

Dekada polskiego Internetu w nauce

Bogdan Lesyng

Internet wywarł olbrzymi wpływ na przyspieszenie procesów globalizacji w wielu obszarach zjawisk społecznych, kulturowych, ekonomicznych oraz w nauce. Prawdziwa nauka miała zawsze międzynarodowy charakter, co zapewniało jej wysoką jakość. Internet przyspieszył i pogłębił proces integracji badań prowadzonych przez międzynarodowe zespoły. Należy tutaj przypomnieć, o czym nie wszyscy pamiętają, że technologia Internetu narodziła się w CERN-ie - można powiedzieć, że narodziła się „na zamówienie” złożonych projektów badawczych w dziedzinie fizyki wysokich energii realizowanych przez duże, międzynarodowe zespoły. Internet w szczególności pozwolił na znacznie bardziej efektywną komunikację zarówno między pracownikami naukowymi jak i całymi zespołami badawczymi oraz na efektywniejsze podejmowanie decyzji w kierowaniu zespołami. W chwili obecnej możliwości stwarzane przez technologie internetowe, począwszy od najprostszej poczty komputerowej, poprzez listy dyskusyjne, wideokonferencje i zaawansowane serwisy GRID-owe, o czym będziemy jeszcze mówić, są lub stają się elementem codziennej pracy badacza. Rys. 1 zwraca uwagę na najbardziej istotne zalety wykorzystania Internetu w prowadzeniu badań naukowych.

Wykorzystanie Internetu pozwoliło na stosunkowo łatwe, z technicznego punktu widzenia, i tanie upowszechnianie badań naukowych. W Polsce od 5 lat organizowany jest Festiwal Nauki. Przykładowo, w końcu września 2001 r. w Warszawie w ramach Festiwalu organizowanych jest około 400 imprez naukowych adresowanych do całego społeczeństwa oraz 115 lekcji dla młodzieży szkolnej (rys.2). Festiwal od strony administracyjnej oraz organizacyjnej do poziomu wydziałów i/lub instytutów naukowych obsługują jedynie trzy osoby: dyrektor i sekretarz Festiwalu oraz przewodniczący Rady Programowej. Bez Internetu takie przedsięwzięcie nie mogłoby być realizowane operacyjnie przez tak małą grupę osób. Wydarzenia festiwalowe i niektóre prezentacje są dokumentowane na serwerze WWW o adresie: <http://www.icm.edu.pl/festiwal/>.

W chwili obecnej portale internetowe są podstawowym standardem udostępniania różnorodnych usług lub przekazywania określonych informacji. Gdybym zapytał Państwa, jaki był pierwszy w Polsce bardzo popularny portal (wtedy kiedy to pojęcie nie istniało

jeszcze w języku polskim) (rys.3), wielu z Państwa mogłoby mieć kłopoty z odpowiedzią. Czas technologii internetowych tak szybko płynie! Dla ułatwienia mogę dodać, że było to depozytorium bardzo wielu popularnych zasobów informacyjnych o charakterze naukowym, oprogramowania typu *public domain*, oraz fizycznym miejscem lokalizacji jedynej w tamtych czasach polskiej przeszukiwarki internetowej Infoseek. To był SunSite (rys.4). Należy również dodać, że SunSite był przez pewien czas wykorzystywany jako pierwszy w Polsce serwer W3cache – serwer tymczasowego przechowywania przesyłanych z zagranicy zbiorów, w celu minimalizacji międzynarodowego ruchu internetowego. Twórcą serwerów SunSite oraz W3cache był Wojciech Sylwestrzak z Interdyscyplinarnego Centrum Modelowania Matematycznego i Komputerowego Uniwersytetu Warszawskiego (ICM).

Nie byłoby nowej jakości w prowadzonych w Polsce badań w dziedzinie nauk obliczeniowych i informacyjnych oraz modelowania matematycznego i komputerowego, gdyby nie centra komputerów dużej mocy. Komitet Badań Naukowych ufundował 5 takich centrów w Polsce - trzy większe: w Krakowie (CYFRONET), Poznaniu (PCSS) i Warszawie (ICM) oraz dwa mniejsze w Gdańsku (TASK) i Wrocławiu (WCSS). W centrach tych realizowanych jest kilkaset projektów przez grupy badawcze z całego kraju (rys.5). Centra, połączone są poprzez sieci POL-34 oraz NASK, wykorzystują również innych operatorów. Centra KDM dostarczają zarówno mocy obliczeniowych jak i licencjonowanego oprogramowania w różnorodnych dziedzinach nauk przyrodniczych i obliczeniowych. Nazwa „komputery dużej mocy” jest nieco na wyrost. W skali europejskiej i światowej są to jeszcze stosunkowo małe centra, jednak różnorodność architektur komputerowych (CRAY, IBM, SG, NEC, SUN) pozwala na stosunkowo wygodne prowadzenie badań wymagających specyficznego sprzętu obliczeniowego. Istnieje poza tym szereg lokalnych uczelnianych centrów komputerowych i/lub grup badawczych dysponujących laboratoriami komputerowymi, w których realizowane są prace naukowe na dobrym europejskim poziomie. Wyróżniające się instytucje wymienione są również na rys.5.

