Piotr Brudło

INTERNETOWE SYSTEMY PRZETWARZANIA ROZPROSZONEGO TYPU GRID W ZASTOSOWANIACH BIZNESOWYCH

STRESZCZENIE

Skoncentrowano się na możliwościach wykorzystania oraz integracji rozproszonych mocy obliczeniowych komputerów internautów w globalnej sieci www. Zaprezentowano paradygmaty sieciowego przetwarzania typu grid computing oraz volunteer computing. Podkreślono istotność tego typu przetwarzania w zagadnieniach wymagających bardzo dużych mocy obliczeniowych. Zaprezentowano przykłady rozwiązań systemowych tego typu: system BOINC, będący modelowym reprezentacyjnym systemem referencyjnym w tym zakresie, oraz system Comcute zrealizowany na Wydziale Elektroniki, Telekomunikacji i Informatyki Politechniki Gdańskiej. Pokazano również inne systemy z obszaru rozważanych paradygmatów przetwarzania. Podkreślono znaczenie tych systemów poprzez pokazanie zagadnień, które są aktualnie w ich ramach przetwarzane i rozwiązywane. Zwrócono uwagę na możliwości wykorzystania tego typu podejścia dla rozwiązywania problemów związanych ze złożonymi procesami oraz zagadnieniami o charakterze ekonomicznym i biznesowym.
Słowa kluczowe: BOINC, Comcute, grid computing, volunteer computing

Wstęp

Obecnie do globalnej sieci Internet podłączony jest praktycznie każdy komputer, urządzenie mobilne oraz cały wachlarz urządzeń wspomagających. Stąd też addytywnie występująca sumaryczna moc obliczeniowa całości Internetu jest naprawdę imponująca. Co więcej, większość mocy obliczeniowej Internetu jest zupełnie niewykorzystana. Dla przykładu można podać komputery typowych

internautów, gdzie większość „aktywności" polega głównie na włączeniu komputera oraz ewentualnie przeglądaniu stron internetowych, odbiorze i wysyłaniu poczty, komunikatorach, portalach społecznościowych, używaniu edytora tekstu itp., co w ujęciu statystycznym w bardzo niewielkim stopniu obciąża zasoby komputera (procesor, wykorzystanie pamięci, pamięci dyskowe). Stąd też istnieje potencjalna możliwość „zagospodarowania" (integracji) tych ogromnych zasobów – oczywiście przy zgodzie lub przynajmniej braku sprzeciwu właścicieli tych zasobów.

W wyniku istnienia tego typu potrzeb oraz potencjalnych praktycznych możliwości powstało szereg paradygmatów wykorzystania rozproszonych zasobów obliczeniowych komputerów internautów. Obecnie konstruuje się systemy oparte na podejścia typu grid computing oraz volunteer computing[1].

W sensie definicyjnym przetwarzanie typu grid computing jest sieciową generacją przetwarzania rozproszonego, polegającą na połączeniu dużej liczby heterogenicznych jednostek komputerowych (węzłów) w celu stworzenia koncepcyjnie jednorodnego komputera wirtualnego o bardzo dużej mocy obliczeniowej. Celem jest stworzenie rozproszonego wirtualnego komputera z ogromnej liczby połączonych, niejednorodnych systemów posiadających różnego rodzaju zasoby. Aktualnie standardy dla współdzielenia tych zasobów, wraz z dostępnością większych szerokości pasm transmisyjnych, stanowią szansę technologiczną na dynamiczny rozwój w systemach gridowych. W sensie praktycznym grid computing oznacza masowe przetwarzanie traktowane jako usługa użyteczności ogólnej. Dla klienta nie jest istotne, gdzie są przechowywane jego dane, ani który komputer wykonuje zlecenie. Natomiast koncepcja grid computing widziana od strony usługodawcy oznacza alokację zasobów, współużytkowanie informacji oraz zapewnienie wysokiej dostępności[2].

