Marek MIKA[1], Bartosz BELTER[2], Jan WĘGLARZ[1,2]
[1] Politechnika Poznańska, [2] Poznańskie Centrum Superkomputerowo Sieciowe

PROBLEM SZEREGOWANIA ZADAŃ TRANSMISJI DANYCH WEDŁUG KRYTERIUM ENERGETYCZNEGO

Streszczenie. Oszczędność energii zużywanej przez infrastrukturę obliczeniową w szczególności w dużych centrach danych jest w ostatnich latach jednym z ważniejszych zagadnień badawczych i praktycznych zarazem. W pracy przedstawiamy model problemu, w którym minimalizowane jest zużycie energii przez wyposażenie sieciowe wirtualnego centrum danych. Ze względów praktycznych omawiany model jest dyskretną wersją problemu.

PROBLEM OF DATA TRANSMISSION TASKS SCHEDULING ACCORDING TO THE ENERGETIC CRITERION

Summary. Saving energy consumed by computing infrastructure, especially in large data centers, has been one of the most important research and practical issues in recent years. We present a problem model in which the energy consumption of the virtual data center network equipment is minimized. For practical reasons, this model is a discrete version of the problem..

1. Wprowadzenie

Historia ludzkości zawiera wiele przykładów na to, że człowiek od pradawnych czasów starał się łączyć różne lokalizacje geograficzne w pewnym ściśle określonym celu. Wystarczy wspomnieć słynne szlaki handlowe Bursztynowy i Jedwabny. rzymskie akwedukty transportujące wodę z odległych źródeł, czy też trakty i drogi ułatwiające przemieszczanie się i transport różnorakich dóbr.

Rosnące potrzeby sprawiły, że z czasem zaczęto rozbudowywać istniejące połączenia, a postęp technologiczny doprowadził do tego, że starsze rozwiązania zastąpiono nowszymi efektywniejszymi. Stare szlaki handlowe zastąpiono transportem kolejowym, samochodowym lub samolotowym, akwedukty ustąpiły miejsca rozbudowanym sieciom wodociągowym, a szczególnie intensywnie uczęszczane drogi przybrały kształt autostrad. Nowe wynalazki i odpowiednio skonstruowana dla nich sieć połączeń umożliwiły przesyłanie na duże odległości energii elektrycznej, gazu, ropy naftowej, czy też informacji. Ludzie mogą komunikować się między sobą za pośrednictwem łącz telegraficznych, telefonicznych i sieci komputerowych.

Jednak nadal mamy do czynienia z coraz większą liczbą użytkowników oraz rosnącymi wymaganiami. Można temu sprostać na wiele sposobów, między innymi

przez budowę nowych połączeń lub zastosowanie nowych bardziej efektywnych technologii. Można także wykorzystać istniejącą już infrastrukturę przez zwiększenie przepustowości wybranych połączeń lub też optymalizację ruchu w takiej sieci. Właśnie temu ostatniemu podejściu poświęcona jest ta praca. Przedstawiamy w niej problem szeregowania zadań występujących w sieci połączeń na przykładzie transmisji danych w sieci komputerowej wchodzącej w skład wirtualnego centrum danych. Jako kryterium optymalizacji przyjęto jedno z istotniejszych w ostatnim czasie zagadnień, jakim jest minimalizacja zużytej energii.

Chociaż w pracy rozpatrujemy konkretny problem praktyczny, to jednak przedstawione w niej podejście może być, po pewnych przeróbkach adaptacyjnych, zastosowane z powodzeniem również do innych problemów, gdzie istotnym elementem jest sieć połączeń, a celem jest uszeregowanie zadań, które mają być wykonane z wykorzystaniem tej sieciowej infrastruktury.

