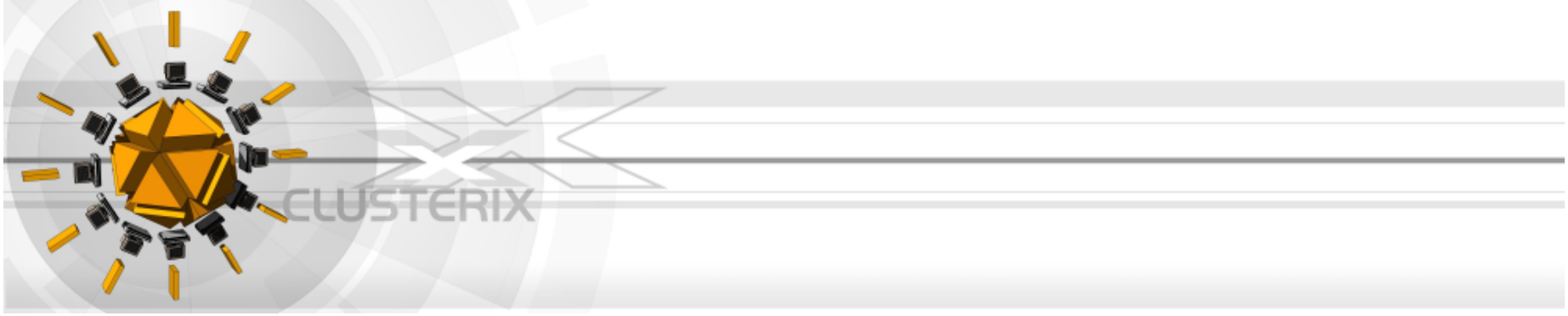


Krajowy Klaster Linuxowy
CLUSTERIX

Roman Wyrzykowski

Instytut Informatyki Teoretycznej i Stosowanej
Politechnika Częstochowska

SIECI NOWEJ GENERACJI

POLSKIE PROJEKTY SIECIOWE – PIONIER (1Q05)

Podobne rozwiązanie USA:
National Lambda Rail

Status projektu CLUSTERIX

- Projekt celowy MNiI „**CLUSTERIX** - Krajowy Klaster Linuxowy (National Cluster of Linux Systems)"
- Uczestniczyło 12 jednostek z całej Polski (MAN-y i KDM-y)
- Politechnika Częstochowska jako koordynator
- Faza badawczo-rozwojowa: 01.01.2004 - 30.09.2005
- Faza wdrożeniowa: 01.10.2005 – 30.06.2006

Partnerzy

- **Politechnika Częstochowska (koordynator)**
- Poznańskie Centrum Superkomputerowo-Sieciowe (PCSS)
- Akademickie Centrum Komputerowe CYFRONET AGH, Kraków
- Trójmiejska Akademicka Sieć Komputerowa w Gdańsku (TASK)
- Wrocławskie Centrum Sieciowo-Superkomputerowe (WCSS)
- Politechnika Białostocka
- Politechnika Łódzka
- UMCS w Lublinie
- Politechnika Warszawska
- Politechnika Szczecińska
- Uniwersytet Opolski
- Uniwersytet w Zielonej Górze

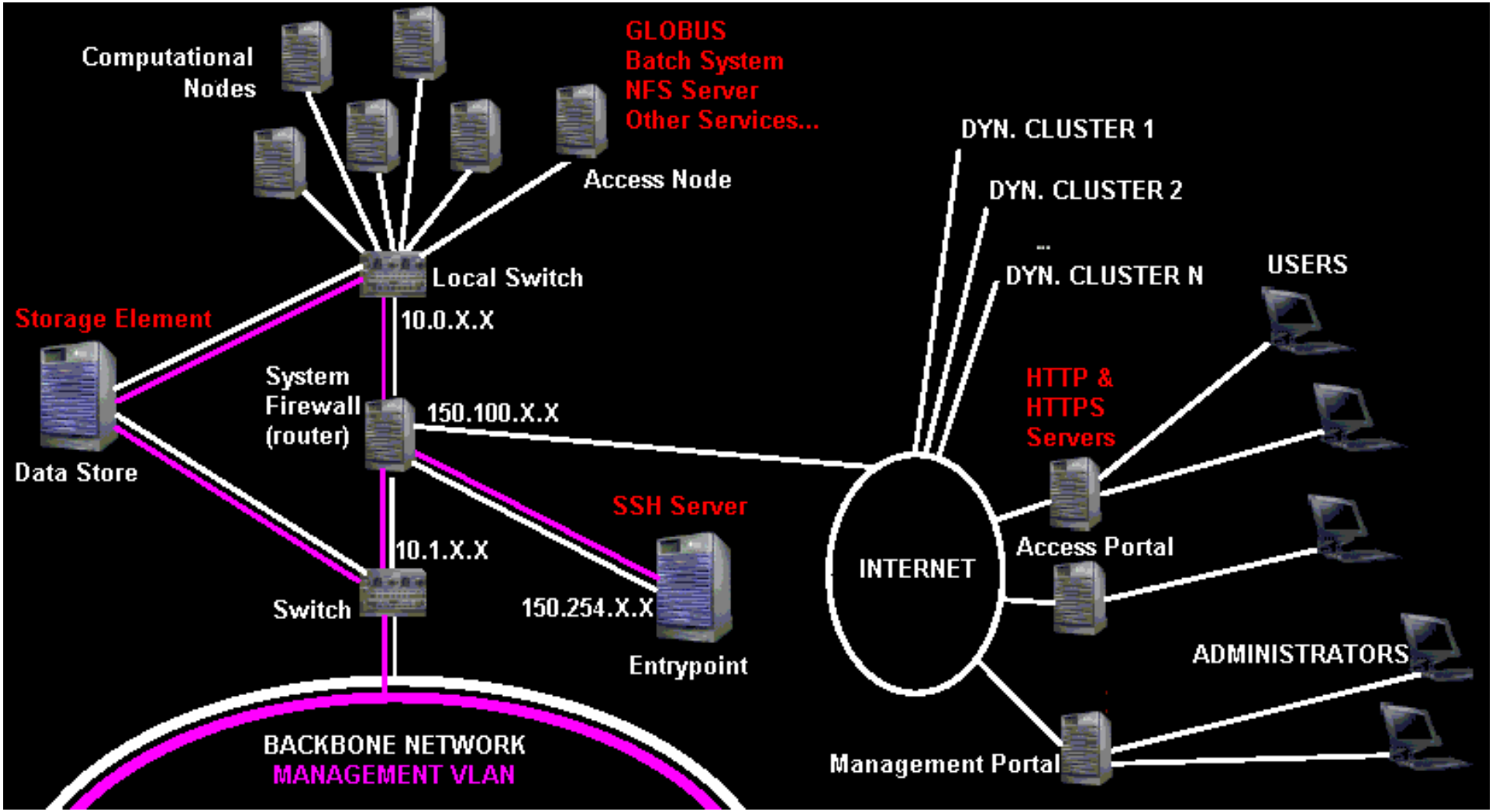
Założenia projektu

- Głównym celem realizacji projektu był opracowanie narzędzi oraz mechanizmów umożliwiających wdrożenie **produkcyjnego środowiska gridowego,**
- W konfiguracji bazowej udostępnia ono infrastrukturę lokalnych klastrów PC-Linux o architekturze 64-bitowej, połączonych szybką siecią kręgosłupową zapewnianą przez sieci PIONIER
- Do infrastruktury bazowej mogą być podłączane zarówno istniejące, jak i nowe klastry, o zróżnicowanej architekturze 32- i 64-bitowej
- W wyniku powstaje rozproszony klaster PC nowej generacji o dynamicznie zmieniającej się wielkości, w pełni operacyjny i zintegrowany z innymi usługami