Volunteer computing jest to rodzaj obliczeń rozproszonych, w którym użytkownicy komputera (zazwyczaj świadomie) udostępniają swoje zasoby obliczeniowe (czas pracy procesora, pamięć, miejsce na dysku itp.) do wykonania określonych zadań obliczeniowych – obecnie zazwyczaj w ramach projektów badawczych lub naukowych. Woluntariuszami są przede wszystkim internauci dobrowolnie zgadzający się na obliczeniowe wykorzystanie ich komputerów. W przetwarzaniu typu volunteer computing można zidentyfikować szereg charakterystycznych aspektów. Woluntariusze są zwykle anonimowi. Można co prawda wymagać rejestracji oraz formalnego przystąpienia do jakiegoś projektu, ale identyfikacja taka ma znaczenie tylko w ramach określonego przetwarzania. Z powodu anonimowości oraz paradygmatu tego podejścia woluntariusz nie jest w żaden sposób dyscyplinowany przez całość systemu, udostępnia swoje zasoby na zasadzie pełnej dobrowolności. Ponadto woluntariusze muszą mieć zapewniony wysoki stopień własnego bezpieczeństwa, co oznacza, że zadania

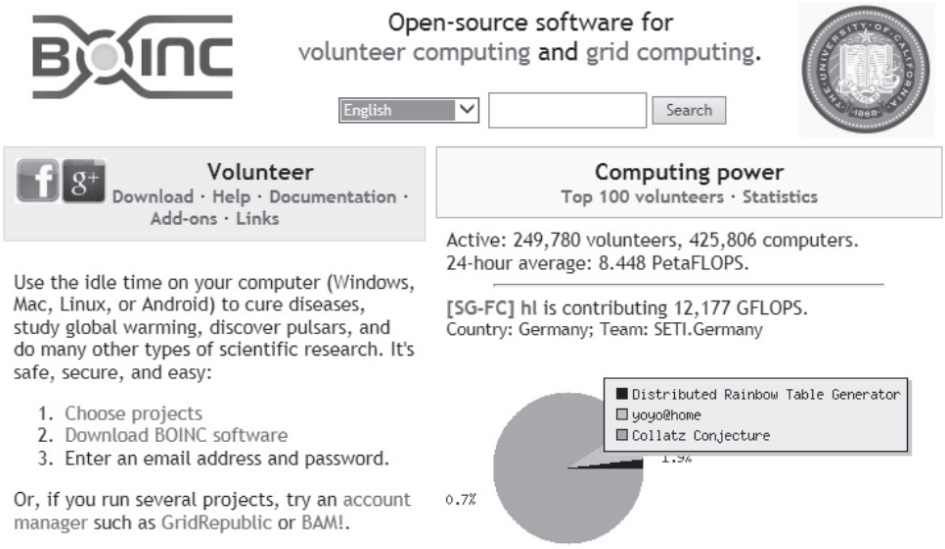
[1] J. Balicki, H. Krawczyk, E. Nawarecki (red.), *Grid and Volunteer Computing*, Wydawnictwo Politechniki Gdańskiej, wyd. I, Gdańsk 2012, s. 33–36, 53–55.

[2] J. Balicki, J. Kuchta (red.), *Obliczenia rozproszone w systemach komputerowych o architekturze klasy grid*, Wydawnictwo Politechniki Gdańskiej, Gdańsk 2012, s. 39–41, 46–48.

do przetworzenia nie mogą wyrządzać im krzywdy (niszczenie zasobów, wyłudzanie informacji, wirusy, trojan itd.). W praktyce realizowane jest to przez budowanie zaufania pomiędzy stroną systemową a internautami. W przypadku realizacji dużych projektów (np. naukowych lub badawczych) strona zlecająca przetwarzanie u woluntariuszy sprawdza, publikuje, weryfikuje oraz gwarantuje zakres oraz rodzaj przetwarzania, jak również stopień wykorzystania zasobów u internautów. Ponadto strona zlecająca dokłada staranności w zakresie ogólnie pojętego bezpieczeństwa sieciowego. Technologicznie współczesne systemy typu volunteer computing są zazwyczaj oparte na systemach gridowych.

