

Kilka słów prawdy o EGEE (*Enabling Grids for E-science*)

Nie, wbrew tytułowi, nie będzie tu żadnych sensacji w stylu spiskowej historii dziejów. Wokół projektu narosło jednak sporo niejasności, nawet wśród ludzi, którzy mają z nim do czynienia. Jako jedna z osób biorących udział w projekcie, postanowiłem w kilku zdaniach uporządkować i opisać pewne pojęcia, chronologię i mechanizmy związane z projektem. Przy okazji EGEE pojawiły się takie nazwy jak LCG, Tier2, Atlas, BalticGrid, by przytoczyć tylko niektóre z nich. Poniżej postaram się je wyjaśnić i uporządkować, a szerzej ujmując, przybliżyć w bardzo nieformalny i uproszczony sposób wszystko to, co dotyczy EGEE.

Koncepcja komputerowego gridu – a więc prowadzenia obliczeń na rozproszonej infrastrukturze komputerowej – pojawiła się w połowie lat 90-tych w Stanach Zjednoczonych. Nieco wcześniej, bo na początku lat 90-tych, w **CERN** (*Europejski Ośrodek Badań Jądrowych*) pod Genewą zrodziły się plany budowy nowego, potężnego akceleratora cząstek elementarnych. Fizycy mają nadzieję, że przy jego pomocy uda się rozwiązać kilka fundamentalnych, wciąż nierozwiązanych w fizyce problemów. Nowy akcelerator **LHC** (*Large Hadron Collider*) składa się z tunela w kształcie okręgu (o długości ok. 27 km), w którym rozpędzane będą przeciwbieżne wiązki cząstek, i kilku wydzielonych miejsc, w których dochodzi do ich zderzeń. W każdym z nich umieszczone zostaną potężne detektory wychwytyjące produkty owych zderzeń i urządzenia analizujące ich parametry. W zależności od typu zderzeń i sposobu ich analizy z każdym takim miejscem stowarzyszone są różne eksperymenty: **ATLAS**, **CMS**, **LHCb**, **ALICE** i zajmujące się nimi grupy fizyków.

Bardzo szybko zdano sobie sprawę, że ilość danych produkowanych w eksperymentach przekracza możliwości obliczeniowe jednego, nawet bardzo dużego ośrodka komputerowego. Liczbę niezbędnych procesorów szacuje się bowiem na 100 tys. procesorów typu Pentium 4, a danych do składowania w ciągu jednego roku pracy akceleratora na 20 PB (petabajtów, czyli 20 tysięcy gigabajtów). Pomijając trudności finansowe, trudno sobie wyobrazić budowę i funkcjonowanie takiego molocha. Sięgnięto więc po ideę gridu komputerowego rozproszonego po ośrodkach na świecie biorących udział w projekcie. W samym Krakowie, tylko w eksperymencie ATLAS, uczestniczy w nim grupa kilkudziesięciu fizyków, co stanowi ok. 2% wszystkich biorących udział w tym eksperymencie. Powstał pomysł projektu, który miałby budowę takiej gridowej infrastruktury nadzorować i koordynować. Jego nietypowa nazwa – **LCG** pochodzi od *LHC Computing Grid* i zawiera w sobie znany skrót, LHC. Stworzono hierarchiczną strukturę, tzw. Tierów (ang. tier – warstwa, poziom). CERN stanowi Tier-0, największe ośrodki, np. FZK w Karlsruhe czy RAL w UK – Tier-1, mniejsze, np. Cyfronet – **Tier-2**, a na najniższej warstwie znajduje się np. krakowski IFJ – Tier3. Do ośrodków europejskich dołączyły ośrodki z innych części świata: USA, Tajwanu, Japonii i nazwę LCG przekształcono w 2005 r. w **WLCG** (*Worldwide LCG*). Rola ośrodków Tier-2, mniejszych, ale i liczniejszych od Tier-1, jest niebagatelna – planuje się, że wykonają ok. połowy wszystkich obliczeń.

