Program PLGrid – infrastruktura i projekty

Nowoczesna nauka wymaga do prowadzenia badań naukowych stosowania najnowszych osiągnięć technologii informatycznych. Narzędziem, które coraz bardziej limituje szybkie osiągnięcia naukowe w wielu dziedzinach, są komputery dużej mocy obliczeniowej. Jednakże sama dostępność mocy obliczeniowej nie jest wystarczająca dla efektywnego wspierania lub prowadzenia badań naukowych. Rozproszone dane, oprogramowanie, współpraca pomiędzy międzynarodowymi zespołami badawczymi – wymagają specjalnego podejścia. Zasoby obliczeniowe, dopiero w połączeniu z ułatwiającymi do nich dostęp nowoczesnymi technologiami informatycznymi i platformami obliczeń oraz opartymi na nich usługami dziedzinowymi, w pełni odpowiadają na aktualne potrzeby badaczy.

Jako autorskie rozwiązanie tego problemu Cyfronet zainicjował Program PLGrid. W 2007 roku podpisane zostało porozumienie powołujące Konsorcjum PL-Grid, w skład którego weszło pięć akademickich centrów superkomputerowych w Polsce: ICM UW w Warszawie, PCSS IChB PAN w Poznaniu, WCSS PWr we Wrocławiu, TASK PG w Gdańsku, z ACK Cyfronet AGH jako jego koordynatorem. Pierwszym celem Programu PLGrid była budowa rozproszonej infrastruktury obliczeniowej dla nauki. Zadanie to zostało zrealizowane w oparciu o projekt „*Polska Infrastruktura Informatycznego Wspomagania Nauki w Europejskiej Przestrzeni Badawczej – PL-Grid*" (2009–2012). W jego ramach powstała sfederowana infrastruktura obliczeniowa przyjazna dla użytkowników. Ta uwspólniona infrastruktura oferuje użytkownikom wygodny Portal, z którego możliwe jest aplikowanie o zasoby obliczeniowe oraz uzyskanie dostępu do szeregu usług, zarówno ogólnych jak i dziedzinowych. Dostęp do zasobów jest całkowicie automatyczny, a autoryzacja użytkowników, pracowników naukowych lub ich podopiecznych, następuje poprzez record ID bazy OPI. Poprzez Portal infrastruktura udostępnia system grantów obliczeniowych, dzięki któremu użytkownicy uzyskują gwarancję rezerwacji określonych przez nich zasobów obliczeniowych, niezbędnych do przeprowadzenia obliczeń i uzyskania wyników naukowych. Co ważne, dostęp do wszystkich zasobów, usług i narzędzi infrastruktury odbywa się za pomocą tego samego konta i powiązanego z nim hasła.

Usługi dziedzinowe

Drugi etap realizacji Programu PLGrid to budowa tzw. gridów dziedzinowych – specyficznych środowisk obliczeniowych, czyli rozwiązań, usług i poszerzonej infrastruktury obliczeniowej wraz z oprogramowaniem, dostosowanych do potrzeb różnych grup naukowców. Usługi rozwijane przez gridy dziedzinowe powstają przy ścisłej współpracy Konsorcjum PLGrid z naukowcami reprezentującymi wybrane dziedziny nauki. Usługi te mają za zadanie integrowanie specjalistycznego sprzętu wymaganego w badaniach w poszczególnych dziedzinach, integrację danych, na których opierają się obliczenia, umożliwienie dostępu do specjalizowanego oprogramowania oraz przygotowanie i wdrożenie narzędzi wspierających typowe scenariusze pracy użytkownika. Efektem wdrożenia nowych usług dziedzinowych będzie możliwość szybszego uzyskania wyników naukowych oraz usprawnienie i zautomatyzowanie pracy naukowców i grup badawczych.

