Nowoczesne techniki programowania heterogenicznego na przykładzie wyznaczania śladów przelotu cząstek

Dr Grzegorz Korcyl

1. Motywacja

Nowoczesne systemy komputerowe projektowane są z uwzględnieniem układów obliczeniowych o różnych architekturach. Każda klasa procesorów posiada wady i zalety. Układy ogólnego zastosowania CPU są uniwersalne i łatwe do programowania co jest okupione kosztem wydajności. Procesory graficzne GPGPU posiadają architekturę zoptymalizowaną do prowadzenia obliczeń wektorowych i strumieniowych. Jednak aby w pełni je wykorzystać, należy starannie zorganizować dostęp do danych oraz sekwencję instrukcji. Potężna moc obliczeniowa tych układów, często nie w pełni wykorzystywana, pociąga za sobą znaczne zużycie energii. Bezpośrednio programowalne macierze bramek FPGA, pozwalają na strumieniowe przetwarzanie danych w czasie rzeczywistym co jednak wymaga od programisty dogłębnej znajomości tej technologii.

Większość problemów obliczeniowych składa się z szeregu zadań, z których każde może wykorzystać unikalne cechy danego rodzaju procesora i tym samym wykonać się w sposób bardziej optymalny. Jednocześnie, systemy komputerowe składające się z większej ilości różnych rodzajów układów stają się coraz bardziej złożone, znacznie podnosząc poziom skomplikowania rozwijania oprogramowania.

Przykładem zaawansowanych systemów obliczeniowych są systemy przetwarzania danych z eksperymentów fizycznych dużej skali. Znaczna ilość danych produkowanych przez detektory wymaga przetworzenia w locie w celu selekcji najbardziej istotnych fragmentów. Selekcja jest wynikiem działania wielu różnych algorytmów, na różnego rodzaju danych z poszczególnych elementów detektora. Jest to zatem miejsce gdzie heterogeniczna infrastruktura obliczeniowa doskonale znajduje swoje zastosowanie.

Rozwój elektroniki musi w związku z tym iść w parze z rozwojem środowisk programistycznych pozwalających na wykorzystanie możliwości systemów heterogenicznych. Do niedawna, każdy rodzaj układów obliczeniowych posiadał swój własny zestaw metod oraz języków programowania, kompilatorów i narzędzi. Obecnie powstają inicjatywy takie jaki OneAPI od Intel czy SYCL od Khronos Group. Są to wysokopoziomowe modele programowania unifikujące metody oraz dyrektywy dla kompilatorów, dzięki którym ten sam kod źródłowy może być w sposób optymalny kompilowany i wykonywany na układach o różnych architekturach.

Badania prowadzone nad różnymi architekturami obliczeniowymi, metodami programistycznymi oraz optymalizacją oprogramowania pod kątem sprzętu jest fundamentalną aktywnością w ramach obszaru badawczego DigiWorld, w którym każda domena badawcza może czerpać korzyści z ich rezultatów. W szczególności, prace realizowane w ramach tego projektu wpisują się w domenę „Zaawansowane metody obliczeniowe i sztuczna inteligencja" poprzez wytworzenie nowych technik obliczeniowych oraz w domenę „AI w naukach ścisłych i przyrodniczych" poprzez akcelerację sieci neuronowych będących jednym z algorytmów używanych do wyznaczania śladów przelotu cząstek.

2. Cel projektu

Celem projektu jest zbadanie możliwości heterogenicznych systemów komputerowych oraz wysokopoziomowych środowisk programistycznych na przykładzie algorytmów wyznaczania śladów przelotu cząstek w eksperymentach fizycznych dużej skali.

Detektory śledzące są jednymi z ważniejszych elementów eksperymentu PANDA budowanego w ośrodku FAIR w Niemczech. Ich skalę podkreśla fakt, że w trakcie działania będą produkować w sumie prawie 30 GB danych na sekundę. Dane te wymagają natychmiastowej obróbki w celu odrzuceniu nieistotnych danych oraz ekstrakcji śladów przelotu cząstek. Ta informacja pozwoli na realizację kolejnych algorytmów, których celem jest rekonstrukcja pełnego zdarzenia fizycznego, które zostało zarejestrowane przez cały system detekcyjny.

W tym celu projektowany jest rozbudowany system komputerowy składający się z wysokowydajnych serwerów o heterogenicznej architekturze. Zadaniem systemu jest odbieranie surowych danych z detektorów, wykonanie sekwencji algorytmów analizujących oraz podjęcie decyzji o przesłaniu danych do magazynowania. Ponieważ dane z detektorów spływają w trybie ciągłym, wymagane jest aby przetwarzanie odbywało się w sposób strumieniowy. W związku z tym, architektura systemu oraz oprogramowanie musi zostać w pełni zoptymalizowane, tak aby być w stanie obliczyć rezultaty zanim nowa porcja danych zostanie odebrana.

W ramach tego projektu, przeprowadzone zostaną badania nad dwoma algorytmami wyznaczania śladów przelotu cząstek. Jeden algorytm bazuje na transformacie Hougha [1] natomiast drugi wykorzystuje do tego celu sieci neuronowe [2]. Oba algorytmy zostaną zaimplementowane przy użyciu wysokopoziomowego środowiska programistycznego SYCL, oraz następnie przeprowadzone zostaną badania nad sposobem ich wykonywania na 3 głównych architekturach układów obliczeniowych: CPU, GPGPU oraz FPGA.