Koszty budowy europejskiej infrastruktury gridowej są bardzo duże, rzędu kilkuset milionów Euro. WLCG nie posiada własnego finansowania czy dodatkowych subwencji – pamiętajmy, że budowa LHC i całego oprzyrządowania pochłonie już ogromne nakłady finansowe, ok. 5 mld CHF. Postanowiono więc do idei gridu zachęcić inne środowiska naukowe. Korzyści byłyby obustronne: fizycy zyskują dodatkowe możliwości finansowania, naukowcy z innych dyscyplin mogą korzystać z potężnej infrastruktury komputerowej tworzonej dla fizyków. Na kanwie tych pomysłów powstaje w marcu 2001 r. **DataGrid**, 3-letni projekt finansowany

przez EU w ramach 5-tego Programu Ramowego – tworzy się infrastruktura gridowa, rodzi nowe oprogramowanie gridowe, tzw. middleware. Grid otwiera się na inne grupy naukowców: chemików, biologów, geofizyków. Nie wszystkie kraje, w tym m.in. Niemcy i Hiszpania, zdążyły dołączyć do DataGrid. Rok później, dla wszystkich „spóźnialskich” stworzono nowy 3-letni europejski projekt, **CrossGrid**, finansowany z tych samych źródeł (europejski Program 6) co DataGrid. Jego inicjatorem i głównym koordynatorem był prof. Turała, a ośrodkiem koordynującym – krakowski Cyfronet. Oba bliźniacze projekty skończyły się wielkim sukcesem, a ich największym osiągnięciem było stworzenie, jeszcze niedoskonałej, ale już funkcjonującej infrastruktury gridowej, obejmującej ok. 70 ośrodków na całym świecie.

Sukces ten umożliwił powstanie w kwietniu 2004 r. kolejnego wielkiego europejskiego projektu – **EGEE**. Początkowa nazwa *Enabling Grids for E-science in Europe* szybko, ze względu na jego światowy wymiar, przekształca się w *Enabling Grids for E-science*. Główny nacisk w projekcie położono na dalsze doskonalenie infrastruktury, mniej na rozwój oprogramowania. Do gridu dołączają kolejne ośrodki (w chwili obecnej jest ich już ponad 200), uaktywniają się kolejne grupy użytkowników pogrupowanych w tzw. **VO** (*Virtual Organisation*). Fizycy przeprowadzają intensywne obliczenia przypadków symulowanych, nie „próznąją” też inne VO. EGEE ma odmienną strukturę niż WLCG, nie ma tu hierarchii Tier, a koordynujące ośrodki pogrupowane są w tzw. **CIC** (*Core Infrastructure Center*) i **ROC** (*Regional Operations Center*). Trzy ośrodki w Polsce: Cyfronet, ICM i PCNS stanowią ROC Federacji Europy Środkowej (CE) koordynujący pracę ośrodków we wszystkich krajach tej części Europy (Austria, Czechy, Polska, Słowacja, Słowenia, Węgry), do których ostatnio dołączyła też Chorwacja.

I w tym przypadku projekt kończy się dużym sukcesem. W kwietniu 2006 rusza jego kolejna, 2-letnia faza o nazwie **EGEE-II**. Zmiany organizacyjne są stosunkowo niewielkie. Jeszcze większy nacisk położono na zarządzanie istniejącą infrastrukturą, CIC przekształcają się w duże ROC, dochodzą kolejne ośrodki, a dotychczasowe ROC, m.in. Cyfronet, przejmują obowiązki dawnych CIC. Doceniono zasługi i pracę naszego cyfronetowego zespołu: dwukrotnie zwiększono finansowanie, zwiększając odpowiednio liczbę naszych obowiązków. Powstają kolejne projekty mocno powiązane z EGEE, np. **BalticGrid** – wspomagany przez Cyfronet – skupiający głównie „nowe” kraje bałtyckie: Litwę, Łotwę, Estonię, **Int.eu.grid** – będący kontynuacją projektu Crossgrid. Trwa gorączkowe „szlifowanie” infrastruktury w oczekiwaniu na pierwsze prawdziwe dane pochodzące z akceleratora, którego rozruch planowany jest na jesień 2007 r.