2. Wirtualne centrum danych

Rozproszone systemy komputerowe, obecne w informatyce od kilku dekad, stopniowo ewoluowały, doprowadzając między innymi do powstania koncepcji chmury obliczeniowej. Znalazło to swoje odzwierciedlenie między innymi w postaci nowego wyposażenia i usług, które odmieniły pracę użytkowników końcowych. W popularnym obecnie modelu przetwarzania, dane użytkownika przechowywane są w dużych centrach danych, które oferują również dostęp do dużych mocy obliczeniowych dla uprawnionych do tego podmiotów. Postęp technologiczny w zakresie technologii sieciowych umożliwił łączenie centrów danych, użytkowników i różnorodnych urządzeń na odpowiednim poziomie wydajności i jakości, dając szansę na tworzenie tzw. wirtualnych centrów danych [2]. Najprościej rzecz ujmując jest to rozwinięcie koncepcji klasycznego centrum danych, w którym pojedyncze centra danych łączy się za pomocą bardzo wydajnych sieci komputerowych. Stosując odpowiednie techniki wirtualizacji można wirtualne zasoby obliczeniowe połączyć za pośrednictwem wirtualnych łącz. Rosnące zapotrzebowanie na usługi w modelu oferowanym przez centra danych prowadzi do wzrostu ilości używanego sprzętu, a przez to do zwiększonego zużycia energii elektrycznej. W sukurs właścicielom i operatorom sieciowym przychodzą tu nowoczesne rozwiązania technologiczne, takie jak programowalna sieć komputerowa (SDN – Software Defined Networking) i protokół OpenFlow implementowane w wielu nowych elementach wyposażenia sieciowego. Umożliwiają one administratorom i/lub właścicielom zasobów sieciowych stosowanie własnych reguł kierowania ruchem sieciowym [1]. Jeżeli jednym z istotnych kryteriów kierowania takim ruchem jest oszczędność energii, to dobrym sposobem sterowania siecią jest możliwość dynamicznego wyłączania i włączania poszczególnych elementów infrastruktury sieciowej.

3. Sformułowanie problemu

Na rozważany w tej pracy problem szeregowania zadań transmisji danych można spojrzeć z co najmniej kilku punktów widzenia. Jednym z najczęściej rozważanych w literaturze jest punkt widzenia właściciela wszystkich zasobów zarówno sieciowych,

jak i obliczeniowych. W tym wypadku podmiot odpowiedzialny za szeregowanie zadań ma pełną wiedzę na temat dostępnych zasobów oraz zadań do wykonania. Jednakże model ten nieczęsto występuje w praktyce, ponieważ rzadko kiedy zasoby obliczeniowe i sieciowe mają tego samego właściciela/operatora. Innym, praktycznie nie stosowanym, aczkolwiek prawdopodobnym w niedalekiej przyszłości, jest model z punktu widzenia właściciela zadań. Obecnie właściciel zadań nie ma zbyt dużego wpływu na oszczędność energii zużytej do przetwarzania zgłoszonych przez niego zadań. Jednak, łatwo można sobie wyobrazić sytuację podobną do tej, z jaką mamy obecnie do czynienia na rynku mediów, gdzie użytkownik może wybrać operatora, oferującego najkorzystniejszą dla niego wersję dostarczania energii elektrycznej, gazu, wody. Możliwe, że za kilka lat zgłaszając do wykonania zbiór zadań obliczeniowych, będziemy mogli porównać oferty kilku konkurencyjnych operatorów i wybrać tę najtańszą, gdzie cena będzie zależna między innymi od przewidywanego zużycia energii elektrycznej. Z jeszcze innym modelem mamy do czynienia, gdy spojrzymy na problem z punktu widzenia właściciela zasobów sieciowych. W tym wypadku właścicielem zasobów obliczeniowych jest inny podmiot, ale obydwa podmioty współpracują ze sobą w celu dostarczenia usług o odpowiedniej jakości. Właściciel zasobów obliczeniowych odpowiedzialny jest za wykonanie zadań obliczeniowych, a właściciel zasobów sieciowych za wykonanie zadań transmisji danych. Każdy z nich kieruje się własną polityką szeregowania zadań i przydzielania im zasobów. Jednak żaden z nich nie może robić tego nie uwzględniając drugiej strony i drugiego rodzaju zadań. To, w jakiej kolejności przystąpią oni do szeregowania zadań i jak będzie wyglądała interakcja między nimi, w dużej mierze zależy od wzajemnych ustaleń obu stron.

