

Wydajne algorytmy rozproszone i równoległe dla dużych i dynamicznych danych

Wszechobecność technologii informatycznych skutkuje powstawaniem ogromnych ilości danych składowanych w wielkich centrach obliczeniowych lub rozproszonych w węzłach rozległych sieci. Ograniczenia technologiczne i fizyczne uniemożliwiają lub utrudniają zebranie powstających danych w jednym miejscu dla szybkiego ich przetwarzania. Ograniczenia czasowe wymagają przetwarzania takich danych w sposób zrównoleglony, polegający na podziale zadań obliczeniowych na podzadania wykonywane jednocześnie na wielu jednostkach obliczeniowych. Choć problemy te pojawiają się od lat, ostatnio zmieniały się wiodące rozwiązania technologiczne dla obliczeń na masowych danych (np. koncepcje Map Reduce, Spark) i dla infrastruktury komunikacyjnej do wymiany gromadzonych danych (np. Internet of Everything, 5G). Celem projektu jest budowanie podstaw algorytmicznych dla architektur obliczeniowych i komunikacyjnych wykorzystywanych do przesyłania i przetwarzania takich dużych ilości danych. Obejmuje to zarówno budowanie algorytmów dla konkretnych kluczowych problemów obliczeniowych jak i identyfikowanie strukturalnych cech badanych środowisk obliczeniowych.

Rozwiązania dla obliczeń na zebranych dużych danych tworzyć będziemy konstruując algorytmy i identyfikując trudności w modelu masowych obliczeń równoległych (MPC, ang. Massive Parallel Computing) i blisko powiązanego z nim modelu obliczeń rozproszonych Congested Clique (CC). Modele MPC i CC odzwierciedlają najistotniejsze cechy architektur obliczeniowych w nowoczesnych centrach danych, np. MapReduce, Spark czy Hadoop. Choć istnieją już rozwiązania dla niektórych problemów obliczeniowych, brakuje wydajnych rozwiązań dla wiele kluczowych problemów a istniejące rozwiązania rzadko są dopasowane do danych dynamicznych i sytuacji, w których pamięć lokalna jest istotnie ograniczona względem rozmiaru danych. Wykorzystując zdobyte doświadczenia zamierzamy konstruować nowe wydajne rozwiązania w takich scenariuszach obliczeniowych.

Podstawowym studiowanym od lat modelem algorytmicznym do budowania rozwiązań na potrzeby komunikacji bezprzewodowej na poziomie lokalnym jest kanał wielodostępowy (MAC od ang. multiple access channel), w którym wielu uczestników próbuje przesyłać wiadomości w obrębie jednego kanału komunikacyjnego, gdzie wysłana wiadomość może być przesłana tylko wtedy, gdy żaden inny nadawca nie próbuje przesyłać wiadomości w tym samym czasie. Choć badanie MAC i algorytmów dla niego ma długą historię (np. analizy protokołów backoff), większość algorytmicznych analiz nie uwzględnia istotnych dla współczesnych zastosowań aspektów, np. zależności między urządzeniami, ograniczonej mocy obliczeniowej i innych zasobów miniaturowych sensorów czy braku lub ograniczeń dostępu takich urządzeń do możliwości ładowania baterii. W projekcie będziemy rozwijać techniki uwzględniające te aspekty, budując rozwiązania algorytmiczne i podstawy dla rozwiązań podstawowych problemów komunikacyjnych np. planowaniem komunikacji w sieci z dynamicznie napływającymi pakietami, czy konstrukcją struktur kombinatorycznych do cyklicznego stosowania. Uogólnienia MAC są wykorzystywane do rozwiązywania problemów z innych obszarów, np. grupowego testowania. Wykorzystując własne doświadczenia w konstrukcji algorytmów testowanie grupowego zamierzamy budować rozwiązania uwzględniające wymagania zastosowań np. we współczesnym testowaniu medycznym.

Opisane powyżej aspekty gromadzenia i przetwarzania dużych danych współdzielą wiele technik algorytmów i obliczeń rozproszonych i równoległych. Naturalnym rozszerzeniem tych modeli są sieci rozległe, w których węzły wykonujące obliczenia połączone są w rozległe skomplikowane sieci połączeń. Rozległość sieci powoduje jednak, że tylko niektóre pary węzłów połączone są bezpośrednio. W naszych badaniach zajmować będziemy się również takimi rozszerzeniami, z naciskiem na specyfikę zastosowań w gromadzeniu danych przez ograniczone obliczeniowo sensory i przetwarzaniu danych w ilościach uniemożliwiających szybkie uzyskanie wyników bez korzystania z równoległości obliczeń.

Efektom naszych badań będą nowe rozwiązania algorytmiczne dla badanych problemów a także identyfikacja ograniczeń modeli i problemów trudnych obliczeniowo. Spodziewanym efektem badań jest uzyskanie nowych szybszych rozwiązań, lepiej dopasowanych do współczesnych środowisk obliczeniowych i komunikacyjnych. Choć nasze badania mają charakter podstawowy, celem będzie uzyskania rozwiązań dających potencjał do praktycznych zastosowań, dzięki uwzględnieniu w modelu kluczowych cech współczesnych środowisk obliczeniowych i komunikacyjnych .