Zrealizowane zadania

- **Utworzenie środowiska produkcyjnego**
  - Dostarczenie narzędzi i mechanizmów pozwalających na dynamiczną i automatyczną rekonfigurację infrastruktury
  - Umożliwi dołączanie kolejnych klastrów lokalnych 32- lub 64-bitowych
  - Instalacja pilotowa stanowi bazową infrastrukturę szkieletową połączoną siecią PIONIER
- **Rozwój oprogramowania narzędziowego**
  - Zarządzanie danymi
  - System rozdziału zasobów z predykcją, monitorowanie
  - Zarządzanie kontami użytkowników i organizacjami wirtualnymi
  - Bezpieczeństwo
  - Zarządzanie zasobami sieciowymi, IPv6
  - Interfejs użytkownika oraz administratora
  - Dynamiczne równoważenie obciążenia, mechanizm punktów kontrolnych
- **Rozwój aplikacji gridowych**
  - Modelowanie zjawisk termomechanicznych w krzepnących odlewach
  - Symulacja przepływów transonicznych wokół skrzydła samolotu
  - Symulacje wielkiej skali przepływu krwi w mikrokapilarach metodą cząstek
  - Modelowanie molekularne
  - Dynamiczna wizualizacja terenu 3D z danych przestrzennych

CLUSTERIX: Architektura



Instalacja pilotowa (1)

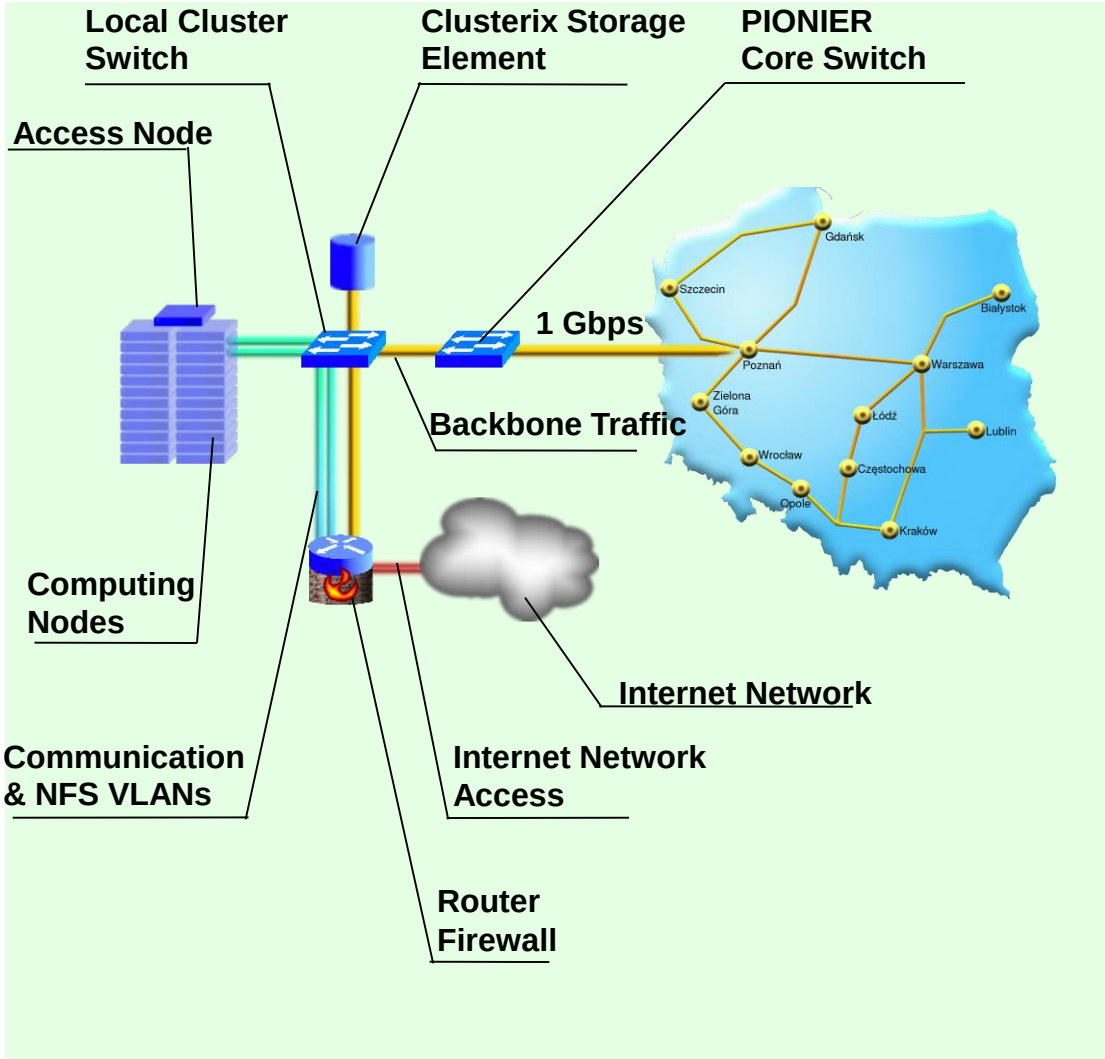
- 12 klastrów lokalnych stanowi szkielet systemu
- Klastry lokalne budowane w oparciu o 64-bitowe procesory Intel Itanium2 1,4 GHz wyposażone w pamięci podręczne o wielkości 3 MB
- 2 lub 4 GB pamięci operacyjnej RAM na procesor
- Komunikacja wewnątrz klastrów lokalnych oparta o Gigabit Ethernet i InfiniBand (lub Myrinet)
- Komunikacja pomiędzy klastrami lokalnymi odbywa się za pomocą dedykowanych kanałów 1 Gb/s udostępnianych przez sieć PIONIER

Instalacja pilotowa (2)

- 252 x IA-64 w szkielecie
- 800 x IA-64 w konfiguracji testowej - 4,5 TFLOPS
- 3000+ km włókien optycznych o przepustowości 10Gbps (technologia DWDM)
- w pełni wdrożone protokoły IPv4 oraz IPv6