Zdarza się, chociaż jest już to zjawisko coraz rzadsze, że kwestionuje się celowość prowadzenia badań naukowych z wykorzystaniem metod obliczeniowych i symulacyjnych. Chciałbym zatem zwrócić uwagę, że celem tego typu badań naukowych jest przede wszystkim pełniejsze zrozumienie złożonych układów i procesów w fizyce, chemii, biologii czy ekonomii. Bez zaawansowanych technik symulacyjnych postęp w tych dziedzinach byłby niemożliwy. Rys. 6 obrazuje dyskutowany problem.

Przykładem wartościowego projektu, wykorzystującego nietrywialną fizykę atmosfery oraz metody obliczeniowe, jest projekt numerycznych prognoz pogody realizowany w ICM (rys.7). Inne przykłady pochodzą np. z klasycznej i kwantowej teorii układu wielu ciał czy modeli hydrodynamicznych astrofizyki (rys.8). Rys 9 przedstawia przybliżoną statystykę obciążania komputerów w centrach KDM na realizację projektów badawczych w określonych dziedzinach nauki. Najbardziej obciążające czasowo obliczenia realizowane są w dziedzinie chemii kwantowej (około 36 %).

Istotną rolę w realizacji badań naukowych z wykorzystaniem dostępnych poprzez Internet serwerów zasobów informacyjnych odgrywają naukowe bazy danych. Przegląd zasobów informacyjnych zawarty jest m.in. w wydawnictwach TASK-u, zawierających materiały konferencyjne poświęcone naukowym bazom danych (rys. 10). W TASK-u została stworzona wartościowa baza danych o dostępnych w Polsce bazach, mówiąc w skrócie „baza baz”.

W Polsce największe zasoby informacyjne i obliczeniowe, w szczególności różnorodne bazy danych, stanowiące podstawę tworzonej naukowej biblioteki wirtualnej i krajowe licencje oprogramowania, zlokalizowane są w ICM (rys.11). W szczególności ICM zapewnia dostęp zarówno lokalny jak i sieciowy do takich aplikacji jak *Advanced Visualization System (AVS)*, *Molecular Simulation (MSI)*, TRIPOS oraz do baz danych chemicznych i fizyko-chemicznych Beilsten/Gmelin. I tak MSI jest biblioteką przeszło 200 pakietów symulacji w dziedzinie molekularnego modelowania i/lub projektowania układów biomolekularnych oraz materiałowych, optoelektroniki, fizyki polimerów, przejść fazowych, itd. W bazie Beilstein można przykładowo znaleźć przepis realizacji jednej z kilkudziesięciu milionów opisanych w literaturze reakcji chemicznych.

Istotną wartość naukową mają dostępne sieciowo genetyczne bazy danych znajdujące się w Instytucie Biochemii i Biofizyki PAN w Warszawie, który jest jednym z węzłów EMBnet-u (*European Molecular Biology Network*), rys.12 i 13. Kopie (*mirrors*) genetycznych baz zawierają m.in. sekwencje ludzkiego genomu oraz innych organizmów, jak również biblioteki programów do analizy sekwencji kwasów nukleinowych (nośników informacji genetycznej) i białek (rys. 14). Nie trzeba przekonywać o ważności i wartości tego rodzaju zasobów informacyjnych. Sekwencje białek, czyli kolejność aminokwasów w genach białek, zawierają informację o strukturze przestrzennej tych białek (rys.15). Przewidywanie dokładnej,

trójwymiarowej struktury białek, od której tylko krok do praktycznych zastosowań, np. projektowania leków, jest jednym z wyzwań współczesnej biologii i biofizyki molekularnej oraz bioinformatyki. Objętość genetycznych baz danych rośnie w chwili obecnej wykładniczo (rys. 16) i proces ten potrwa jeszcze przynajmniej kilkanaście lat. Przechowywanie tak dużych zbiorów informacyjnych oraz techniczne aspekty związane z ich analizą w rozproszonym środowisku sieciowym w skali Europy i całego świata będzie jednym z wyzwań współczesnej informatyki stosowanej.