W rozważanym modelu założono, że szeregowanie odbywa się w dwóch fazach. W pierwszej kolejności odbywa się szeregowanie zadań obliczeniowych przy założeniu, że zadania transmisji danych są wykonywane z szybkością co najmniej równą wartości wcześniej uzgodnionej pomiędzy właścicielem zadań i właścicielami zasobów, a następnie przeprowadzane jest szeregowanie zadań transmisji danych w oparciu o niektóre dane będące wynikiem szeregowania zadań obliczeniowych. Kolejność ta podyktowana jest praktycznymi obserwacjami, z których wynika, że zaledwie 20-30 % energii konsumowanej przez całą infrastrukturę centrum danych przypada na wyposażenie sieciowe, a reszta zużywana jest przez zasoby obliczeniowe. Jeżeli kryterium stosowanym przez obydwa podmioty jest oszczędność energii, to zasadnym wydaje się udzielenie pierwszeństwa tej stronie, która dzięki optymalizacji może osiągnąć większe korzyści. Operator/właściciel zasobów obliczeniowych zatem jako pierwszy szereguje zadania obliczeniowe zakładając przy tym, że rozmiar zadań transmisji danych jest znany i będzie można je wykonać z szybkością nie niższą niż uzgodniona. W wyniku szeregowania zadań obliczeniowych otrzymujemy informację o tym kiedy i na jakich zasobach obliczeniowych będą one wykonane. Dane te są jednocześnie parametrami dla szeregowania zadań transmisji danych. Znane są bowiem miejsca wykonania kolejnych zadań obliczeniowych, które będą stanowiły punkty źródłowy i docelowy transmisji danych, oraz terminy rozpoczęcia i zakończenia tych zadań, definiujących okna czasowe dla poszczególnych zadań transmisji. Dzięki temu właściciel zasobów sieciowych, który nie ma wpływu na szeregowanie zadań obliczeniowych, może dostosować własne procedury

szeregowania do typu informacji, jakie otrzymuje od właściciela zasobów obliczeniowych. W rezultacie powstanie uszeregowanie zadań transmisji, które będzie uwzględniało łącza wchodzące w skład ścieżki pomiędzy węzłami źródłowym i docelowym każdego z zadań transmisji, szybkość transmisji oraz terminy rozpoczęcia i//lub zakończenia poszczególnych zadań. Zauważmy, że ze względu na przyjęte kryterium szeregowania, nie jest wskazane, by algorytmy szeregujące działały zbyt długo, ponieważ zużycie energii związane z ich wykonaniem mogłoby przewyższyć oszczędności wynikające z ich zastosowania. Nie przewiduje się zatem dalszej interakcji pomiędzy właścicielami zasobów po fazie szeregowania zadań transmisji danych. Z tego samego powodu stosowanie algorytmów szeregowania zadań ma sens jedynie dla zadań polegających na transmisji dużych wolumenów danych, takich jakie spotykane są między innymi w aplikacjach typu workflow [3]. Aplikacje takie składają się ze zbioru wzajemnie powiązanych zadań obliczeniowych, które najczęściej polegają na przetwarzaniu zbiorów danych. Zadania te muszą być wykonane w odpowiedniej kolejności wynikającej ze struktury aplikacji, a pomiędzy parą zadań obliczeniowych występujących bezpośrednio po sobie najczęściej dochodzi do transmisji danych, które pojawiają się na wyjściu jednego zadania i stanowiących dane wejściowe dla kolejnego zadnia.