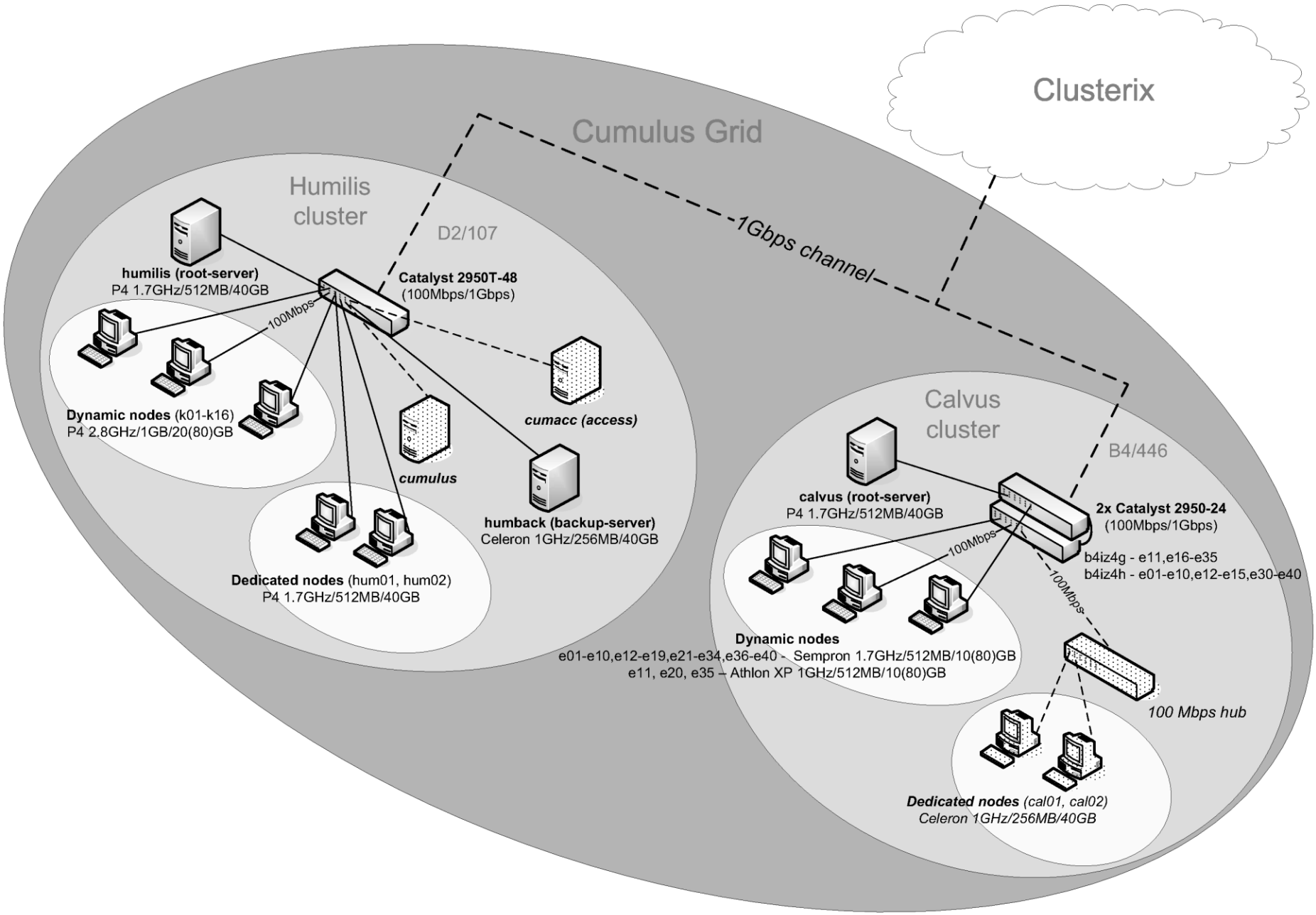
CLUSTERIX: Architektura sieci

- Bezpieczny dostęp do sieci zapewniają routery ze zintegrowaną funkcjonalnością firewalla
- Użycie dwóch VLAN-ów umożliwia separację zasobów obliczeniowych od szkieletu sieci
- Wykorzystanie dwóch sieci z dedykowaną przepustowością 1 Gbps pozwala poprawić efektywność zarządzania ruchem w klastrach lokalnych

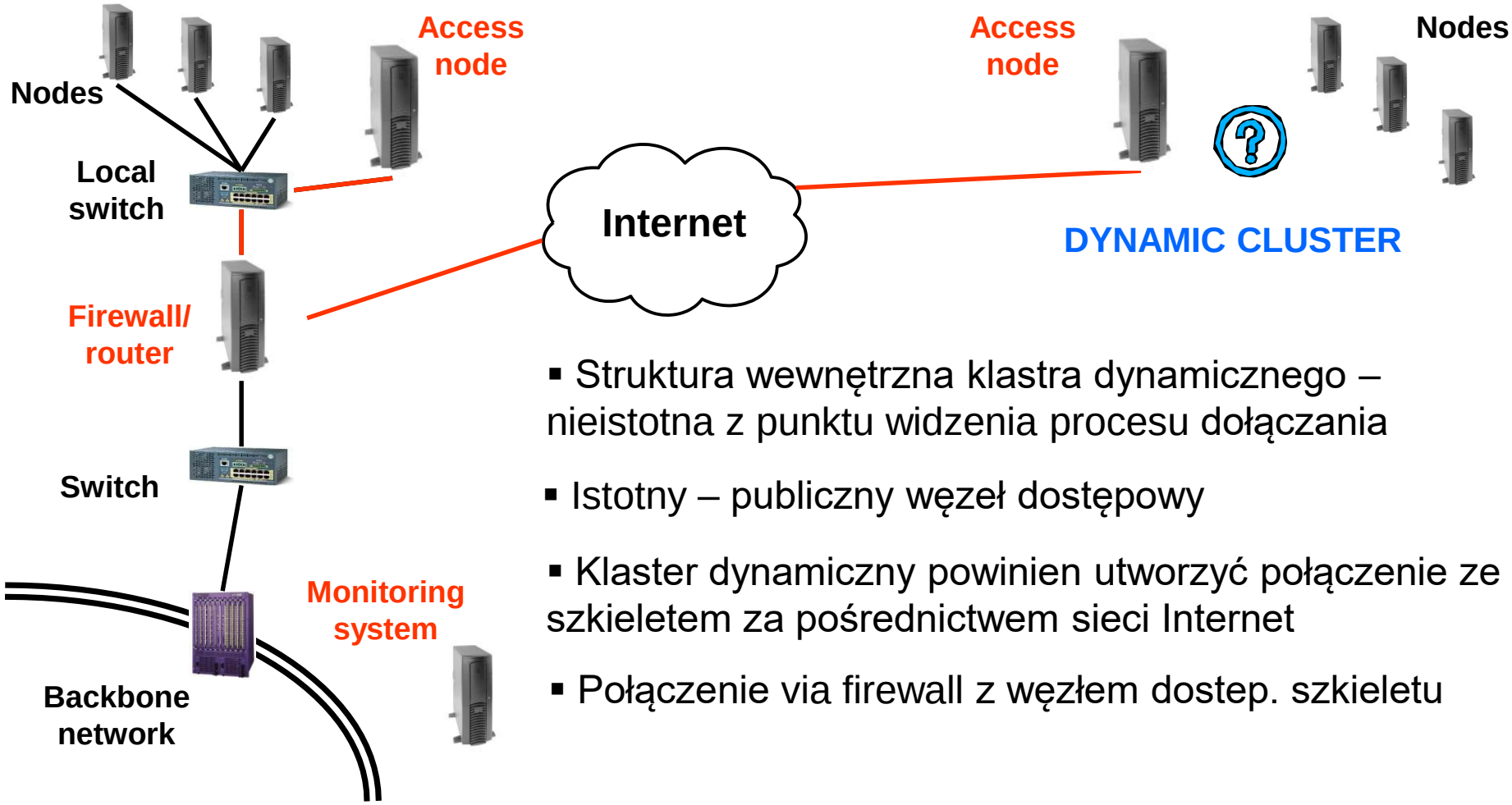


Integracja klastrów dynamicznych

- Klastry dynamiczne (zewnętrzne) mogą być w prosty sposób (automatycznie) dołączane do szkieletu systemu CLUSTERIX aby:
  - zwiększyć dostępną moc obliczeniową
  - wykorzystać zasoby klastrów zewnętrznych w momentach, gdy są one nieobciążone (noce, weekendy …)
  - zapewnić skalowalność infrastruktury systemu


Clusterix
Cumulus Grid
1Gbps channel
Humilis cluster
D2/107
humilis (root-server)
P4 1.7GHz/512MB/40GB
Catalyst 2950T-48
(100Mbps/1Gbps)
100Mbps
Dynamic nodes (k01-k16)
P4 2.8GHz/1GB/20(80)GB
cumacc (access)
cumulus
humback (backup-server)
Celeron 1GHz/256MB/40GB
Dedicated nodes (hum01, hum02)
P4 1.7GHz/512MB/40GB
Calvus cluster
B4/446
calvus (root-server)
P4 1.7GHz/512MB/40GB
2x Catalyst 2950-24
(100Mbps/1Gbps)
100Mbps
b4iz4g - e11,e16-e35
b4iz4h - e01-e10,e12-e15,e30-e40
100Mbps
Dynamic nodes
e01-e10,e12-e19,e21-e34,e36-e40 - Sempron 1.7GHz/512MB/10(80)GB
e11, e20, e35 – Athlon XP 1GHz/512MB/10(80)GB
100 Mbps hub
Dedicated nodes (cal01, cal02)
Celeron 1GHz/256MB/40GB