Pracę nad domenowo-specyficznymi rozwiązaniami rozpoczęto w ramach projektu „*Dziedzinowo zorientowane usługi i zasoby infrastruktury PL-Grid dla wspomagania Polskiej Nauki w Europejskiej Przestrzeni Badawczej – PLGrid Plus*" dla 13 grup użytkowników spośród strategicznych dziedzin polskiej nauki: AstroGrid-PL, HEPGrid, Nanotechnologie, Akustyka, Life Science, Chemia kwantowa i fizyka molekularna, Ekologia, SynchroGrid, Energetyka, Bioinformatyka, Zdrowie, Materiały oraz Metalurgia. Obecnie, w ramach kolejnego projektu „*Dziedzinowe Usługi Nowej Generacji w Infrastrukturze PL-Grid dla Polskiej Nauki – PLGrid NG*" realizowane są kolejne obszary wsparcia informatycznego dla nastę-

fot. ACK Cyfronet AGH

pujących dziedzin: Biologia, Chemia Obliczeniowa, Complex Networks, eBalticGrid, Energetyka Jądrowa i CFD, Geoinformatyka, Hydrologia, Matematyka, Medycyna, Medycyna Spersonalizowana, Meteorologia, OpenOxides, Technologie Przetwarzania Metali oraz UNRES. Podejmowane są także działania związane z rozbudową infrastruktury w kierunku zaawansowanych aplikacji i usług wraz z jednoczesnym rozwojem bazy sprzętowej.

Wraz z rozwojem usług dziedzinowych w infrastrukturze PLGrid podjęto działania, które mają na celu uzyskanie satysfakcjonującego i gwarantowanego poziomu jakości opracowanych usług. Efektem takiego podejścia było przeszkolenie pracowników oraz wdrożenie standardu Fit SM, jednej z metodologii w zarządzaniu usługami IT. Fakt, iż w stosunkowo krótkim czasie udostępniono w infrastrukturze PLGrid dużą liczbę nowych narzędzi i platform, potwierdził konieczność rozwoju i utrzymania tej infrastruktury informatycznej zgodnie z najlepszymi praktykami zarządzania usługami IT (ang. *IT Service Management*), takimi jak ITIL czy ISO-2000. Aktywności realizowane w tym zakresie obejmują np.: zapewnienie bezpieczeństwa infrastruktury przy wdrażaniu nowych usług, dostosowanie oprogramowania tych usług do infrastruktury PLGrid, a także szkolenia i odpowiedni poziom wsparcia dla użytkowników.

Centrum Kompetencji infrastruktur typu gridowego

W wyniku zawiązania ścisłej współpracy ze środowiskami naukowymi nastąpił ważny efekt synergii tworzonych środowisk. Na podstawie wspólnych doświadczeń oraz obserwacji zmian zachodzących w stylu prowadzenia badań, Cyfronet zaplanował szereg działań, aby sprostać tym wymaganiom. Środkiem do realizacji tej inicjatywy jest projekt „*Centrum kompetencji w zakresie rozproszonych infrastruktur typu gridowego – PLGrid Core*". Ten projekt to kolejny krok w implementacji Programu PLGrid. Koncentruje się on przede wszystkim na działaniach w zakresie obliczeń chmurowych oraz na analizie dużych i rozproszonych zasobów danych. Cele projektu obejmują również:

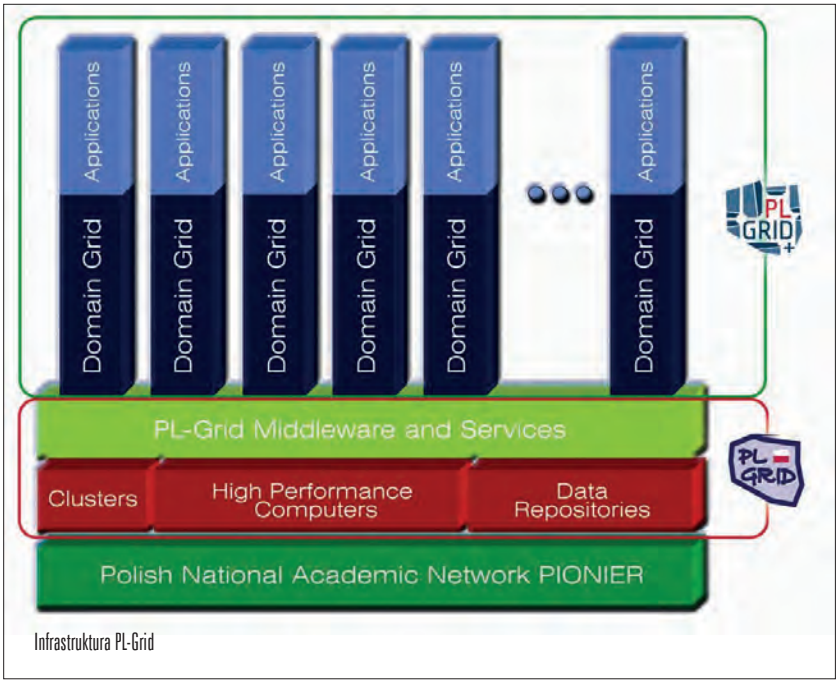
- rozwój centrum kompetencji, wyspecjalizowanego w dostarczaniu nowoczesnych technologii ICT oraz wsparcia ukierunkowanego na potrzeby naukowców korzystających z infrastruktury PLGrid,
- rozszerzenie dotychczasowej e-infrastruktury PLGrid, w celu zbudowania światowego poziomu nowoczesnych zasobów sprzętowych i implementowanych na nich innowacyjnych, wysokiej jakości usług.