Sieć komputerowa składa się z węzłów, odpowiadających na ogół lokalizacji poszczególnych centrów danych, a od strony wyposażenia sieciowego składających się z tzw. bazowego elementu węzła, w skład którego wchodzą karty wyposażone w interfejsy. Interfejsy stanowią punkty przyłączenia łącz optycznych, za pośrednictwem których łączą się one ze swoimi odpowiednikami w innych węzłach sieciowych. Każdy z wymienionych elementów wyposażenia sieciowego podczas pracy, czyli w trybie aktywnym zużywa pewną ilość energii. Natomiast zużycie energii po przejściu w tryb nieaktywny jest pomijalnie małe. Interfejs jest w stanie aktywnym, jeśli za jego pośrednictwem odbywa się transmisja danych. W przeciwnym wypadku przechodzi on w tryb nieaktywny. Zużycie energii w kartach składa się z niezależnej od wielkości ruchu wartości stałej oraz wartości zmiennej proporcjonalnej do wielkości ruchu na interfejsach karty. Przejście karty w tryb nieaktywny następuje wyłącznie wtedy, gdy żaden interfejs tej karty nie bierze udziału w transmisji danych. Bazowy element węzła sieciowego może być w trybie nieaktywnym wyłącznie wtedy, gdy wszystkie karty do niego należące również są w trybie nieaktywnym. W przeciwnym wypadku jest on włączony, a jego zużycie energii jest stałe i niezależne od wielkości ruchu sieciowego. Z pewnym uproszczeniem można przyjąć, że łącza światłowodowe zużywają tyle energii, ile konsumują wzmacniacze optyczne rozmieszczone na takich łączach co 80 km.

Zakłada się, że zadania transmisji danych są niepodzielne, co oznacza, że muszą być one wykonane bez jakichkolwiek przerw przy użyciu tych samych zasobów. Innymi słowy, jeśli pomiędzy węzłem źródłowym i docelowym zostanie zestawiona ścieżka dla tego zadania, to cała transmisja z tym związana musi odbyć się tą ścieżką. Zadania te są również nieskalowalne, czyli wykonywane są w całości z tą samą szybkością i nie może ona ulec zmianie. Szybkość transmisji wyznaczana jest w trakcie szeregowania. Jednak musi się ona mieścić w przedziale, którego granice wyznaczone są przez minimalną szybkość, ustaloną pomiędzy stronami, oraz jej maksymalną wartość, określoną przez najwolniejsze łącze występujące na rozważanej

ścieżce pomiędzy węzłem źródłowym a docelowym. Założono również, że opóźnienia transmisji można pominąć, ponieważ są dużo mniejsze niż czas samej transmisji danych.

W rozważanym problemie zbiór K zawierający n zadań transmisji danych należy wykonać przy użyciu zasobów tworzących sieć komputerową łączącą lokalizacje, w których znajdują się zasoby obliczeniowe pewnego wirtualnego centrum danych tak, aby zminimalizować energię zużytą podczas wykonywania tych zadań. Topologia sieci reprezentowana jest przez nieskierowany multigraf $G = (V, E)$. Zbiór V wierzchołków tego grafu składa się z dwóch rozdzielnych podzbiorów V_1 oraz V_2, reprezentujących odpowiednio zbiór bazowych elementów węzła oraz zbiór kart wyposażonych w interfejsy sieciowe. Pomiędzy parą wierzchołków $i, j \in V$ może wystąpić więcej niż jedna krawędź $(i,j)_l \in E$, gdzie $l = 1,2, \ldots$, co odpowiada często spotykanej sytuacji, gdy pomiędzy dwoma węzłami występują co najmniej dwa różne łącza fizyczne. Ponadto, zbiór E jest również podzielony na dwa rozłączne podzbiory $E_1 = \{(i,j)_l : i, j \in V_2\}$ oraz $E_2 = \{(i,j)_l : i \in V_1, j \in V_2\}$. Podzbiór E_1 zawiera krawędzie odpowiadające fizycznym łączom pomiędzy węzłami obliczeniowymi. Natomiast podzbiór E_2 reprezentuje połączenia pomiędzy kartami a bazowymi elementami węzła. Każda krawędź $(i,j)\_l \in E_2$ oznacza, że karta j należy do bazowego elementu węzła i. Dla każdego łącza fizycznego reprezentowanego przez krawędź $(i,j)_l \in E_1$ określone są dwa parametry: szybkość F_{ij}^l oznaczająca maksymalny współczynnik transmisji przypadającej na jednostkę czasu oraz jego długość L_{ij}^l. Każde zadanie transmisji danych $k \in K$ opisane jest następującymi parametrami: a_k – termin gotowości, d_k – linia krytyczna, i_k – węzeł źródłowy ($i_k \in V_2$), j_k – węzeł docelowy ($j_k \in V_2$), D_k – rozmiar pliku danych, który ma być przesłany między węzłami i_k oraz j_k. Obydwa parametry czasowe a_k i d_k definiują okno czasowe $[a_k, d_k]$ w którym musi być wykonane zadanie $k \in K$.