1. System przetwarzania rozproszonego BOINC

BOINC (*Berkeley Open Infrastructure for Network Computing*)[3] to niekomercyjne rozwiązanie z dziedziny obliczeń rozproszonych, które pierwotnie powstało na potrzeby projektu SETI@home[4], aktualnie wykorzystywane jest również w projektach innych niż SETI. Jest to niekomercyjne oprogramowanie pośredniczące pozwalające na udział komputera zwykłego użytkownika w naukowych projektach. BOINC jest rozwijany na Uniwersytecie Kalifornijskim w Berkeley, jest wolnym i otwartym oprogramowaniem wydawanym na licencji GNU LGPL i jest wspierany finansowo przez amerykańską rządową agencję naukową National Science Foundation (rys. 1).



Rysunek 1. BOINC – Berkeley Open Infrastructure for Network Computing

Źródło: BOINC, University of California, National Science Foundation, boinc.berkeley.edu (dostęp: 11.2014).

[3] BOINC – Berkeley Open Infrastructure for Network Computing, University of California, National Science Foundation, boinc.berkeley.edu (dostęp: 11.2014).

[4] SETI@home – Search for Extraterrestrial Intelligence, University of California, National Science Foundation, BOINC Project, setiathome.berkeley.edu (dostęp: 11.2014).

Oprogramowanie BOINC dzieli się na oprogramowanie pracujące po stronie serwera projektu oraz na oprogramowanie uruchamiane przez wolontariuszy na swoich komputerach. Do najważniejszych aplikacji pracujących po stronie serwera należy scheduler (serwer harmonogramów). Zajmuje się on dystrybucją fragmentów danych do obliczeń pomiędzy komputery uczestników projektu. W swoim działaniu scheduler uwzględnia między innymi możliwości komputerów uczestników (moc obliczeniowa, ilość pamięci RAM) oraz średni czas w ciągu doby, jaki komputery te przeznaczają na pracę z BOINC. W ten sposób unika się nadmiernego obciążenia słabych komputerów oraz pozwala się na pełniejsze wykorzystanie mocnych maszyn. Jeżeli na komputerze otrzymującym dane do przetwarzania nie została jeszcze zainstalowana aplikacja mająca je przetwarzać, jest ona również przesyłana do uczestnika projektu.

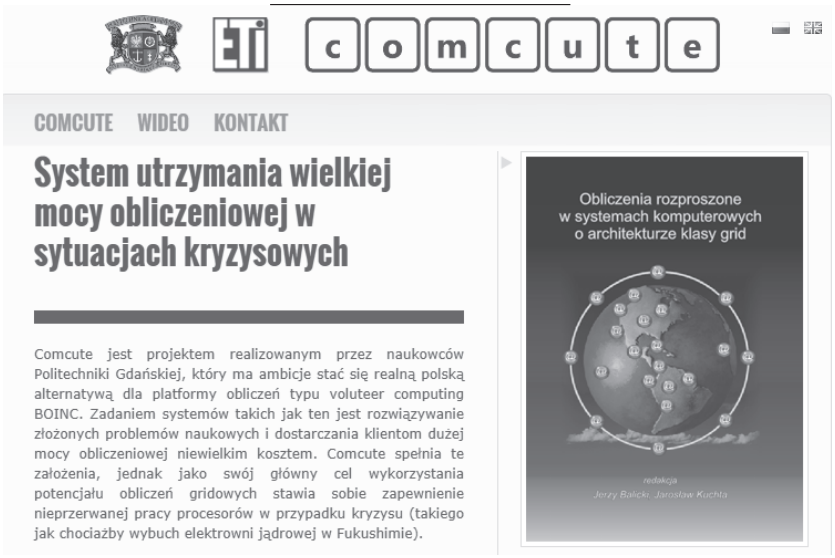
Gdy na komputerze uczestnika znajdą się zarówno dane do przetwarzania, jak i odpowiednia aplikacja, rozpoczyna się przetwarzanie danych. Czas przetwarzania jednej porcji danych jest różny w zależności od projektu i waha się od kilkunastu sekund do wielu godzin. Dzięki okresowemu zapisywaniu wykonanej pracy obliczenia nie muszą odbywać się w jednym nieprzerwanym ciągu, lecz mogą być zawieszane, gdy zachodzi potrzeba przeznaczenia mocy obliczeniowej na inne zadania (np. lokalne zadania internauty) lub po prostu wyłączenia komputera.