Dołączenie klastra dynamicznego



- Struktura wewnętrzna klastra dynamicznego – nieistotna z punktu widzenia procesu dołączania
- Istotny – publiczny węzeł dostępowy
- Klaster dynamiczny powinien utworzyć połączenie ze szkieletem za pośrednictwem sieci Internet
- Połączenie via firewall z węzłem dostep. szkieletu

Oprogramowanie zarządzające systemem CLUSTERIX - technologie

- Tworzone oprogramowanie bazuje na Globus Toolkit 2.4 oraz WEB serwisach, korzystając z Globusa dostępnego w dystrybucji Globus 3.2

  - możliwość powtórnego wykorzystania tworzonego oprogramowania

  - zapewnia interoperacyjność z innymi systemami gridowymi na poziomie serwisów

- Technologia *Open Source*, w tym LINUX (Debian, jadro 2.6.x) oraz systemy kolejkowe (Open PBS, SGE)

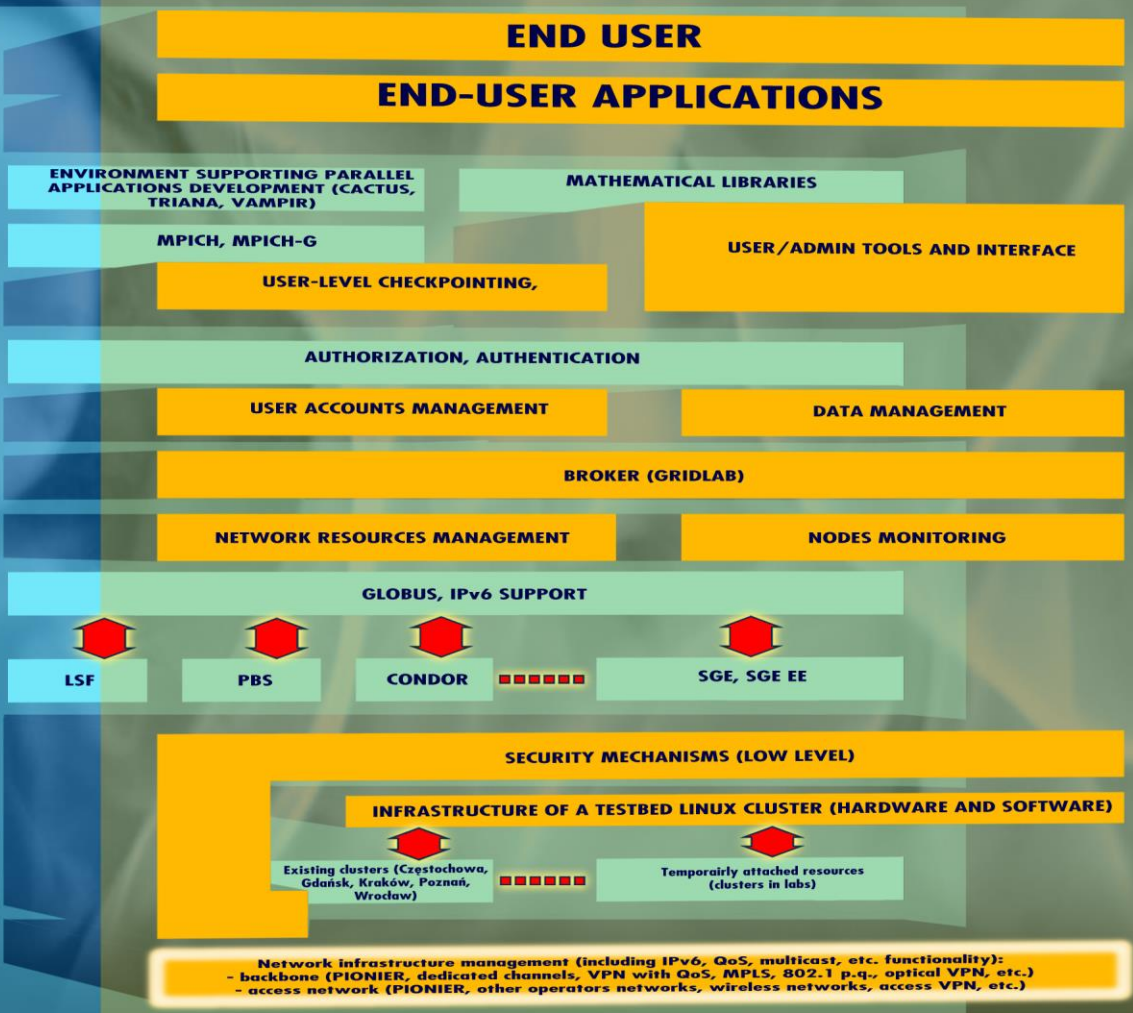
  - oprogramowanie *Open Source* jest bardziej podatne na integrację zarówno z istniejącymi, jak i z nowymi produktami

  - dostęp do kodu źródłowego projektu ułatwia dokonywania zmian i ich publikację

  - większa niezawodność i bezpieczeństwo

- Szerokie wykorzystanie istniejących modułów programowych, np. brokera zadań z projektu *GridLab*, po dokonaniu niezbędnych adaptacji i rozszerzeń

SOFTWARE ARCHITECTURE



GRMS

- GRMS (Grid Resource Management System) jest systemem zarządzania zasobami i zadaniami w projekcie CLUSTERIX
- Został stworzony w ramach projektu GridLab
- Głównym zadaniem systemu GRMS jest zarządzanie całym procesem zdalnego zlecania zadań obliczeniowych do różnych systemów kolejkowych obsługujących klastry lokalne
- Wszystkie wymagania użytkowników są specyfikowane przy pomocy specjalnych dokumentów XML, nazywanych opisami zadań GJD (GRMS Job Description), przy czym są one wysyłane do GRMS'a jako komunikaty SOAP poprzez połączenia GSI

Job Description - MPI example

```
<grmsjob appid="psolidify">
    <simplejob>
        <resource>
            <localrmname>pbs</localrmname>
        </resource>
        <executable type="mpi" count="8">
            <file name="exec" type="in">
                <url>gsiftp:////access.wcss.clusterix.pl/~/myapp/psolidify/</url>
            </file>
            <arguments>
                <value>250000.prl</value>
                <file name= "250000.prl" type="in">
                    <url>gsiftp://access.wcss.clusterix.pl/~/data/250000.prl</url>
                </file>
                </arguments>
                <stdout>
                    <url>gsiftp://access.wcss.clusterix.pl/~/app1.out</url>
                </stdout>
        </executable>
    </simplejob>
</grmsjob>
```

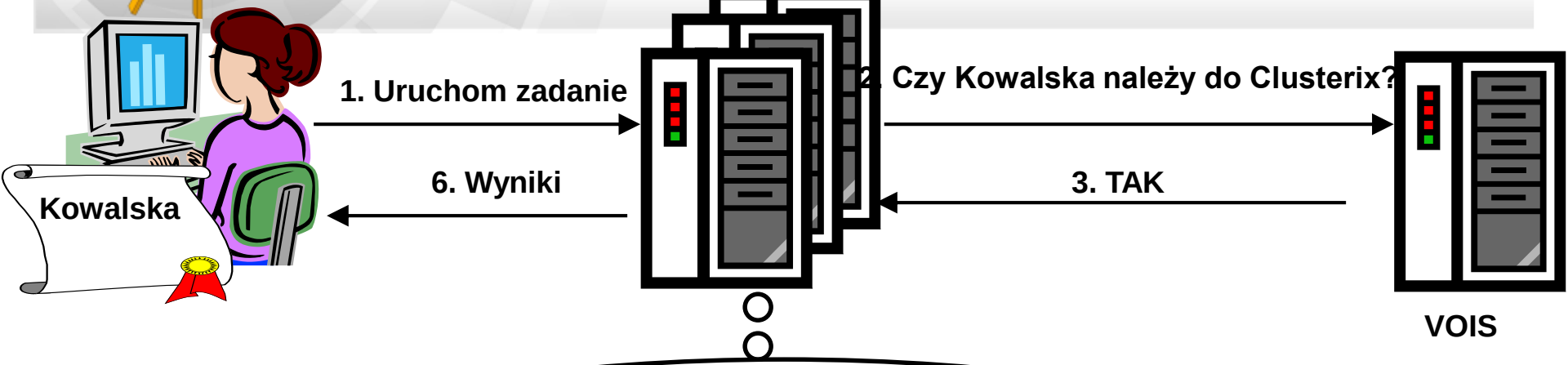
System Wirtualnych Użytkowników (VUS)

Dotychczas użytkownik musiał posiadać oddzielne konto fizyczne na każdej maszynie

VUS dla Gridu

- Zbiór ogólnodostępnych kont, które mogą być przyporządkowane kolejnym zadaniom
- Możliwość zgłaszania zadań do innych maszyn i klastrów lokalnych
- Uproszczona administracja kontami użytkowników
- Pełna informacja rozliczeniowa o wykorzystaniu kont
- Wspiera różnorodne scenariusze dostępu do Gridu – dla użytkownika końcowego, właściciela zasobów, managera organizacji

VUS - jak to działa?