Celem jest znalezienie optymalnego uszeregowania, które minimalizuje ilość energii zużytej podczas wykonywania całego zbioru K zadań transmisji danych. Szeregowanie w tym wypadku polega więc na: i) znalezieniu zasobowo i czasowo dopuszczalnych przydziałów zasobów do zadań, a następnie ii) wyznaczeniu terminów rozpoczęcia tych zadań. Dany przydział zasobów dla zadania $k \in K$ jest dopuszczalny zasobowo, o ile przydzielone łącza sieciowe tworzą ścieżkę pomiędzy węzłami i_k oraz j_k, i dopuszczalny czasowo, jeśli dostępna część szybkości łącz tworzących tę ścieżkę umożliwia wykonanie w całości zadania $k \in K$ w jego oknie czasowym $[a_k, d_k]$.

3. Dyskretny model matematyczny

Jak wspomniano w poprzednim rozdziale, szeregowanie zadań transmisji danych polega na znalezieniu dla każdego zadania przydziału zasobów oraz terminów rozpoczęcia i/lub zakończenia, które są zarówno czasowo, jak i zasobowo dopuszczalne oraz optymalizują przyjęte kryterium szeregowania. Przydział zasobów oznacza w tym wypadku wyznaczenie ścieżki pomiędzy węzłami źródłowym i docelowym oraz określenie części szybkości łącz tworzących tę ścieżkę, która zostanie przeznaczona do wykonania zadania transmisji danych. Każdy z wymienionych tu parametrów, zarówno termin rozpoczęcia, jak i zakończenia oraz

część prędkości łącza przydzielona do realizacji zadania, ma charakter ciągły. Jednak w tym rozdziale przedstawimy model matematyczny rozważanego problemu, w którym powyższe parametry zostały zdyskretyzowane. Oczywiście, model ten nie gwarantuje znalezienia rozwiązania dokładnego. Jednak stanowi on dobre jego przybliżenie i umożliwia zastosowanie komercyjnych solwerów matematycznych lub też adaptację znanych i efektywnych algorytmów.

W prezentowanym modelu występują następujące zmienne decyzyjne:

x_{kt} – zmienna binarna przyjmująca wartość 1, jeżeli zadanie k rozpoczyna się w chwili t ($t \in [a_k, d_k]$);

y_{ij}^{lk} – zmienna binarna przyjmująca wartość 1, wtedy i tylko wtedy, gdy łącze $(i,j)_l$ między węzłami i oraz j jest użyte do transmisji danych zadania k, tj. stanowi część ścieżki (i_k, j_k) przydzielonej do przeprowadzenia transmisji danych między węzłami i_k oraz j_k;

z_k – nieujemna zmienna całkowita określająca współczynnik szybkość transmisji zadania k.

Model matematyczny problemu wygląda następująco:

Zminimalizować:

$$E = E_N + E_{BF} + E_{BV} + E_L \tag{1}$$

przy ograniczeniach:

$$\sum_{t=a_k}^{d_k} x_{kt} = 1 \qquad k \in K \tag{2}$$

$$\sum_{t=a_k}^{d_k} x_{kt} \cdot t \geq a_k \qquad k \in K \tag{3}$$

$$\sum_{t=a_k}^{d_k} x_{kt} \cdot t + D_k \cdot \frac{q}{z_k} \leq d_k \qquad k \in K \tag{4}$$

