



Linux w rozwiązaniach klastrowych HPC

CI TASK

Michał Białoskórski, Bartosz Pilszka



HPC

HPC - High Performance Computing

Komputery dużej mocy obliczeniowych

*maszyny masywnie równoległe (MPP)
(SGI Altix, IBM BlueGene, Cray T3)*

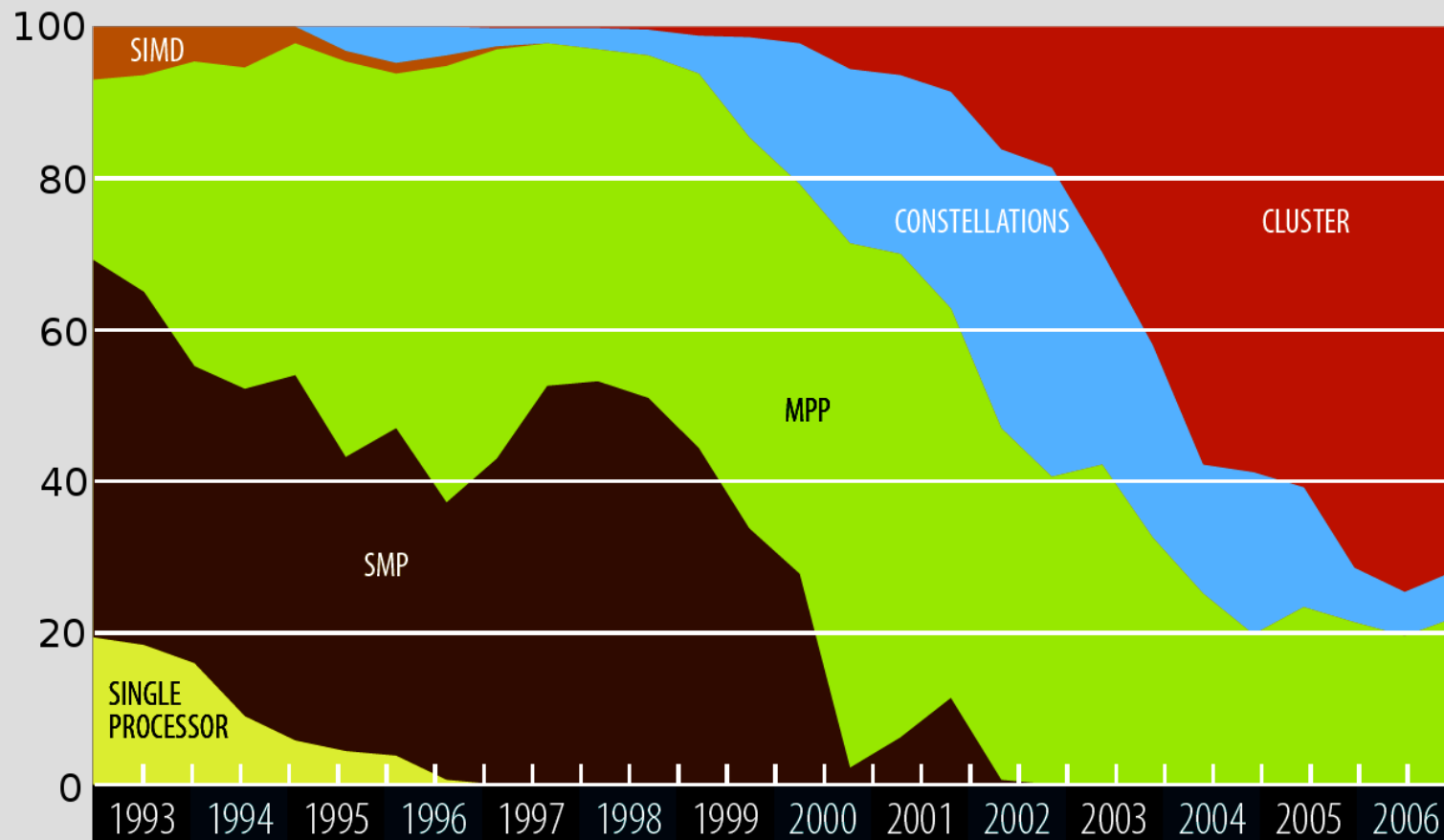
*Klasy komputerów
(PC, serwery x86/ia64, blade)*

Wydajnościowe, a nie niezawodnościowe

Klastry w HPC

www.top500.org

Architektura





Linux w HPC

www.top500.org

Operating system Family	Count	Share %
Linux	376	75.20 %
Unix	86	17.20 %
BSD Based	3	0.60 %
Mixed	32	6.40 %
Mac OS	3	0.60 %
Totals	500	100%



Linux w HPC

www.top500.org

2006/11

2000/11

Operating System	Count	Share %
Linux	53	10.60 %
HP Unix (HP-UX)	5	1.00 %
IRIX	19	3.80 %
Solaris	92	18.40 %
UNICOS	47	9.40 %
Super-UX	16	3.20 %
UXP/V	17	3.40 %
HI-UX/MPP	16	3.20 %
AIX	214	42.80 %
N/A	3	0.60 %
Tru64 UNIX	9	1.80 %
Paragon OS	1	0.20 %
EWS-UX/V	7	1.40 %
SUSE Linux	1	0.20 %
Totals	500	100%