BOINC w sposób automatyczny ściąga i uruchamia na komputerze wolontariusza aplikację projektu, co może budzić uzasadniony niepokój o możliwość uruchomienia złośliwego oprogramowania. Aby temu zapobiec, BOINC korzysta z cyfrowego podpisywania aplikacji, aby nie dopuścić do „podstawienia" fałszywej aplikacji w miejsce oryginalnej. Użytkownicy ze swojej strony powinni jednak zwrócić uwagę, czy projekty, do których chcą się podłączyć, nie są podejrzane, a w razie podejrzeń warto poszukać opinii na forach dyskusyjnych innych projektów.

2. Comcute – system utrzymania wielkiej mocy obliczeniowej w sytuacjach kryzysowych

Comcute[5] jest projektem realizowanym przez naukowców Politechniki Gdańskiej, który ma ambicje stać się realną polską alternatywą dla platformy obliczeń typu volunteer computing BOINC (rys. 2). Zadaniem takich systemów jest rozwiązywanie złożonych problemów naukowych i dostarczania klientom dużej mocy obliczeniowej niewielkim kosztem. Comcute spełnia te założenia, jednak jako swój główny cel dla wykorzystania potencjału obliczeń gridowych stawia sobie zapewnienie nieprzerwanej pracy procesorów w przypadku kryzysu (takiego jak chociażby wybuch elektrowni jądrowej w Fukushimie).

[5] Comcute – System utrzymania wielkiej mocy obliczeniowej w sytuacjach kryzysowych, Politechnika Gdańska, Wydział ETI, comcute.eti.pg.gda.pl (dostęp: 11.2014).



Rysunek 2. Comcute – System utrzymania wielkiej mocy obliczeniowej w sytuacjach kryzysowych

Źródło: Comcute, Politechnika Gdańska, Wydział ETI, comcute.eti.pg.gda.pl (dostęp: 11.2014).

System bazuje na filozofii integracji mocy obliczeniowych komputerów internautów – zwanych wolontariuszami – dobrowolnie udostępniających zasoby swoich maszyn na rzecz całości projektu. Typowy scenariusz w takim środowisku wygląda tak, że każda osoba z dostępem do Internetu może bezpłatnie i w dowolnym momencie przystąpić do badań przez stronę internetową projektu. Tematyka tych badań może być merytorycznie różna, a wybór danego zagadnienia do przetwarzania zależy wyłącznie od internauty, gdyż ten, decydując się na partycypację w badaniach, udostępnia moc obliczeniową swojego komputera i jego pamięć dla konkretnego, jawnie określonego, projektu (lub kilku projektów). Zalety takiego podejścia są oczywiste: setki tysięcy ludzi na całym świecie mogą, nie wychodząc z domu oraz małym nakładem pracy, uczestniczyć w rozwiązywaniu największych zagadnień naukowych, które trapią ludzkość od stuleci.

Comcute niesie ze sobą nieograniczoną liczbę możliwości. W praktyce tylko wyobraźnia ogranicza twórców w tym, co mogą osiągnąć dzięki tej platformie. Symulacje powodzi, modelowanie pogody czy symulacja zachowań tłumu w panice, to tylko przykłady zagadnień, które dzięki przetwarzaniu rozproszonemu na platformie staną się możliwe do zrealizowania. Już na początku istnienia Comcute zaprojektowano oprogramowanie, którego celem jest symulacja pożarów w lasach. Opisane wieloma parametrami (takimi jak prędkość i kierunek wiatru, kaloryczności podłoża i drzew oraz wilgotność powietrza) wirtualne przestrzenie poddawane są regularnym podpaleniom, a uzyskane wyniki mogą przynieść ogromne korzyści, np.: pomóc lepiej chronić tereny narażone na podpalenie oraz przed rozprzestrzenianiem się ognia. Ośrodki naukowe, wojsko, firmy polegające na niezawodności obliczeń, a nawet banki – wszyscy oni mogą czerpać korzyści z platformy Comcute i wykorzystywać jej ogromny potencjał do swoich celów przy stosunkowo niskich kosztach.