$$\sum_{j}\sum_{l} y_{ij}^{lk} - \sum_{j}\sum_{l} y_{ji}^{lk} = \begin{cases} 1 & \text{jeśli } i = i_k \\ -1 & \text{jeśli } i = j_k \\ 0 & \text{wpp} \end{cases} \qquad i,j \in V \tag{5}$$

$$\sum_{k=1}^{n} \sum_{b=t}^{\min\left\{d_k, t+D_k \cdot \frac{q}{z_k}\right\}} \frac{z_k}{q} \cdot y_{ij}^{lk} \cdot x_{kb} \leq F_{ij}^{l} \qquad \begin{matrix} (i,j)_l \in E_2 \\ t = \left[0, \max\limits_{k \in K}\{d_k\}\right] \end{matrix} \tag{6}$$

$$x_{kt} \in \{0,1\} \qquad k \in K, t \in [a_k, d_k] \tag{7}$$

$$y_{ij}^{lk} \in \{0,1\} \qquad k \in K, (i,j)_l \in E_2 \tag{8}$$

$$z_k \geq 0 \qquad k \in K \tag{9}$$

Ograniczenia (2) zapewniają, że każde zadanie zostanie uszeregowane dokładnie raz. Nierówności (3) i (4) dotyczą okien czasowych zdefiniowanych dla każdego z zadań. Ograniczenia (3) gwarantują, że żadne z zadań nie rozpocznie się przed swoim terminem gotowości, a (4), że zakończy się przed linią krytyczną. Zakłada się, że zarówno termin gotowości a_k, jak i linia krytyczna d_k są wartościami dyskretnymi. Jeżeli tak nie jest to a_k jest zaokrąglane do najbliższej wartości dyskretnej większej od a_k, a d_k do najbliższej wartości dyskretnej mniejszej od d_k. Występująca w (4) wartość q jest wynikiem dyskretyzacji szybkości łącz i oznacza czas potrzebny do przesłania określonej (wynikającej z dyskretyzacji) liczby bajtów. Kolejna grupa ograniczeń (5) zapewnia, że dla każdego zadania zostanie zestawiona dokładnie jedna ścieżka pomiędzy węzłami źródłowym i docelowym. Nierówności (6) gwarantują, że dla każdego pojedynczego łącza przydzielonego do wykonania zadań ze zbioru K nie zostanie przekroczony jego limit szybkości. Ograniczenia (7) i (8) definiują binarny charakter zmiennych x_{kt} oraz y_{ij}^{lk}. Natomiast nierówności (9) zapewniają nieujemne wartości dla zmiennych z_k.

Funkcja celu (1) stanowi sumę energii zużytej przez bazowe elementy węzłów (E_N), karty (E_{BF}), interfejsy (E_{BV}) i wzmacniacze na łączach optycznych (E_L), gdy są one w stanie aktywnym. Wzory dla poszczególnych składowych funkcji celu są dość rozbudowane i wobec tego nie będziemy ich tutaj przedstawiać. Jednak, poza jednym wyjątkiem (E_{BV}), stanowią one sumaryczną wartość energii zużytej przez poszczególne elementy wyposażenia sieciowego, gdzie energia zużyta przez pojedyncze urządzenie wyrażona jest iloczynem mocy tego urządzenia i czasu jego aktywności. Energia zużyta przez interfejsy E_{BV}, czyli zmienny składnik energii zużytej przez karty, dla pojedynczego interfejsu jest iloczynem rozmiarów plików przesłanych przez ten interfejs oraz tak zwanego współczynnika mocy interfejsu przypadającej na jednostkę danych i jednostkę czasu