Operating System	Count	Share %
Linux	326	65.20 %
SuSE Linux Enterprise Server 8	3	0.60 %
Redhat Enterprise 3	1	0.20 %
HP Unix (HP-UX)	27	5.40 %
MacOS X	3	0.60 %
Solaris	5	1.00 %
UNICOS	8	1.60 %
Super-UX	3	0.60 %
AIX	43	8.60 %
Tru64 UNIX	3	0.60 %
SuSE Linux Enterprise Server 9	25	5.00 %
UNICOS/Linux	2	0.40 %
CNK/SLES 9	27	5.40 %
SUSE Linux	3	0.60 %
Redhat Linux	4	0.80 %
RedHat Enterprise 4	7	1.40 %
UNICOS/SUSE Linux	3	0.60 %
SUSE Linux Enterprise Server 10	2	0.40 %
SLES10 + SGI ProPack 5	5	1.00 %
Totals	500	100%



Największe: TOP500

- **Mare Nostrum**: BladeCenter Cluster, PPC 970, Myrinet, Nproc=10240, Rpeak= 94TFlops, **5**
- **Thunderbird**: PowerEdge 1850, 3.6 GHz, InfiniBand, Nproc=9024, Rpeak=65TFlops, **6**
- **TSUBAME**: Sun Fire x4600 Cluster, Opteron ClearSpeed Accelerator, InfiniBand, Nproc=11088, Rpeak=82TFlops, **9**

HPC w Polsce

- wykorzystywane do celów naukowych
- 8 największych maszyn to klastry
- kolejne maszyny to SGI Altix i Cray
- zlokalizowane w 5 ośrodkach KDM:
Warszawa, Poznań, Kraków, Wrocław,
Gdańsk
- ok. 50 mniejszych klastrów, budowanych
przez małe grupy

Te duże

- **Holk (CI TASK)** - moc: 3,2 TFlops, 288 procesorów Itanium2 Dual Core, InfiniBand
- **Nova (WCSS)** - moc: 2,9 Tflops, 152 procesory Xeon, Dual Core, dwie płyty gł. w jednej obudowie
- **Zeus (Cyfronet)** - moc 2 Tflops, 384 procesory w czterech generacjach od PIII do Xeon 2,8 GHz

Typowy klaster naukowy w Polsce



- 5-20 węzłów - „dobry PC”
- węzły SMP - dwuprocesorowe
- często heterogeniczny
- kilka etapów rozwoju - od PIII do DualCore
- medium - FastEthernet, GigabitEthernet
- OS - Linux
- dodatki - MPICH, PBS/Torque, Ganglia

Clusterix

- Krajowy Klaster Linuksowy, 2004
- 12 ośrodków, koordynator: Politechnika Częstochowska
- na bazie krajowej sieci Pionier
- integracja rozproszonych małych klastrów w dynamiczny grid obliczeniowy
- rozwój oprogramowania gridowego
- lepsze wykorzystanie zasobów

Zalety Linuksa

- dostępność wielu aplikacji naukowych często Linux jest wspólnym mianownikiem aplikacji dostępnych na różne platformy,
- jednolite środowisko użytkownika na serwerze i stacji roboczej (w pracy, w domu) np. dla programisty
- elastyczność, łatwość dostosowania do specyficznych potrzeb użytkownika i administratora

Zalety Linuksa c.d.

- rozwój sterowników i oprogramowania do sieci klastrowych typ InfiniBand, często odbywa się na Linuksie
- duży wybór dystrybucji

Czemu akurat klastry?

- zapotrzebowanie -> decyzja / cena -> -> zakup klastra
- większość oprogramowania przystosowuje się do obliczeń rozproszonych
- jeżeli oprogramowanie nie da się przystosować, to przystosować musi się sam użytkownik

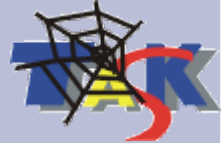


Zalety Linuxa w HPC

- takie jak w innych rozwiązaniach
- dużo informacji w sieci
- łatwość rozwiązywania problemów
- darmowe oprogramowanie
- darmowy support
- szeroka gama oprogramowania

Historia klastrów w TASKu

- Galera(2000r.) -128 x PIII Xeon, SCI
- TurboLinux – „hasłowe” wsparcie dla HPC
- próby z Solarisem 8 – brak zdecydowanej przewagi nad Linuksem (2.4)
- Debian – najbardziej elastyczna dystrybucja



TASK - Holk

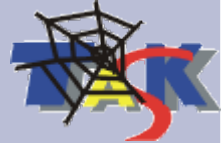
- Holk (2003r.) - 256 Itanium2
- Debian jako jedyna sensowna dystrybucja na ia64 - darmowa, pełna funkcjonalność, duże repozytorium pakietów



