów szeregowania obciążenia zastosowanych w systemach akwizycji danych innych eksperymentów fizyki wysokich energii. Omówiono algorytmy, które zwiększają odporność systemu na awarie poprzez szeregowanie obciążenia i samo-stabilizację.

Trzeci rozdział pracy zawiera szczegółowy opis systemu akwizycji danych eksperymentu CMS. Omówiony został dwustopniowy sposób filtracji rejestrowanych danych dotyczących kolizji, oraz sposób, w jaki system dokonuje rekonstrukcji całości danych opisujących poszczególne kolizje. Dane opisujące pojedyncze kolizje hadronów są odczytywane z milionów kanałów detektora CMS przez system akwizycji danych. Ponieważ ilość napływających danych (ok. 38 TB na sekundę) przekracza możliwości ich dłuższej rejestracji a większość kolizji nie przynosi ciekawych wyników, mało interesujące dane są odfiltrowywane. Następnie kanały detektora są łączone i tworzą ok. 500 źródeł danych (nazywanych w pracy źródłami obciążenia dla rozważanego systemu filtracji danych), z których każde dostarcza część danych każdej z zarejestrowanych kolizji. Wszystkie fragmenty danych z zarejestrowanych kolizji są następnie przesyłane przy pomocy nieblokującej sieci Myrinet do farm komputerów filtrujących i scalających dane opisujące kolizje. Dotychczas stosowana statyczna metoda szeregowania obciążenia oparta jest na założeniu, że przepustowość farm komputerów filtrujących jest z góry znana przed rozpoczęciem procesu akwizycji danych i nie ulega zmianie w trakcie tego procesu. W przypadku wystąpienia awarii któregoś z komputerów tworzących farmy (ok. 1600 komputerów klasy PC, z czego ok. 600 jest krytycznych dla całego systemu, a ich awaria powoduje znaczny spadek przepustowości) następuje zduszenie częstotliwości akwizycji danych i w efekcie część danych nie jest rejestrowana i przepada. Tymczasem su-

maryczne możliwości czynnych w dalszym ciągu komputerów w zupełności wystarczyłyby do filtracji i rejestracji wszystkich napływających danych.

W rozdziale czwartym została przeprowadzona analiza wymagań stawianych przed projektowanym algorytmem szeregowania obciążenia. Przedstawiono również analizę przyczyn utraty danych w systemie akwizycji danych eksperymentu CMS w roku 2011, a także wyznaczono ilość danych, jaką można byłoby uratować przed utratą, gdyby zastosowana została zaproponowana w pracy metoda równoważenia obciążenia.

Rozdział piąty zawiera szczegółową analizę proponowanego algorytmu równoważenia obciążenia. Jako podstawę funkcjonowania tego algorytmu przyjęto prowadzenie bieżącego pomiaru przepustowości poszczególnych farm komputerów filtrujących i scalających dane opisujące kolizje. W tym celu wszystkie węzły dokonujące konkatenacji skorelowanych fragmentów danych wysyłają informacje na temat ich lokalnej wydajności do węzła nadzorującego daną farmę. Następnie węzły nadzorujące wymieniają się danymi, aby uzyskać redundancję danych pomiarowych oraz ustalają moment, w którym należy przekazać dane o wydajności farm (liczba zrekonstruowanych kolizji w jednostce czasu) do źródeł obciążenia. Każde ze źródeł obciążenia podejmuje decyzję o przydzieleniu fragmentu kolizji farmie obliczeniowej niezależnie od innych źródeł. Decyzja jest podejmowana na podstawie danych otrzymanych ze wszystkich farm z danego cyklu pomiarowego.

Szeregowanie obciążenia rozpoczyna się, gdy jedna z farm znajdzie się w stanie niedociążenia. Węzeł nadzorujący niedociążonej farmy komputerów filtrujących wysyła powiadomienie do pozostałych węzłów nadzorujących w celu uruchomienia procesu we wszystkich farmach jednocześnie. Dodatkowo każdy węzeł nadzorujący posiada maskę bitową określającą stan poprawnej pracy lub awarii każdej z farm komputerów filtrujących. Omawiane bitmaski są rozsyłane razem z danymi o obciążeniu, a następnie używane do wykluczenia niedziałających farm z procesu akwizycji danych. W momencie, gdy węzeł otrzymuje powiadomienie, wysyła on dane o obciążeniu swojej farmy, dane o obciążeniu farmy poprzednika w pierścieniu, oraz bitmaskę farm do źródeł obciążenia.