Co dalej? EGEE-II kończy się w kwietniu 2008 r. Warto podkreślić, że oba projekty: WLCG – realizowany dla fizyków i EGEE – realizowany także dla innych grup naukowców, wzajemnie się uzupełniają i w dużym stopniu pokrywają. I choć najważniejszym ich celem jest stworzenie stabilnej infrastruktury gridowej będącej w stanie przetworzyć ogromne ilości informacji spływające z eksperymentów LHC, nie można zapominać o rosnących w siłę innych VO, np. biochem, rzeszach entuzjastów gridowych i związanych z tym grup nacisków. Dla wszystkich staje się więc jasne, że po EGEE pojawi się jego następca, choć nie wiadomo jeszcze w jakiej formie. EGEE z jego elementami integracyjnymi, skupiający większość obecnych i przyszłych członków EU, doskonale wpisuje się też politycznie w szerszy pojęty model integracji europejskiej. Czy będzie to EGO (*European Grid Organisation*) lub EGI (*European Grid Initiative*), a może coś innego, jeszcze nie wiemy, ale już teraz możemy być pewni jego powstania. Mamy uzasadnioną nadzieję, że Cyfronet zajmie w nim równie ważne i prestiżowe miejsce jak dotychczas.

Na koniec kilka słów odnośnie naszej – Cyfronetu – roli w projekcie EGEE i EGEE-II, choć to temat na dłuższy raport. Jesteśmy największą i najważniejszą częścią ROC CE. Do naszych obowiązków należy, m.in. koordynacja pracy wszystkich ośrodków należących do Federacji Środkowoeuropejskiej. Doszły nowe prestiżowe obowiązki, tzw. **COD** (*CIC on Duty*, nazwa już historyczna ze względu na likwidację CIC), polegające na okresowym badaniu funkcjonalności całej infrastruktury EGEE-II (nie tylko jej części europejskiej), wykrywaniu nieprawidłowości i proponowaniu sposobu ich usunięcia. Część ekipy zajmuje się też współpracą z zespołem krakowskich fizyków, instalując lub pomagając w instalacji niezbędnego oprogramowania, np. programów do obsługi storage'u, i uczestnicząc w okresowych testach sprawności infrastruktury organizowanych przez fizyków w ramach tzw. *Service Challenge* (**SC**). Obecnie zaczyna się SC4. Prócz funkcji koordynacyjnej, jesteśmy też ważnym ośrodkiem obliczeniowym w strukturze EGEE-II, udostępniając ponad 200 procesorów. W ciągu najbliższych dwóch lat planujemy udostępnić ok. 500 procesorów i aż 200 TB przestrzeni dyskowej. W projekcie zatrudnionych jest, w różnym wymiarze czasu, jedenaście osób typu *funded* (finansowanych z projektu) i kilka *unfunded*. Ze względu na trudności lokalowe, nasz zespół jest podzielony: część pracuje w budynku głównym, część – w w budynku na ul. Gramatyka. Oto lista najważniejszych osób wraz z krótkim opisem ich podstawowych obowiązków:

Unfunded:

Michał Turała – „ojciec chrzestny” większości projektów gridowych realizowanych w Cyfronecie, koordynator projektu LCG
Aleksander Kusznir – ROC manager, koordynator projektu w Cyfronecie
Andrzej Oziębło – zastępca powyższego

Funded:

Tomasz Szepieniec – koordynacja projektu w Federacji Środkowoeuropejskiej
Marcin Radecki – CIC on duty, koordynacja części operacyjnej projektu w Federacji
Patrik Lasoń – administracja lokalnego klastra
Łukasz Skitał – administracja części klastra
Mariusz Sterzel – nowe aplikacje w EGEE
Witold Marton – sprawy finansowe, administracyjne, raporty

Oraz zatrudnieni na część etatu (*funded*):

Michał Dwużnik – storage, kontakt z fizykami
Marek Magryś (student) – j.w.
Małgorzata Krakowian (student) – CIC on duty
Łukasz Flis (student) – instalacja centralnych serwisów, opracowanie dokumentacji
Wojciech Ziajka (student) – rejestracja użytkowników Federacji, opieka nad stroną www