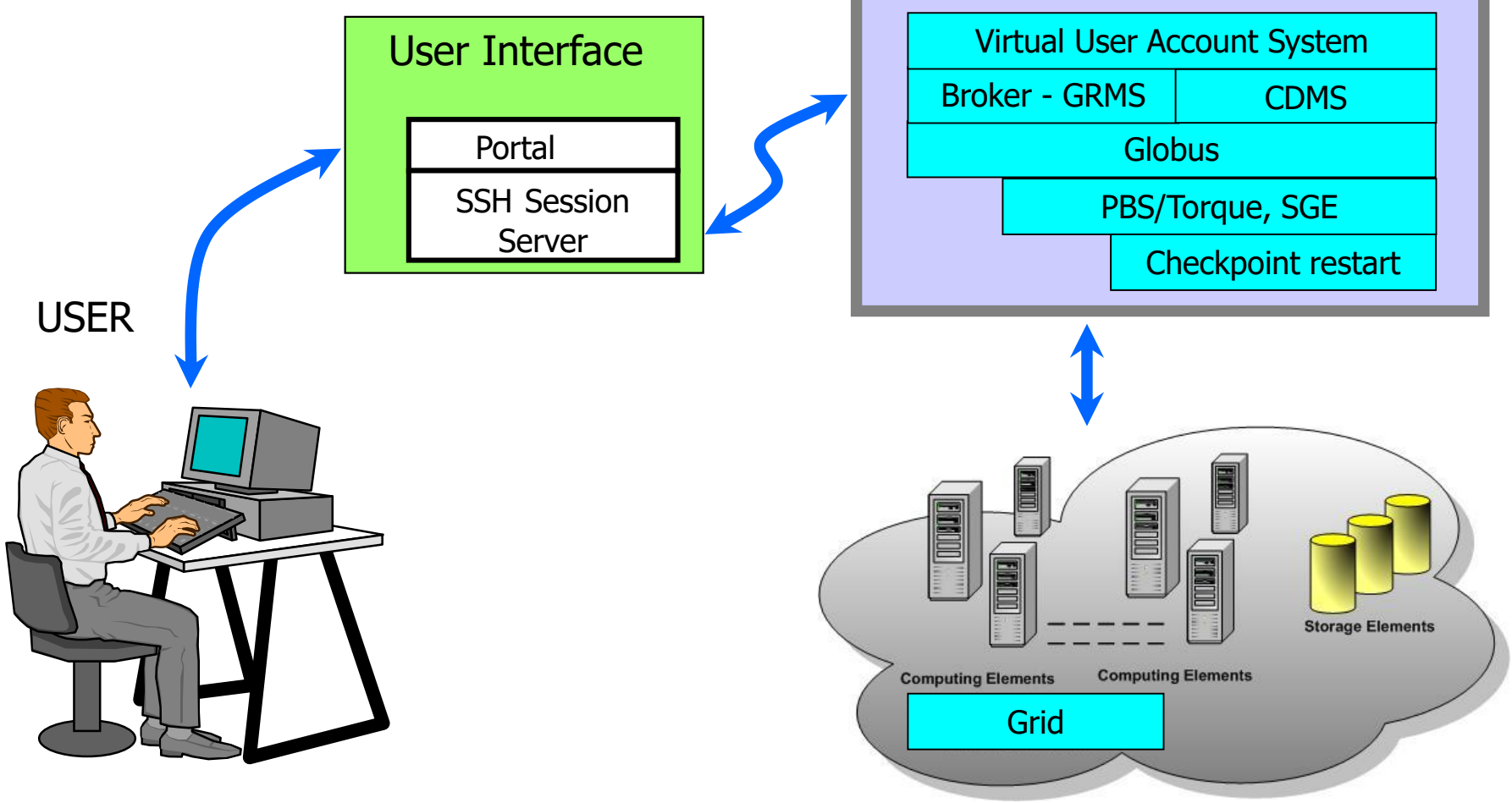


4. Zaloguj użytkownika na jedno z kont Clusterix (zawsze 1 użytkownik na 1 konto w danej chwili)

5. Wykonaj zadanie

7. Jeśli konto nie jest używane, można na nie zalogować innego użytkownika

Wykonanie zadania w systemie CLUSTERIX



Bazowy scenariusz wykonywania zadań w systemie CLUSTERIX

1. Użytkownik zleca zadanie do systemu GRMS za pośrednictwem np. portalu, przekazując opis zadania GJD
2. GRMS wybiera optymalny zasób do uruchomienia zadania, zgodnie z opisem zadania (hardware/software)
3. Transfer danych wejściowych i plików wykonywalnych:

   a) dane wejściowe opisywane przy pomocy adresu logicznego lub fizycznego - pobierane są z systemu CDMS czyli CLUSTERIX Data Management System (lub węzła dostępowego) przy współudziale GRMS

   b) pliki wykonywalne, w tym również skrypty
4. VUS odpowiada za mapowanie uprawnień użytkownika (user credentials) na konta fizyczne w klastrach lokalnych
5. Wykonanie zadania
6. Po zakończeniu zadania wyniki przekazywane są do CDMS; wykorzystywane konta fizyczne są „czyszczone" przez VUS

Aplikacje pilotowe

- Ponad 20 aplikacji użytkownika końcowego
- Wśród zaproponowanych aplikacji można wyróżnić następujące typy zadań:
  - zadania jednoprocesowe
  - sekwencja zadań jednoprocesowych
  - zadania równoległe uruchamiane na jednym klastrze lokalnym
  - sekwencja zadań równoległych uruchamianych na jednym klastrze lokalnym
  - meta-aplikacje rozproszone uruchamiane na więcej niż jednym klastrze lokalnym - **MPICH-G2**