W systemie Comcute nie trzeba instalować specjalnego dedykowanego oprogramowania – Comcute operuje całkowicie w przeglądarce internetowej.

3. Inne systemy przetwarzania gridowego w sieci

Obecnie jest wiele uruchomionych oraz dostępnych systemów tego typu. Dla potrzeb niniejszego przeglądu wybrano dwa: Polską Infrastrukturę Gridową PL-Grid[6] oraz system GPUGrid[7], reprezentujący typ przetwarzania volunteer computing w zastosowaniach biomedycznych.

3.1. PL-Grid – Polska Infrastruktura Gridowa

Polska Infrastruktura Gridowa została zbudowana w ramach projektu PL-Grid w latach 2009–2013 (rys. 3). Powstała w celu dostarczenia polskiej społeczności naukowej platformy informatycznej opartej na klastrach komputerów, służących e-Science w różnych dziedzinach. Infrastruktura wspiera badania naukowe poprzez integrację danych doświadczalnych i wyników zaawansowanych symulacji komputerowych prowadzonych przez geograficznie rozproszone zespoły. Infrastruktura PL-Grid umożliwia polskim naukowcom prowadzenie badań naukowych na podstawie symulacji i obliczeń dużej skali z wykorzystaniem klastrów komputerów oraz zapewnia wygodny dostęp do rozproszonych zasobów komputerowych. Stworzenie infrastruktury PL-Grid nie tylko rozszerzyło ilość zasobów obliczeniowych dostarczanych polskiej społeczności naukowej o ponad

Rysunek 3. PL-Grid – Polska infrastruktura dla przetwarzania gridowego

Źródło: PL-Grid, Polska Infrastruktura Gridowa, www.plgrid.pl (dostęp: 11.2014).

[6] PL-Grid – Polska Infrastruktura Gridowa, www.plgrid.pl, (dostęp: 11.2014).

[7] GPUGrid – volunteer computing for biomedicine, www.gpugrid.net, (dostęp: 11.2014).

230 TFlopów mocy obliczeniowej i ponad 3,6 PBajtów przestrzeni dyskowej, ale – co ważniejsze – ułatwiło efektywne wykorzystanie tych zasobów poprzez dostarczenie użytkownikom innowacyjnych usług gridowych i narzędzi wraz z ciągłym wsparciem technicznym.

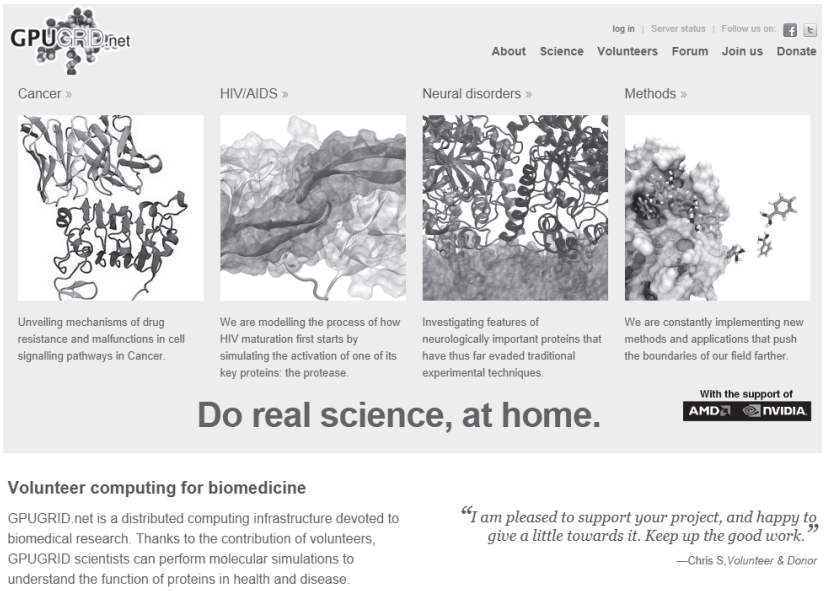
Polska Infrastruktura Gridowa jest częścią paneuropejskiej infrastruktury budowanej w ramach EGI (*European Grid Initiative*), której celem jest integracja narodowych infrastruktur gridowych w jedną trwałą infrastrukturę produkcyjną. PL-Grid jest zarządzana przez Konsorcjum PL-Grid, utworzone w styczniu 2007 roku, w skład którego wchodzą: Akademickie Centrum Komputerowe Cyfronet AGH w Krakowie, Interdyscyplinarne Centrum Modelowania Matematycznego i Komputerowego w Warszawie, Instytut Chemii Bioorganicznej PAN – Poznańskie Centrum Superkomputerowo-Sieciowe w Poznaniu, Centrum Informatyczne Trójmiejskiej Akademickiej Sieci Komputerowej w Gdańsku, Wrocławskie Centrum Sieciowo-Superkomputerowe we Wrocławiu.