Projektem o znaczeniu strategicznym z punktu widzenia rozwoju nauki i technologii w Polsce jest wspomniana już cyfrowa/wirtualna biblioteka nauki rozwijana głównie w ICM we współpracy z Biblioteką Uniwersytetu Warszawskiego oraz w znacznie mniejszej skali w innych centrach KDM (rys.17). Poza bibliotekami oprogramowania i bazami danych bardzo istotnym elementem przyszłej biblioteki są cyfrowe, pełnotekstowe wersje czasopism naukowych. Dostęp sieciowy do wirtualnej biblioteki przez każdego pracownika naukowego w Polsce jest celem, który technicznie może być zrealizowany w bardzo krótkim czasie. Jedynym istotnym ograniczeniem jest finansowanie tego projektu przez KBN, szczególnie dotyczy to funduszy na odpowiednie zakupy licencyjne. Poza wymienionymi już zasobami informacyjnymi na uwagę zasługuje dostępny w ICM przez Internet *Science Citation Index* z abstraktami oraz bazy abstraktów czasopism biomolekularnych i biomedycznych (MEDLINE) oraz fizyki, chemii i nauk technicznych (INSPEC). Korzystanie z baz danych staje się coraz bardziej popularne i systematycznie wzrasta. Przykład statystyki wykorzystania niektórych baz przedstawia rys. 18. Rysunki 19 i 20 zawierają dodatkowe informacje dotyczące aktualnego stanu rozwijanej biblioteki oraz plany na przyszłość. Więcej na ten temat na serwerze ICM: <http://www.icm.edu.pl/>.

Większość przyszłych serwisów sieciowych oraz zdalnego przetwarzania, w szczególności tych realizowanych na rzecz środowiska naukowego, będzie miała miejsce w rozległych sieciach europejskich i światowych w technologii GRID-owej (rys.21). W pewnym uproszczeniu technologie GRID-owe mają doprowadzić do takiej sytuacji, że użytkownik Internetu, w szczególności pracownik naukowy, będzie miał dostęp do dedykowanych zasobów zlokalizowanych w innych centrach, tak jakby były one zlokalizowane na jego lokalnym komputerze - „pod stołem”. Będzie mógł również przetwarzać dane z wykorzystaniem rozproszonych w Europie zasobów obliczeniowych, sterując procesem obliczeniowej z „centralnej konsoli” swojej lokalnej stacji roboczej. Do tego celu potrzebne

jest złożone środowisko sieciowe realizujące przyjęte przez wszystkich użytkowników odpowiednie standardy komunikacyjne oraz protokoły bezpiecznego dostępu do danych i ich transferu. Nietrywialnym problemem jest też odpowiednio zunifikowany system rozliczeń finansowych za wzajemnie świadczone usługi przez centra tworzące odpowiednie konsorcja GRID-owe. Przybliżona, logiczna struktura typowego GRID-u przedstawiona jest na rys.22. Rysunki 23-25 przedstawiają stan europejskich projektów GRID-owych realizowanych lub będących w trakcie tworzenia z udziałem polskich instytucji naukowych (stan na koniec roku 2001). Jeżeli chodzi o GRID-owe środowiska sieciowe, w ramach których realizowanych jest większość projektów, to należy wymienić tutaj amerykański GLOBUS oraz niemiecki UNICOR. Rozwijane są również systemy interfejsów, pozwalające na pracę w jednym systemie i pozwalające na dostęp do zasobów w drugim systemie. Na uwagę zasługuje inicjatywa Argonne National Laboratory oraz firmy IBM stworzenia standardów, jakie mają spełniać przyszłe, rozszerzone środowiska GRID-owe.

Należy podkreślić, że rozwój nauk w obszarze fizyki wysokich energii, astrofizyki, biologii molekularnej, w tym genomiki i proteomiki, będzie silnie zależał od rozwoju nowoczesnego Internetu, który w warstwie logicznej będzie realizował nowoczesne rozwiązania GRID-owe, wykraczające znacznie poza dotychczasowe technologie WWW. Zaangażowanie polskich instytucji naukowych w rozwój tych rozwiązań jest w pełni zadowalający. Jedynym ograniczeniem, jak zwykle, mogą być ograniczenia finansowe, ale to jest już inna historia.