Przesłane dane są podstawą podejmowania decyzji przez źródła obciążenia o przydziale fragmentów danych opisujących kolizje (czyli obciążeniu) do poszczególnych farm. Każde ze źródeł obciążenia podejmuje decyzję przydziału obciążenia asynchronicznie i niezależnie od pozostałych źródeł na podstawie pomiarów obciążenia ze wszystkich farm z danego cyklu pomiarowego. Ponieważ wszystkie źródła obciążenia dostarczają dane dotyczące kolizji w tej samej kolejności oraz dysponują zestawem tych samych pomiarów, decyzja przydziału fragmentarycznych danych tej samej kolizji w każdym ze źródeł będzie taka sama (co zapewnia możliwość skompletowania wszystkich fragmentów danych opisujących kolizję w tym samym komputerze jednej z farm). Dane dotyczące poszczególnych kolizji przydzielane są przy użyciu algorytmu karuzelowego z pominięciem tych farm, do których wszystkie należne im dane zostały już wysłane (w pracy rozważano 3 różne implementacje algorytmu o różnych złożonościach obliczeniowych).

Należy podkreślić, ze opracowany algorytm szeregowania obciążenia działa w sposób asynchroniczny i rozproszony, co znaczy, że każde ze źródeł podejmuje decyzję o przydziale fragmentu danych opisujących kolizje do farmy obliczeniowej bez potrzeby komunikowania się z innymi źródłami, gdyż tego typu synchronizacja wprowadzałaby zbyt wielkie opóźnienia do systemu.

W szóstym rozdziale pracy przedstawiono wyniki badań eksperymentalnych, w których sprawdzano własności opracowanego algorytmu szeregowania obciążenia. W pierwszej kolejności sprawdzono, że wprowadzenie nowego algorytmu (i związanych z nim procesów przesyłania danych) nie wpłynęło negatywnie na dostępną przepustowości sieci nieblokującej Myrinet. Okazało się również, że wydajność procesu filtrowania i scalania fragmentów danych opisujących kolizje została nawet nieznacznie podwyższona. Wykonane eksperymenty potwierdziły również, że omawiany algorytm spełnia wymagania eksperymentu CMS. Po pierwsze wszystkie fragmenty danych dotyczące jednej kolizji zawsze trafiają do tej samej farmy obliczeniowej. Po drugie, przepustowość w pojedynczym węźle odbiorczym farmy obliczeniowej jest większa niż 200 MB/s (co jest warunkiem koniecznym do utrzymania częstotliwości akwizycji danych na poziomie 100 kHz).

Następnie przedstawiono serię przeprowadzonych eksperymentów, które potwierdziły, że system używający dynamicznego algorytmu szeregowania obciążenia jest znacznie bardziej odporny na awarie od dotychczas wykorzystywanego systemu. W kolejnych eksperymentach symulowano awarie węzłów obliczeniowych (w tym węzłów krytycznych dla działania poszczególnych farm obliczeniowych) jak i połączeń sieciowych, a następnie badano odpowiedź systemu na te awarie. W każdym z przeprowadzonych eksperymentów system używający zaproponowanego algorytmu działał znacznie lepiej od standardowego systemu. Dzięki zaproponowanemu algorytmowi przydział danych odbywał się proporcjonalnie do przepustowości każdej z farm i w ten sposób udało się uniknąć negatywnego wpływu uszkodzonych farm na farmy w pełni działające. Wykonane eksperymenty potwierdziły, że udało się oddzielić od siebie poszczególne farmy filtrujące, co oznacza, że krytyczna awaria (włączając w to awarie, które skutkują całkowitą utratą przepustowości) w jednej z farm filtrujących nie ma żadnego negatywnego wpływu na pozostałe farmy.

Należy podkreślić, ze badania eksperymentalne prowadzono zarówno w systemie testowym jaki i produkcyjnym CMS.

Rozdział siódmy zawiera podsumowanie wyników pracy i przedstawia najważniejsze wnioski.