Naukowcy i zespoły badawcze z Polski, zainteresowane wykorzystaniem infrastruktury PL-Grid do obliczeń i symulacji wielkiej skali, mogą korzystać nieodpłatnie z szeregu usług, narzędzi oraz pakietów, między innymi z: dostępu do klastrów o dużej mocy obliczeniowej i dużych pamięci dyskowych oraz pakietów specjalistycznych. Umożliwia to przeprowadzanie obliczeń naukowych z dziedziny biologii, chemii kwantowej, fizyki, obliczeń numerycznych i symulacji, przy wykorzystaniu zaawansowanych narzędzi do organizacji eksperymentów obliczeniowych. Ponadto udzielana jest pomoc przy uruchamianiu specjalistycznych pakietów naukowych na rozległych zasobach obliczeniowych oraz wsparcie (przez szkolenia i system helpdesk) przy projektowaniu własnych aplikacji naukowych przeznaczonych do obliczeń na infrastrukturze PL-Grid.

Infrastruktura sieciowa PL-Grid, utworzona w projekcie PL-Grid, jest obecnie dynamicznie rozbudowywana w ramach projektu PLGrid Plus (projekt na lata 2012–2015). W przyszłości planowana jest również kontynuacja oraz dalszy rozwój tego projektu.

3.2. GPUGrid – volunteer computing w zastosowaniach biomedycznych

Jest to rozproszona infrastruktura obliczeniowa przeznaczona i zorientowana na badania o charakterze biomedycznym. Projekt jest realizowany na Uniwersytecie w Barcelonie i reprezentuje typowy przykład sieciowego systemu rozproszonego w paradygmacie volunteer computing (rys. 4). GPUGrid integruje w jedną strukturę obliczeniową karty graficzne komputerów, tworząc w ten sposób wirtualny zasób obliczeniowy o dużej wydajności. Całość systemu jest ukierunkowana na symulacje molekularne. Zakres symulacji rozpoczyna się od typowych prostych symulacji modelowych, a kończy się na zadaniach o dużych wymaganiach wydajnościowych – takich jakie zwykle przetwarzane są na superkomputerach. Należy zwrócić uwagę, że wykorzystanie systemu GPUGrid pozwala na uruchamianie masowych aplikacji o bardzo wysokich wymaganiach na platformie o stosunkowo niskich kosztach, co w przypadku obliczeń biome-



Rysunek 4. GPUGrid – *volunteer computing* w zastosowaniach biomedycznych

Źródło: GPUGrid – volunteer computing for biomedicine, www.gpugrid.net (dostęp: 11.2014).

dycznych wprowadza nową jakość. Dzięki dobrej woli i zaangażowaniu wolontariuszy można na platformie wykonywać symulacje oraz uzyskiwać rezultaty dla istotnych problemów z tej dziedziny, takich jak: modelowanie struktury białek, analiza biomolekularna w chorobach typu rak, HIV/AIDS czy schorzeniach neurologicznych oraz porównywanie nieprawidłowości struktur proteinowych z opracowanymi modelami referencyjnymi.