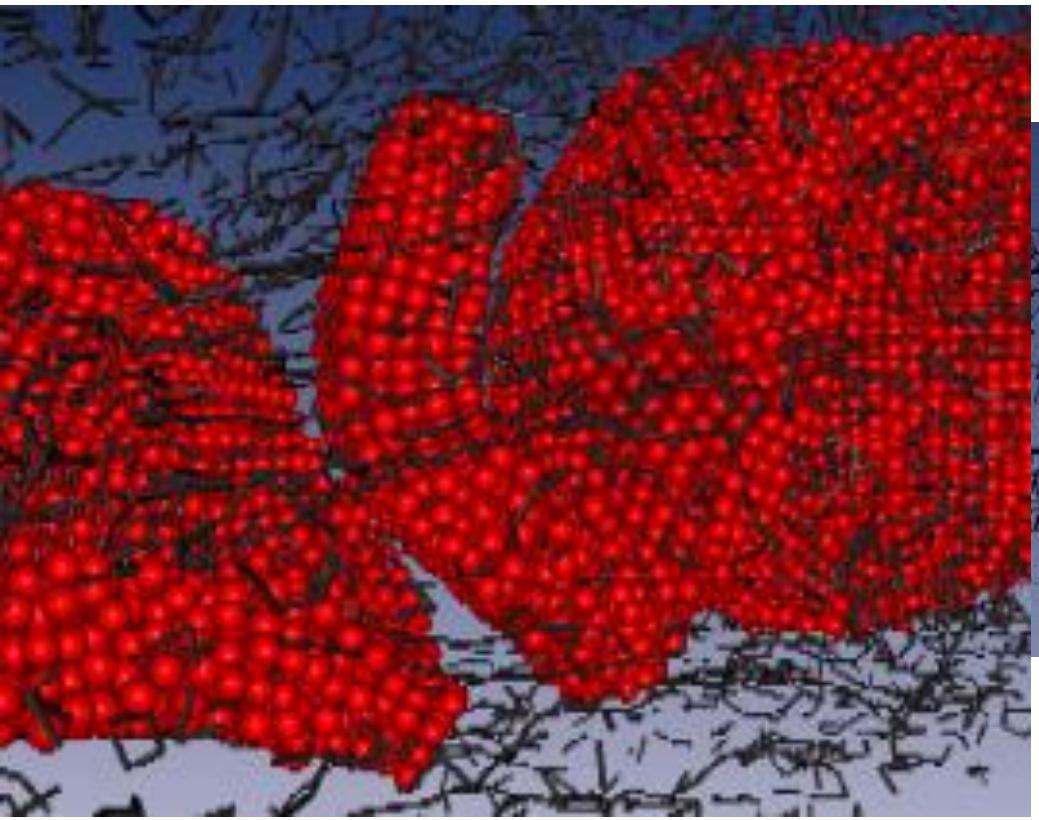
Białystok | Częstochowa | Gdańsk | Łódz | Lublin | Kraków | Opole | Poznań | Szczecin | Warszawa | Wrocław | Zielona Góra

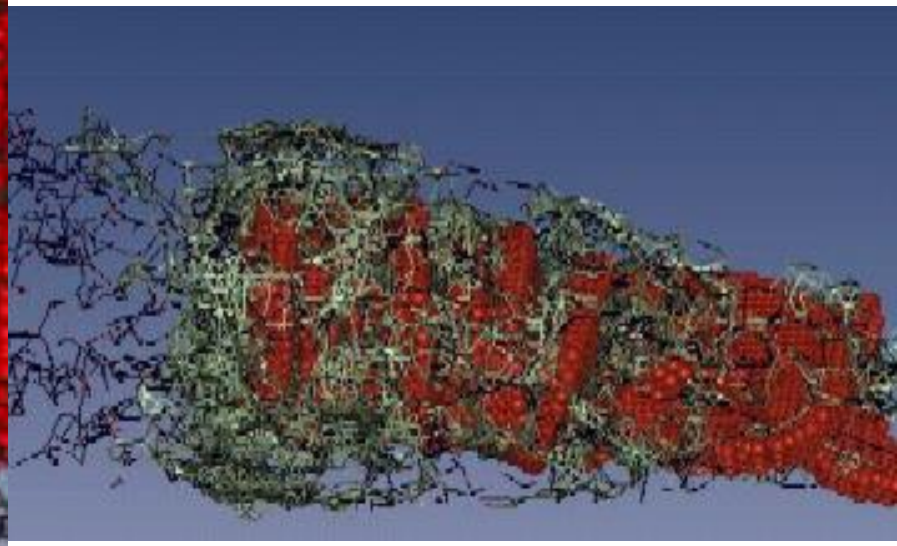
Symulacje wielkiej skali przepływów krwi w mikrokapilarach (dyskretne modele cząstkowe)

W.Dzwinel, K.Boryczko

AGH, Institute of Computer Science

Powstawanie skrzepów krwi

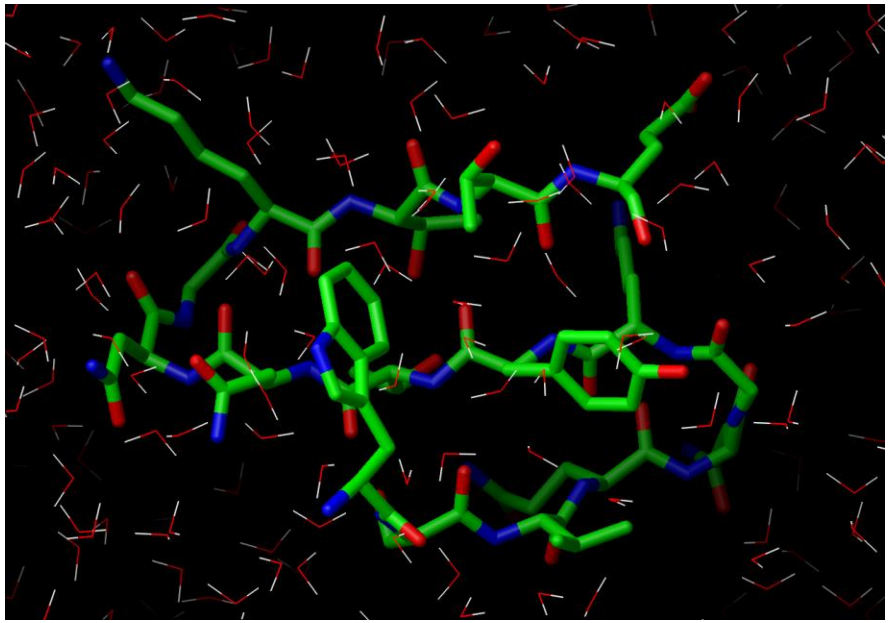


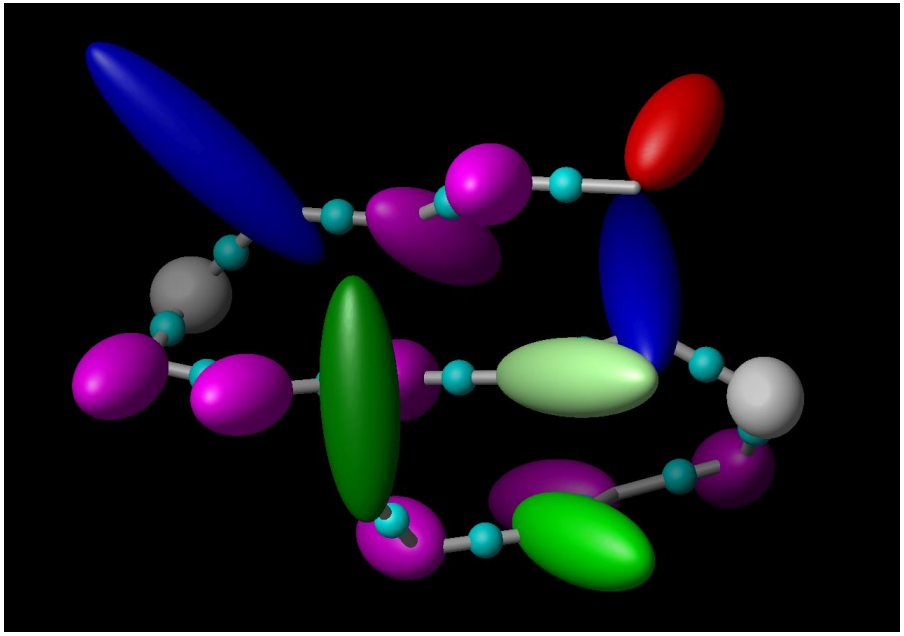


(5x10^6 cząstek, 16 procesorów)

Przewidywanie struktur białek

Adam Liwo, Cezary Czaplewski, Stanisław Ołdziej
Department of Chemistry, University of Gdansk