Oryginalne wyniki pracy – można do nich zaliczyć :

- przeprowadzenia analizy funkcjonowania systemu akwizycji danych mierzonych w detektorze CMS i określenie słabych punktów tego systemu,
- opracowanie metody pomiaru bieżącego obciążenia i przepustowości farm filtrujących,

- opracowanie algorytmu dynamicznego szeregowanie obciążenia, wykorzystującego zaproponowaną metrykę obciążenia farm obliczeniowych,
- implementacja prototypu algorytmu szeregowania obciążenia, przetestowanie algorytmu w systemie testowym jak i produkcyjnym CMS.
- Przeprowadzenie badań potwierdzających, że algorytm spełnia wymagania eksperymentu CMS (przepustowość nie spada poniżej dopuszczalnej minimalnej wartości; wszystkie dane dotyczące tej samej kolizji trafiają zawsze do tej samej farmy)
- Przeprowadzenie eksperymentów oceniających niezawodność nowego systemu akwizycji danych w przypadku awarii węzłów obliczeniowych jak i połączeń sieciowych.

Autor pokazał, że zaproponowany algorytm szeregowania obciążenia zwiększa niezawodność systemu w przypadku awarii węzłów obliczeniowych lub połączeń sieciowych. Zmniejszenie przepustowości systemu jest na tyle niewielkie, że spełnia on dalej wymagania czasowe. Zaobserwowano wzrost wydajności (ok. 3,5%) systemu mierzony liczbą zrekonstruowanych opisów kolizji w jednostce czasu, spowodowany bardziej wydajnym zarządzaniem zasobami. W przypadku wystąpienia awarii (niezależnie czy dotyczy ona połączenia sieciowego czy węzła obliczeniowego) system używający zaproponowanego algorytmu wykazuje się znacząco większą przepustowością. Wykorzystano asynchroniczną, rozproszoną metodę szeregowania obciążenia, a więc każde ze źródeł obciążenia podejmuje decyzję o przydziale opisów kolizji do farmy obliczeniowej bez potrzeby komunikowania się z innymi źródłami. Tezy pracy zostały w ten sposób potwierdzone.

Jak zauważa Autor, słabym punktem zaproponowanej metody okazał się sposób komunikowania się węzłów EVM między sobą. Jak wynika z przeprowadzonych eksperymentów, mimo że farmy komputerów filtrujących są identycznie skonstruowane, nigdy nie osiągają stanu niedociążenia dokładnie w tym samym momencie. Z tego powodu rozważana w pracy metoda komunikacji okazała się zawsze sekwencyjna i mogłaby zostać przyśpieszona poprzez zrównoleglenie (można tu np. zaimplementować mechanizm *multicast*), a także zoptymalizować procedurę decyzyjną.

Praca jest napisana przejrzyście, jej struktura jest dobrze dostosowana do założonych celów, język nie budzi zastrzeżeń. Wyniki prac zostały opublikowane – bibliografia zawiera 3 prace Doktoranta: w materiałach (i) *25th IEEE International Parallel & Distributed Processing Symposium, Anchorage,* (ii) *Journal of Physics: Conference Series,* (iii) *Communications in Computer and Information Science.*

Inne uwagi.

Autor świetnie zna system, którym się zajmuje, a proponowane ulepszenia sprawdza doświadczalnie w sposób metodyczny, obserwując działanie systemu. Z praktycznego punktu widzenia jest to najlepsze rozwiązanie. Jednakże praca zyskałaby na ogólności, a jej rezultaty mogłyby być szerzej wykorzystane, gdyby problem dynamicznego szeregowania i rozdziału obciążenia przeanalizować w sposób bardziej ogólny i formalny, za pomocą modeli matematycznych lub symulacyjnych (symulacja zdarzeń dyskretnych). Literatura dotycząca szeregowania zadań obliczeniowych czy równoważenia obciążenia jest wyjątkowo bogata. Istnieją też

odpowiednie narzędzia analityczne i programowe, w czym Autor dobrze się orientuje. Chętnie poznałbym poglądy Doktoranta na ten temat w czasie obrony pracy.

Podsumowanie:

Uważam, że rozprawa doktorska mgr inż. Michała Simona w pełni spełnia warunki stawiane rozprawom doktorskim przez ustawę o stopniu i tytułach naukowych. Autor wykazał się bardzo dobra znajomością problemów szybkiego przetwarzania danych w bardzo wyspecjalizowanym systemie, stanowiącym jedno z najbardziej zaawansowanych w świecie rozwiązań tego typu, potrafił zaproponować oryginalne i konstruktywne rozwiązania. Wnioskuję o przyjęcie tej pracy jako rozprawy doktorskiej i dopuszczenie jej do publicznej obrony.