To głównie dzięki pracom prowadzonym w ramach PLGrid Core, użytkownikom zostanie udostępniony Prometheus wraz z infrastrukturą towarzyszącą oraz znaczne zasoby dyskowe do przechowywania danych. Co więcej, badacze uzyskają dostęp do zestawu nowych platform bazowych, pozwalających na łatwą integrację ich specyficznych, dziedzinowych rozwiązań, z infrastrukturą PLGrid.

Nowe platformy usług obejmują:

1) **Jednolity dostęp do rozproszonych danych**, oparty o autorską platformę Onedata, która ułatwia dostęp do informacji zgromadzonych w różnych systemach ich składowania, niezależnie od miejsca odczytu danych. Usługa uwzględnia wymagania zarówno administratorów zarządzających systemami składowania danych, dostępnymi w obrębie infrastruktury ośrodków, np. w obrębie infrastruktury PLGrid, jak i jej użytkowników. Unikalną cechą platformy jest zdolność do elastycznego integrowania danych z różnych źródeł (publicznych i prywatnych, systemów plików, baz danych, zasobów chmurowych i gridowych) w jedną logiczną, zdefiniowaną przez użytkownika, przestrzeń danych, przy zachowaniu niezależności zarządzania zasobami przez wielu administratorów.

2) **Chmurę typu Platform as a Service (PaaS)**. W jej ramach opracowywana



Infrastruktura PL-Grid

fot. ACK Cyfronet AGH

jest metodyka oraz prototypowa instalacja platformy, w której różne środowiska obliczeniowe oferowane przez gridy dziedzinowe będą udostępnione w modelu PaaS na potrzeby użytkowników końcowych, z uwzględnieniem takich aspektów jak elastyczne skalowanie i bardziej efektywne wykorzystanie infrastruktury IT. Prace obejmują analizę wymagań wraz z przygotowaniem architektury, oddanie eksperymentalnej platformy oraz przeprowadzenie testów.

3) Środowisko obsługi aplikacji typu MapReduce, ułatwiające prowadzenie obliczeń z zastosowaniem paradygmatu Big Data, przy jednoczesnym uwzględnieniu technologii takich jak Hadoop i innych powiązanych, jak np. Hive i Pig. Rozwiązanie będzie oparte o otwarte oprogramowanie i realizowane na żądanie użytkownika na odpowiednich zasobach infrastruktury.

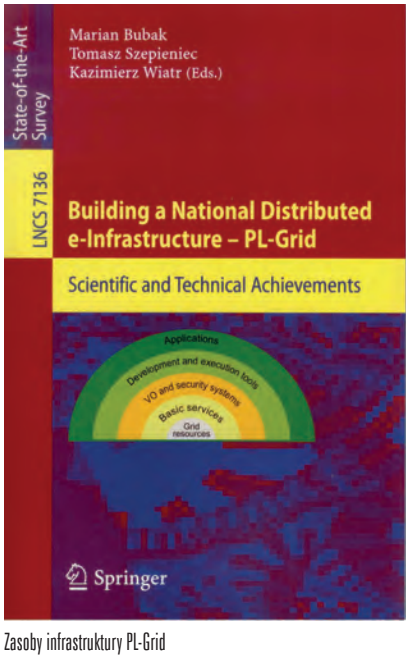
Opracowywany jest także zestaw usług przeznaczonych dla użytkowników końcowych:

- Środowisko obliczeniowe do interaktywnego przetwarzania danych, w którym użytkownik będzie miał dostęp do odpowiednio skatalogowanych i opisanych danych ze swojej dziedziny. Dane te mogą pochodzić z własnych eksperymentów użytkownika, mogą mu być udostępnione przez innych naukowców bądź znajdować się w odpowiedniej bazie wiedzy, dostępnej również z tego środowiska. Obliczenia będą mogły być prowadzone zarówno w sposób interaktywny lub zautomatyzowany – uruchamianie przygotowanego przez system zadania lub kaskad zadań sparametryzowanych, jedynie odpowiednio przez użytkownika. Interaktywne obliczenia mogą być również częściowo zautomatyzowane i oparte o gotowe komponenty, ale umożliwiają też kontrolę użytkownika na każdym etapie: zatrzymywanie części lub całości obliczeń, wybieranie elementów, które mogą przejść do następnego etapu, cofanie się do poprzednich.
- Platformę do tworzenia i uruchamiania aplikacji dużej skali zorganizowanych w *workflow*, czyli zadań obliczeniowych połączonych zależnościami. Unikalną cechą tej platformy jest wsparcie na poziomie rozwoju oraz wykonywania. Platforma przeznaczona jest dla aplikacji dużej skali, łączących przetwarzanie intensywne obliczeniowo, gdzie istnieje potrzeba optymalizacji przydziału zasobów obliczeniowych oraz analitykę dużych zbiorów danych (Big Data).
- Technologie i środowiska realizujące paradygmat Open Science. Przykładem jest usługa Rimrock, która pozwala na zarządzanie procesami i zadaniami w infrastrukturze obliczeniowej za pomocą interfejsów REST. Może być wykorzystana przez dostawców usług dziedzinowych lub zaawansowanych użytkowników do uruchamiania i monitorowania procesów lub zadań w infrastrukturze poprzez delegację certyfikatu proxy użytkownika. Udostępniane metody REST obejmują: zarządzanie procesami, interaktywnymi procesami oraz zadaniami.
- Środowisko wspierające obliczenia typu *data farming*. Proponowana platforma *Scalarm* wspiera wykonywanie tego typu eksperymentów oraz badań parametrycznych, czyli eksperymentów wymagających uruchamiania tego samego kodu z różnymi wartościami parametrów wejściowych. Celem platformy *Scalarm* jest ułatwienie wykonywania tego typu eksperymentów poprzez wsparcie: 1) specyfikacji wartości parametrów wejściowych, dla których należy uruchomić obliczenia, 2) zarządzania wykonaniem wielu uruchomień aplikacji z różnymi wartościami parametrów wejściowych na różnych infrastrukturach, 3) zbierania i analizy rezultatów obliczeń. Wykonywanie obliczeń w ramach prowadzonych eksperymentów odbywa się we wskazanych przez użytkownika infrastrukturach obliczeniowych, np. infrastrukturze PLGrid, chmurze PLGrid, chmurach zewnętrznych (Amazon EC2, Google Compute Engine) lub wskazanych serwerach dostępnych przez sieć. Platforma umożliwia uruchamianie dowolnych aplikacji, ponieważ to użytkownik specyfikuje (przy pomocy dodatkowych skryptów), w jaki sposób aplikacja ma być uruchamiana, w jaki sposób należy przekazać parametry wejściowe i jaki jest rezultat wykonywanych obliczeń. Wyniki uruchamianych obliczeń są automatycznie zbierane i przesyłane do platformy *Scalarm*, skąd mogą być pobrane i poddane dalszej analizie.

Oprócz rozbudowy sprzętu do obliczeń i magazynowania danych oraz tworzenia nowego oprogramowania, prace w projekcie PLGrid Core koncentrują się także na rozbudowie niezbędnej infrastruktury towarzyszącej oraz budowie nowego, zapasowego, centrum danych. Dzięki temu, dane naukowe przechowywane w postaci wielostopniowych kopii zapasowych, będą dodatkowo archiwizowane w oddzielnej lokalizacji. Takie geograficzne rozdzielenie pozwoli na znaczące zwiększenie ich bezpieczeństwa.