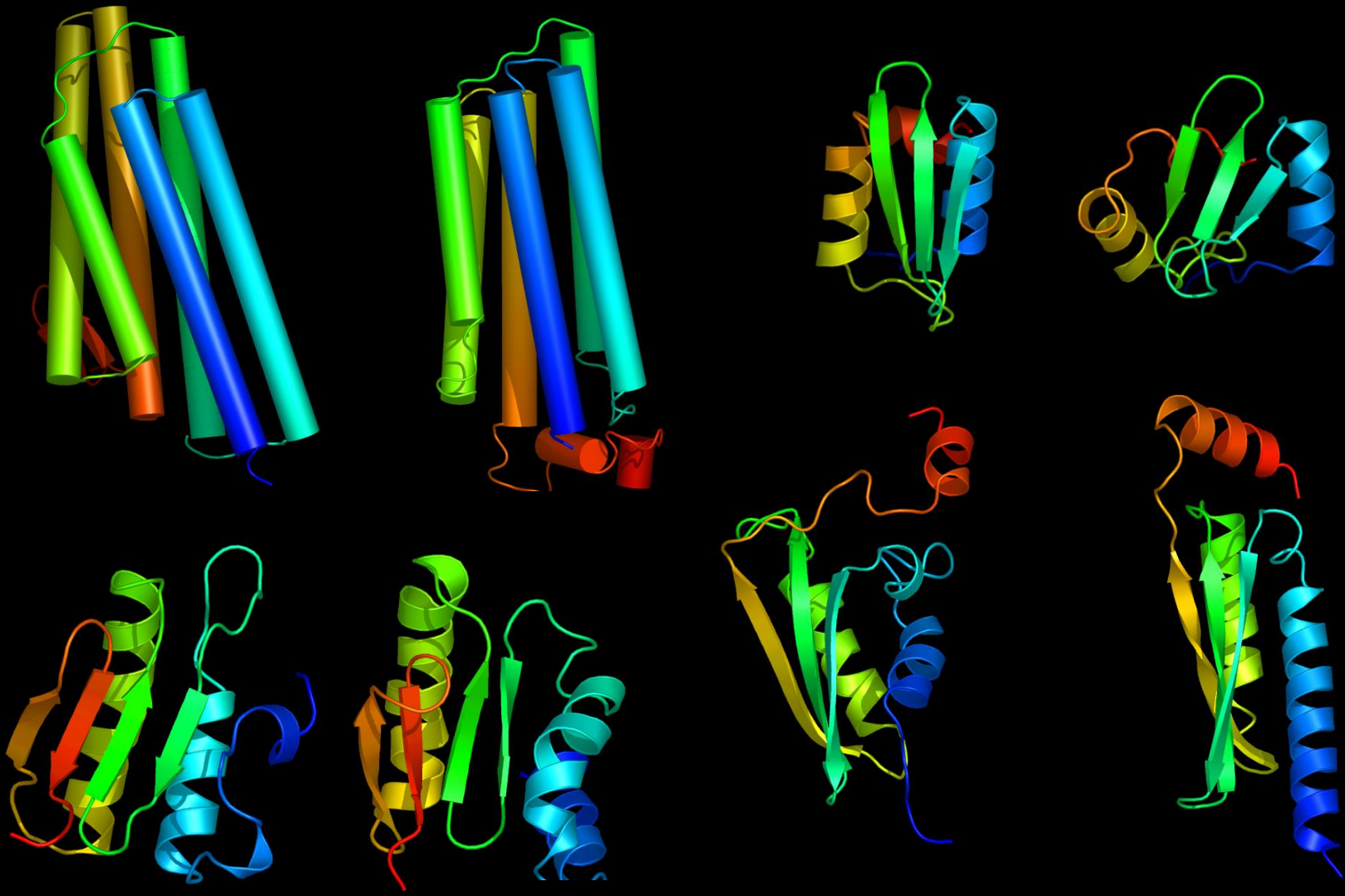




Selected UNRES/CSA results from 6th Community Wide Experiment on the

Critical Assessment of Techniques for Protein Structure Prediction

December 4-8, 2004

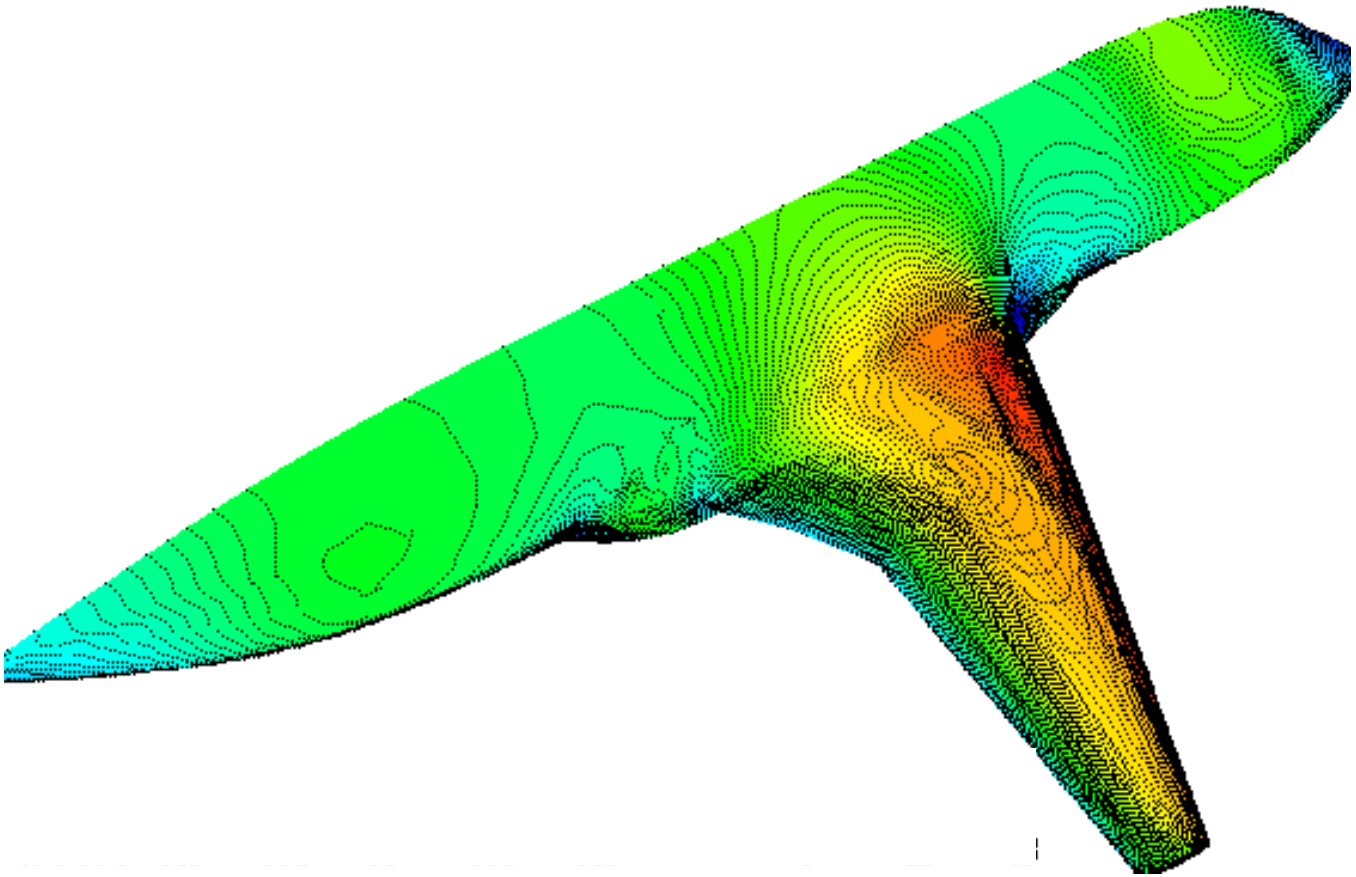


left - experimental structure, right - predicted structure

Symulacje przepływów w aeronautyce
–
autorskie oprogramowanie HADRON

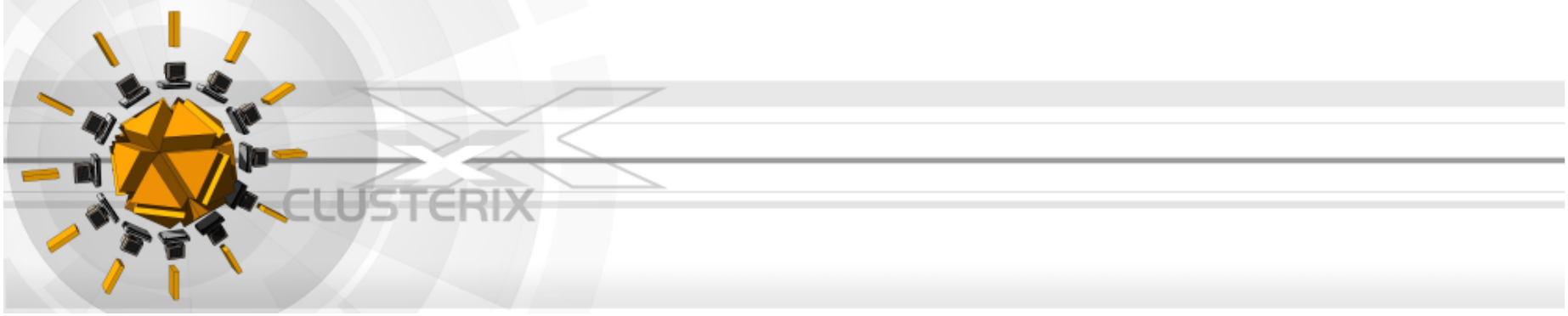
Prof. Jacek Rokicki
Politechnika Warszawska

Symulacje wielkiej skali zagadnień 3D mechaniki płynów



$3 \div 6 \times 10^6$ węzłów

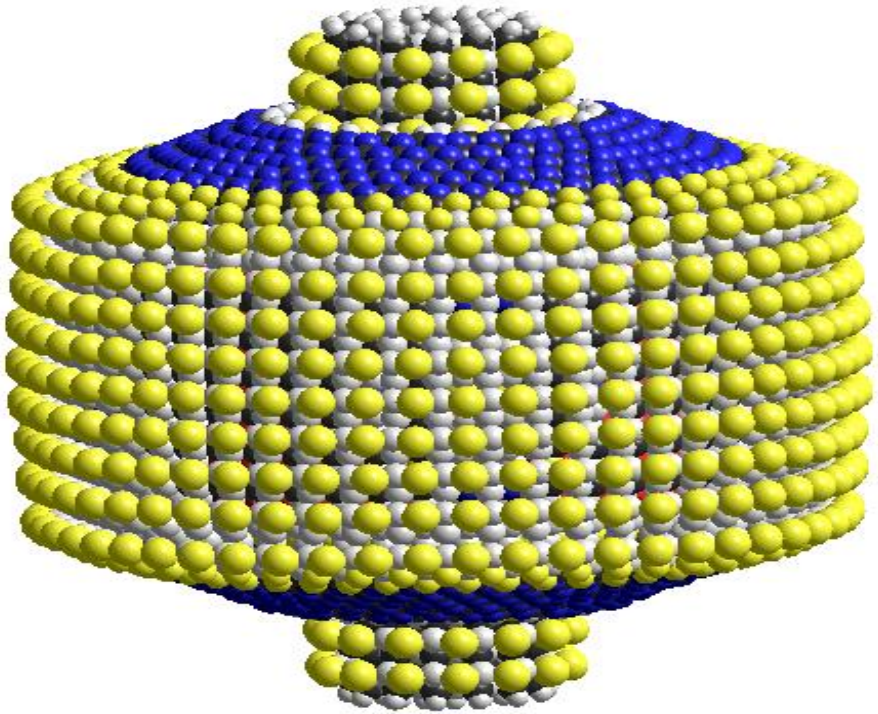
$30 \div 60 \times 10^6$ równań nielin.



Nano-Technologie

Michał Wróbel, Aleksander Herman
TASK & Politechnika Gdańska

XMD testing target:
a planetary gear device
containing 8297 atoms
(C, F, H, N, O, P, S and Si)
designed by
K. E. Drexler and R. Merkle



- XMD - pakiet oprogramowania *Open Source* do symulacji zagadnień dynamiki molekularnej dla nano-urządzeń oraz nano-systemów

Projekt ClusteriX-II: system obliczeń kampusowych

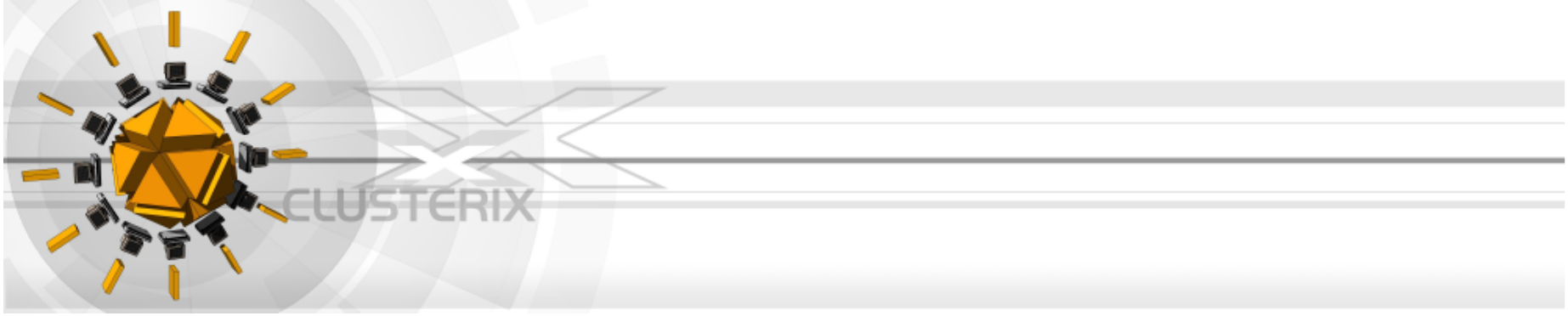
- system klastrowo-gridowy dynamicznie skalowalny:
  - ewolucja z architektury ClusteriX do ClusteriX-II
  - zarządzanie zadaniami, zasobami i użytkownikami w dynamicznym środowisku gridowym