Powyższy model w zależności od stopnia dyskretyzacji daje różne korzyści. Duże wartości parametru q sprawiają, że maleje liczba dopuszczalnych wartości dla zmiennych z_k, co jest oczywiście korzystne, gdy istotny jest czas obliczeń. Niestety, ma to również wpływ na dokładność otrzymanych wyników, ponieważ mniejsze wartości q lepiej przybliżają rozwiązanie dokładne, ale wymagają dłuższego czasu obliczeń. Podobnie jest z parametrami czasowymi, im dłuższa jednostka zdyskretyzowanego czasu, tym mniej zmiennych x_{kt} występuje w problemie, czyli skraca się czas obliczeń. Jednak zbyt duże wartości prowadzą do mniejszej dokładności rozwiązania. Zauważmy, że dyskretyzacja czasu może doprowadzić również do zawężenia okien czasowych $[a_k, d_k]$. W skrajnym przypadku może to spowodować, że niemożliwe będzie znalezienie rozwiązania dopuszczalnego, mimo że dla problemu niezdyskretyzowanego takie rozwiązanie istnieje. Odrębną kwestią jest znalezienie granicy opłacalności stosowania zaprezentowanego modelu. Dokładne lecz czasochłonne obliczenia mogą wymagać tyle energii, że oszczędności, jakie można osiągnąć na poziomie sterowania ruchem w sieci komputerowej, staną się niezauważalne. Z drugiej strony obliczenia nawet niedokładne, a trwające krótko, mogą dać przybliżenie rozwiązania dokładnego na tyle dobre, że zastosowanie go w praktyce da istotne oszczędności w zużyciu energii. Jednakże właściwy dobór poziomu dyskretyzacji wybranych parametrów ciągłych na ogół musi się dokonać

drogą eksperymentów, w których uwzględni się strukturę sieci wirtualnego centrum danych oraz charakter realizowanych w nim zadań.

4. Podsumowanie

Rozwój systemów komputerowych z jednej strony sprawia, że dostępna jest coraz większa moc obliczeniowa, a z drugiej strony prowadzi do coraz większego zapotrzebowania na energię elektryczną. Z roku na rok pojawia się więcej informacji na temat tego, że przy obecnie stosowanej technologii półprzewodnikowej granica możliwości energetycznych zostanie wkrótce osiągnięta. Już teraz jednym z głównych ograniczeń, o jakich wspomina się przy okazji planowania systemów komputerowych w tzw. exaskali, jest dostarczenie do takiego systemu energii o wystarczającej mocy. Oszczędność energii staje się zatem jednym z kluczowych celów projektantów i operatorów systemów komputerowych dużej skali. Oszczędności te można przede wszystkim osiągnąć w infrastrukturze obliczeniowej i na tym skupia się większość prac, ale można je osiągnąć również w infrastrukturze sieciowej. W przypadku już istniejących sieci zarządzanie energią można osiągnąć wprowadzając do niej odpowiednie urządzenia, które pozwalają na stosowanie własnych reguł sterowania ruchem sieciowym (np. urządzenia wykorzystujące koncepcje SDN i OpenFlow). Opracowując własne reguły sterowania ruchem w sieci warto zwrócić uwagę na możliwości oszczędzania energii. W pracy przedstawiono dyskretny model matematyczny opracowany dla wirtualnego centrum danych. Model ten właściwie zastosowany, przy odpowiedniej dyskretyzacji wybranych parametrów ciągłych problemu, może dać istotne korzyści. Skrócenie czasu obliczeń, nawet kosztem dokładności rozwiązania w kontekście zużycia energii może doprowadzić do istotnych oszczędności. Warto zauważyć też, że prezentowany model można zaadaptować również dla innych problemów, gdzie zadania jako jeden z zasobów traktują połączenia pomiędzy różnymi lokalizacjami.

LITERATURA

1. Celenlioglu M.R., Goger S.B., Mantar H.A.: An SDN-based energy-aware routing model for intra-domain networks, Proceedings 22nd International Conference on Computer Networks, 2014, p. 61-66.
2. Faizul Bari M., Boutaba R., Esteves R., Granvilley L., Podlesny M., Rabbani M, Zhang Q., Zhani M.: Data Center Network Virtualization: A Survey, IEEE Communications Surveys & Tutorials, 2013, Vol. 15, p. 909-928.
3. Mika M., Waligóra G., Węglarz J.: Modelling and solving grid resource allocation problem with network resources for workflow applications, Journal of Scheduling, 2011, Vol. 14, p. 291-306.