Zakończenie i wnioski

Obecnie sieciowe systemy przetwarzania rozproszonego są dynamicznie rozwijane. Podstawowymi paradygmatami stosowanymi w tym zakresie są grid computing oraz volunteer computing. W analizowanym obszarze można podać szereg zaawansowanych przykładów zarówno zagranicznych, jak i zrealizowanych z sukcesem w Polsce. Na aktualnym etapie systemy te pełnią zwykle rolę platform do zastosowań badawczych oraz akademickich. Tym niemniej czynione są próby szerszego, komercyjnego wykorzystania tych systemów[8]. Ze względu na potencjalne ogromne możliwości przetwarzające oraz dużą elastyczność i adaptacyjność, naturalne wydaje się zastosowanie systemów typu grid oraz volunteer computing do przetwarzania złożonych i zaawansowanych problemów natury ekonomicznej oraz biznesowej. Pośród wielu zagadnień natury ekonomicznej oraz biznesowej mogących w naturalny sposób być podmiotem przetwarzania w omawianych paradygmatach należałoby wymienić: modelowanie makroeko-

[8] Grid Cafe, E-Science City, The Place to Explore Grid Computing, www.gridcafe.org (dostęp: 11.2014).

nomiczne, modelowanie mikroekonomiczne, zagadnienia ekonometrii, analizę kursów notowań giełdowych, sektorową analizę fundamentalną, zagadnienia szacowania ryzyka inwestycyjnego, szacowanie ryzyka ubezpieczeniowego, matematykę finansową oraz wiele, wiele innych. Natura wymienionych zagadnień pod względem koncepcyjnym pasuje do masowego przetwarzania rozproszonego, a jednocześnie moce obliczeniowe systemów typu grid oraz volunteer computing mogą zapewnić wymagane zasoby do efektywnego przetwarzania w tym zakresie. W najbliższym czasie należy spodziewać się dynamicznego rozwoju tego typu zastosowań.

Literatura

1. Balicki J., Krawczyk H., Nawarecki E. (red.), *Grid and Volunteer Computing*, Wydawnictwo Politechniki Gdańskiej, wyd. I, Gdańsk 2012
2. Balicki J., Kuchta J. (red.), *Obliczenia rozproszone w systemach komputerowych o architekturze klasy grid*, Wydawnictwo Politechniki Gdańskiej, wyd. I, Gdańsk 2012
3. BOINC – Berkeley Open Infrastructure for Network Computing, University of California, National Science Foundation, boinc.berkeley.edu
4. Comcute – System utrzymania wielkiej mocy obliczeniowej w sytuacjach kryzysowych, Politechnika Gdańska, Wydział ETI, comcute.eti.pg.gda.pl
5. GPUGrid – volunteer computing for biomedicine, www.gpugrid.net
6. Grid Cafe, E-Science City, The Place to Explore Grid Computing, www.gridcafe.org
7. *PL*-Grid – Polska Infrastruktura Gridowa, www.plgrid.pl
8. SETI@home – Search for Extraterrestrial Intelligence, University of California, National Science Foundation, BOINC Project, setiathome.berkeley.edu

INTERNET DISTRIBUTED PROCESSING SYSTEMS OF THE GRID TYPE IN BUSINESS APPLICATIONS

SUMMARY

Application alternatives and integration possibilities of distributed processing powers of Internet users' computers in global worldwide network are highlighted. Paradigms of networked processing of the types of grid computing and volunteer computing are presented. Importance of the paradigms for processing of tasks requiring high computation power is underscored. Instances of system solutions are presented: system BOINC – the representative referential model system of this kind, and system Comcute – designed and implemented at the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology. Examples of other systems of the considered processing paradigms are described. Importance of the systems is highlighted by showing the present vital issues the systems can potentially process and solve. Attention is driven to applica-

tion possibilities of the approach in solving the problems connected to complex issues and processes of economic and business nature.
Keywords: BOINC, Comcute, grid computing, volunteer computing