Zebrane w ten sposób doświadczenie oraz wiedza pozwoli skutecznie zaprojektować oraz zrealizować system przetwarzania danych dla eksperymentu PANDA. Polska jest jednym z głównych członków konsorcjum FAIR, natomiast Instytucją Koordynującą jest Uniwersytet Jagielloński. W szczególności, Wydział Fizyki, Astronomii i Informatyki Stosowanej posiada wieloletni, aktywny udział w budowie eksperymentu PANDA poprzez zaprojektowanie oraz budowę detektorów śledzących oraz opracowanie systemu akwizycji i przetwarzania danych [3,4]. Realizacja przedstawionego projektu będzie kolejnym, bardzo istotnym wkładem naszego Wydziału w ten eksperyment.

Badania opisane w ramach tego projektu, będą realizowane przy współpracy z głównymi podmiotami odpowiedzialnymi za system przetwarzania danych w eksperymencie PANDA, to jest: instytut GSI Darmstadt oraz Forschungszentrum Juelich w Niemczech. Opracowane rozwiązania wejdą w pakiet oprogramowania FairRoot i wspomogą procesy symulacji oraz analizy danych prowadzone przez wiele grup badawczych wchodzących w skład kolaboracji PANDA.

3. Organizacja pracy

Projekt będzie realizowany w ramach aktywności wirtualnego laboratorium Hardware Acceleration Lab. Utworzenie laboratorium przyniosło natychmiastowe korzyści takie jak: skoordynowanie aktywności osób pracujących w podobnej domenie, nawiązanie współpracy z Akademickim Centrum Komputerowym Cyfronet AGH i uzyskanie dostępu do infrastruktury obliczeniowej oraz współdzielenie dostępnego sprzętu i licencji na oprogramowanie.

Prace w ramach projektu zostaną podzielone pomiędzy dwóch wykonawców, z których każdy podejmie się implementacji oraz badań jednej wersji algorytmu. Pierwszym krokiem będzie przeniesienie algorytmu do środowiska SYCL a następnie przeprowadzenie optymalizacji pod kątem każdej z 3 głównych architektur obliczeniowych. Na każdym etapie będą przeprowadzane testy sprzętowe oraz zbierane będą pomiary wydajności. Pozwoli to w sprawny sposób zredagować raport oraz publikacje naukowe.

Budżet projektu zostanie w całości przeznaczony na wynagrodzenia dla wykonawców. Poniżej przedstawiono harmonogram oraz budżet projektu rozpisany na 6 miesięcy:

Miesiąc	**Zadania**	**Budżet [PLN]**
1	Przygotowanie stanowiska do pracy Implementacja podstawowej wersji algorytmu	7 000
2-5	Optymalizacja i ewaluacja algorytmu na CPU Optymalizacja i ewaluacja algorytmu na GPGPU Optymalizacja i ewaluacja algorytmu na FPGA	28 000
6	Sporządzenie raportu z badań oraz redagowanie publikacji	7 000
		42 000

4. Oczekiwane rezultaty

Skuteczna realizacja projektu zaowocuje zestawem wypracowanych metod i technik rozwijania oprogramowania na platformy heterogeniczne. W szczególności, wyniki będą stanowiły istotny wkład w projekt pełnego systemu przetwarzania danych dla eksperymentu PANDA.

Dwa opracowane w ramach badań algorytmy, będą stanowiły przykłady, na podstawie których zademonstrowane zostaną opracowane techniki. Kompleksowy przegląd ich implementacji na różnego rodzajach architektur sprzętowych, stanowić będzie materiał na interesujące publikacje naukowe.

Mierzalnymi wskaźnikami realizacji projektu będą:

- 2 implementacje algorytmu śledzenia cząstek udostępnione na publicznych repozytoriach kodu źródłowego

- 2 kompletne raporty opisujące szczegóły implementacyjne, zastosowane techniki oraz wyniki przeprowadzonych pomiarów

- 2 publikacje naukowe przygotowane do wysłania do anglojęzycznego czasopisma o wysokim współczynniku cytowalności i zasięgu międzynarodowym

Pośrednim efektem przeprowadzonych prac, będzie wytworzenie się grupy badawczej w ramach Hardware Acceleration Lab, zwiększenie udziału Uniwersytetu Jagiellońskiego w międzynarodowym projekcie FAIR i zacieśnienie współpracy z międzynarodowymi grupami badawczymi wchodzącymi w jego skład.

Badania przeprowadzone w ramach projektu oraz zawiązana współpraca będą również stanowić podstawę do aplikacji o grant naukowy dedykowany projektowi oraz realizacji pełnego systemu przetwarzania danych dla eksperymentu PANDA.

[1] M. J. Galuska, et al., “Hough transform based pattern recognition for the PANDA Forward Tracking System”, PoS Bormio 2013 023

[2] W. Esmail, T. Stockmanns, J. Ritman, “Machine Learning for Track Finding at PANDA”, arXiv: 1910.07191

[3] G. Korcyl, et al., „Readout Electronics and Data Acquisition for Gaseous Tracking Detectors”, IEEE Transactions on Nuclear Science, vol. 65, 2018

[4] G. Korcyl, et al. PANDA Collaboration, “Technical Design Report for the PANDA Forward Tracker”