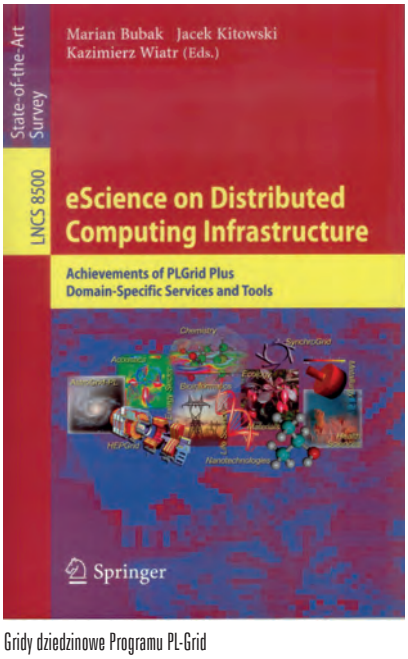
Usługi dodatkowe w infrastrukturze PLGrid

Infrastruktura informatyczna dla nauki to nie tylko wielkie moce obliczeniowe, ogromne magazyny dla danych cyfrowych, usługi gridow dziedzinowych i nowoczesne platformy bazowe. To również różnorakie oprogramowanie wspierające organizację pracy badawczej i planowanie zadań w ramach danego projektu lub zespołu badawczego. Dzięki Infrastrukturze PLGrid możliwe jest skorzystanie z narzędzi pracy zespołowej. Stanowią one zintegrowaną platformę do zarządzania projektami naukowców. Dzięki niej możliwe jest założenie własnego projektu w narzędziach do śledze-



Zasoby infrastruktury PL-Grid

fot. ACK Cyfronet AGH



Gridy dziedzinowe Programu PL-Grid

nia i organizowania współpracy, takich jak: Confluence, JIRA i Stash oraz korzystanie z zaawansowanego systemu do tworzenia wideokonferencji Adobe Connect. Platforma umożliwia także ustawienie niezbędnych uprawnień do korzystania z projektu dla skojarzonej grupy użytkowników.

Na potrzeby użytkowników oddany został również serwis Katalog Aplikacji, czyli system zbierający i udostępniający informacje na temat różnorakich aplikacji naukowych, narzędzi programistycznych oraz bibliotek oferowanych w ramach infrastruktury PLGrid. Pozwala on na wyszukiwanie aplikacji, sprawdzanie stanu ich dostępności w poszczególnych ośrodkach i na konkretnych maszynach obliczeniowych, udostępnia informacje o zmianach i pojawiających się nowościach, a także udostępnia dokumentację i przykłady użycia. Katalog Aplikacji w jednolity i prosty sposób prezentuje pełną i aktualną ofertę oprogramowania. Nawigację po serwisie ułatwia podział na kategorie: zastosowania oraz dziedziny naukowe. Narzędzie rekomendowane jest w szczególności dla nowych użytkowników, którzy dzięki niemu będą mogli wybrać odpowiednie dla swoich potrzeb aplikacje.

– Warto podkreślić, że sukces Programu PLGrid to efekt wspólnych wysiłków wszystkich pięciu polskich centrów superkomputerowych – podkreśla Dyrektor Cyfronetu prof. Kazimierz Wiatr. Trwała współpraca członków Konsorcjum, połączenie ich wiedzy i doświadczenia, a także udostępnianych mocy obliczeniowych, magazynów danych i profesjonalnego oprogramowania otworzyła ogromne możliwości przed zespołami naukowców. Dzięki szeregom wspólnych inicjatyw udało się stworzyć nową jakość w polskiej nauce. Potężne zasoby obliczeniowe udostępniane naukowcom, usługi z gwarancją wysokiej jakości i bezpieczeństwa realnie wpływają na rozwój badań wspieranych przez infrastrukturę informatyczną. Dostęp do tych usług został bardzo uproszczony. Aby skorzystać z wszystkich zasobów wystarczy jedynie konto w serwisie Portal PLGrid. Dodatkowym elementem tego sukcesu jest bliska współpraca z użytkownikami. Za pomocą platformy Helpdesk oraz poprzez organizowane spotkania i szkolenia tworzy się środowisko, w którym poznajemy profil użytkownika i kształtujemy Infrastrukturę tak, aby była ona wygodnym narzędziem, po które będą sięgać kolejni naukowcy – konkluduje Dyrektor.

Szczegółowe informacje o infrastrukturze i ofercie PLGrid dostępne są na stronie

www.plgrid.pl

Jacek Kitowski, Robert Pająk, Mariusz Sterzel