Kamil Figiela

Streszczenie pracy magisterskiej pt.
„Optimization of Resource Allocation on the Cloud"

W obecnych czasach praktycznie wszystkie dziedziny nauki wymagają infrastruktury obliczeniowej do przeprowadzania symulacji, planowania eksperymentów lub analizy ich wyników. Chmury obliczeniowe są nowym modelem dostarczania zasobów, który zdobywa coraz większą popularność w środowisku naukowym ze względu na to, że zasoby są łatwo dostępne na żądanie i rozliczane w modelu 'pay-per-use'. Problem alokacji zasobów dla obliczeń naukowych był przedmiotem badań dla infrastruktur takich jak klastry i systemy gridowe. Specyfika chmur obliczeniowych nie pozwala jednak na bezpośrednie zastosowanie znanych już metod rozwiązywania tego problemu, stąd konieczność poszukiwania nowych rozwiązań.

W ramach pracy magisterskiej rozwiązujemy problem alokacji zasobów w chmurze obliczeniowej. Jest to jedna z pierwszych prób rozwiązywania problemu alokacji zadań w chmurze za pomocą całkowitoliczbowego programowania nieliniowego (mixed integer non-linear programming). W celu rozwiązania tego problemu został zdefiniowany model chmury obliczeniowej oraz modele dwóch rodzajów aplikacji: zbioru zadań (bag of tasks) oraz grafu zadań (workflow). Zdefinowano także optymalizowaną funkcji celu oraz przeprowadzono testy. Modele definiowane są w języku AMPL, który pozwala na zastosowanie najlepszych solverów optymalizacyjnych takich jak CPLEX, czy Cbc.

Model infrastruktury zakłada wiele publicznych oraz prywatnych chmur obliczeniowych udostępniających wiele typów instancji wirtualnych maszyn, różniących się cenami oraz wydajnością. Celem optymalizacji jest minimalizacja całkowitego kosztu wykonania obliczeń. Opracowane modele optymalizacji uwzględniają wiele założeń pomijanych w istniejących rozwiązaniach, takich jak godzinowe rozliczanie zużycia zasobów, koszty transferu danych, czy też to, że w praktyce zasoby udostępniane przez chmury są skończone. Dzięki temu udało się stworzyć model opisujący dość dokładnie charakterystykę chmury obliczeniowej typu Infrastructure-as-a-Serice (IaaS).

Przyjęte założeia dotyczące modelu chmury obliczeniowej, jak również samych modeli aplikacji naukowych, powodują, że proponowany w pracy model optymalizacji wraz z użytymi narzędziami pozwala uzyskać wyniki w akceptowalnym czasie pozwalającym na jego wdrożenie w praktyce.

Proponowane w pracy rozwiązanie zostało przetestowane na danych opisujących rzeczywistą infrastrukturę obliczeniową (np. Amazon EC2, RackSpace) na podstaie publicznie dostępnych benchmarków pochodzących z CloudHarmony. Parametry aplikacji zostały zaczerpnięte z rzeczywistych aplikacji z Pegasus Workflow Gallery, takich jak Montage, SIPHT czy CyberShake, które rozwiązują rzeczywiste problemy naukowe z dziedzin takich jak astronomia, czy sejsmologia.

Przeprowadzenie testów dla różnych parametrów (rozmiaru zadań, ograniczeń czasowych) pozwoliło na sprawdzenie stabilności rozwiązania oraz zaobserwowanie ciekawych zależności ceny i czasu obliczeń. Otrzymane rezultaty przedstawiają typowe problemy związane z podejmowaniem decyzji o tym jak planować wdrażanie aplikacji naukowych na chmurę obliczeniową oraz jak te problemy mogą być rozwiązywane przy użyciu technik optymalizacyjnych. Okazuje się, że w miarę skracania limitów czasu wykonania zadań rośnie koszt ich wykonania. Przyrost ten nie jest liniowy, a jego charakterystyka zmienia się w pewnych punktach, których wykrycie jest istotne. Zbiory rozwiązań generowane przez model optymalizacyjny przedstawiony mogą być wykorzystywane zarówno przez użytkownika jak i sprzedawcę usług chmury obliczeniowej. W przypadku użytkownika, rezultaty pozwalają na wybranie optymalnego pod względem czasu i kosztu obliczeń planu alokacji zasobów. Te same dane mogą również zostać wykorzystane przez sprzedawcę usług chmurowych do maksymalizacji jego zysku.

W przyszłości autor planuje przeprowadzenie eksperymentów na rzeczywistej infrastrukturze chmury obliczeniowej. Pozwoli to na lepsze zrozumienie efektów zachodzących przy uruchamianiu aplikacji naukowych oraz obserwacji tego jak dynamizm środowiska wpływa na faktyczny przebieg obliczeń. Ponadto model infrastruktury zostanie rozszerzony o usługi dostępne w modelu SaaS oraz wsparcie dla chmur rozliczanych w krótszych okresach (np. 5 minutowych). Usprawnienia obejmą również metody optymalizacyjne: eksperymenty z innymi solverami oraz zastosowanie innych technik modelowania.

W wyniku prowadzonych badań powstały dwie publikacje, dołączone w formie załączników do pracy magisterskiej, stanowią one przyczynek do badań nad alokacją zasobów w chmurach obliczeniowych.

Badania były prowadzone we współpracy z University of Notre Dame, USA oraz University of Southern California, USA.