- platforma sprzętowa: procesory wielordzeniowe (4 rdzenie i więcej) wsparte akceleratorami sprzętowymi (w niektórych ośrodkach), co zapewni przyspieszenie obliczeń w węzłach klastra
- zaawansowane partycjonowanie infrastruktury gridowej:
  - grid produkcyjny/grid badawczy/...
  - możliwość wykorzystania różnego middleware
  - zastosowanie technik wirtualizacji

Projekt ClusteriX-II: (2)

- wybór bazowego middleware gridowego:
  - Globus Toolkit 4 ?
- jak najszersze wykorzystanie komponentów middleware opracowanych w projekcie celowym ClusteriX:
- architektura orientowana na usługi (koncepcja SOA)
- gwarantowana jakość usług obliczeniowych
- integracja ze środowiskiem Eclipse:
  - wsparcie tworzenia aplikacji, debugowania, testowania

Projekt ClusteriX-II: (3)

- wsparcie dla zaawansowanych aplikacji rozproszonych
  - dodanie workflows, dynamizm workflows
  - praca interaktywna
- dodatkowe usługi dla społeczności naukowej:
  - usługa zaawansowanej wizualizacji
  - udostępnienie usługi składowania danych
- wsparcie dla uruchamiania aplikacji komercyjnych
  - problem licencjonowania oprogramowania w środowisku rozproszonym
- stworzenie środowiska testującego na bieżąco dostępność i poprawność funkcjonowania infrastruktury sprzętowo-programowej systemu



Dziękuję za uwagę !

https://clusterix.pl

Roman Wyrzykowski
roman@icis.pcz.pl