Dlaczego Debian

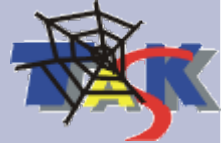
W **2003** jedyny darmowy dobrze rozwinięty Linux na ia64

Duże repozytorium pakietów



Instalacja klastra

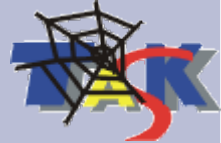
- 1 - instalacja serwera dostępowego
- 2 - instalacja wzorcowego węzła
- 3 - klonowanie



Automatyczna instalacja

Mechanizmy automatycznej instalacji OS zawarte w dystrybucjach:

- Kickstart - RedHat
- FAI - Fully Automatic Installation - Debian



Zarządzanie sprzętem

- setki, tysiące węzłów
- zarządzanie niezależnie od systemu operacyjnego
- wymagana możliwość zdalnego włączania/wyłączania węzłów
- monitorowanie podstawowych parametrów pracy: temperatura, napięcia



Zarządzanie - rozwiązania

- **IPMI** -Intelligent Platform Management Interface,
- **BMC** -Baseboard Management Controllers,
- zarządzalne listwy zasilające
- bootowanie po sieci
- grub, elilo, EFI, OpenBIOS
- ganglia
- **własne skrypty**

Zarządzanie kontami

- specyfika użytkowników HPC:
- stałe dane: (login, dane osobowe, uid, gid)
- stała liczba: ruch ~10%

Użytkownicy

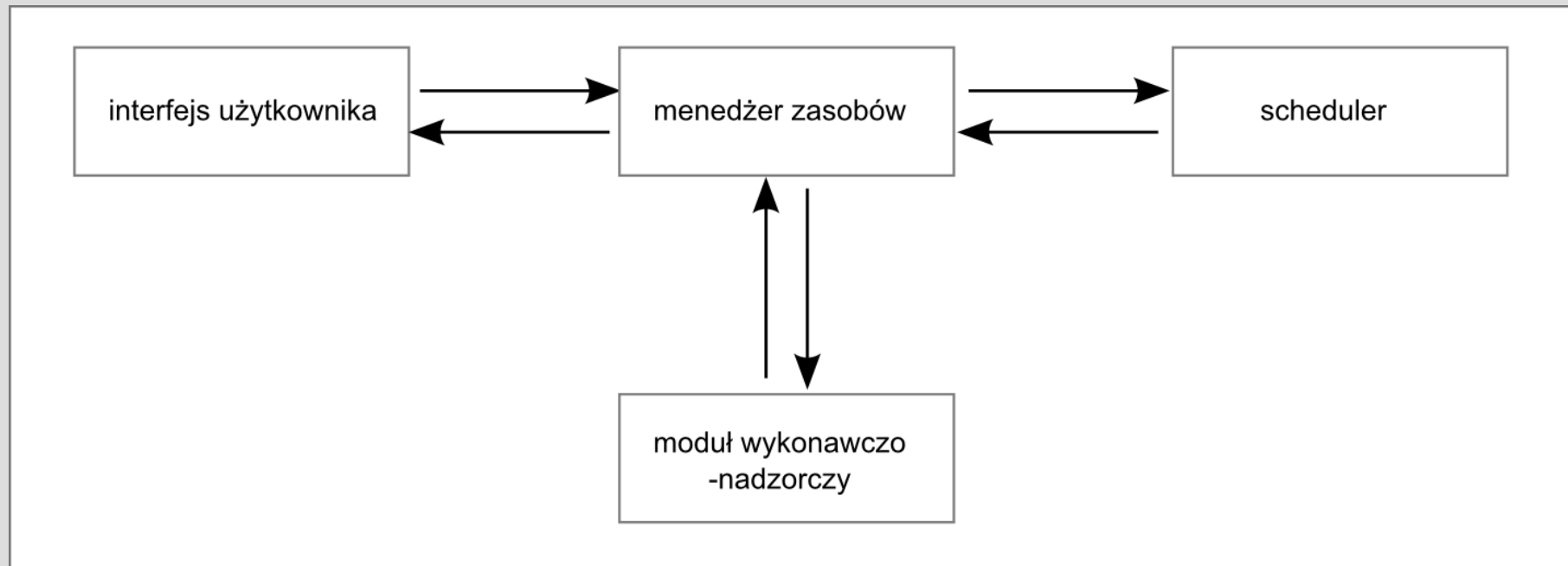
- specyficzni – środowisko naukowe
- pierwszy kontakt z Linuksem/Unixem – męczarnia ;)
- użytkownicy aplikacji naukowych
- zaawansowani programiści – (bardziej rozgarnięci niż administratorzy? ;))
- różne możliwości, różne potrzeby
- względnie wyrozumiali

System kolejkowy



- zarządzanie zasobami obliczeniowymi
- monitoring klastra
- przydzielanie zasobów
- interfejs użytkownika,
- rozliczanie wykorzystania zasobów

Schemat systemu kolejkowego

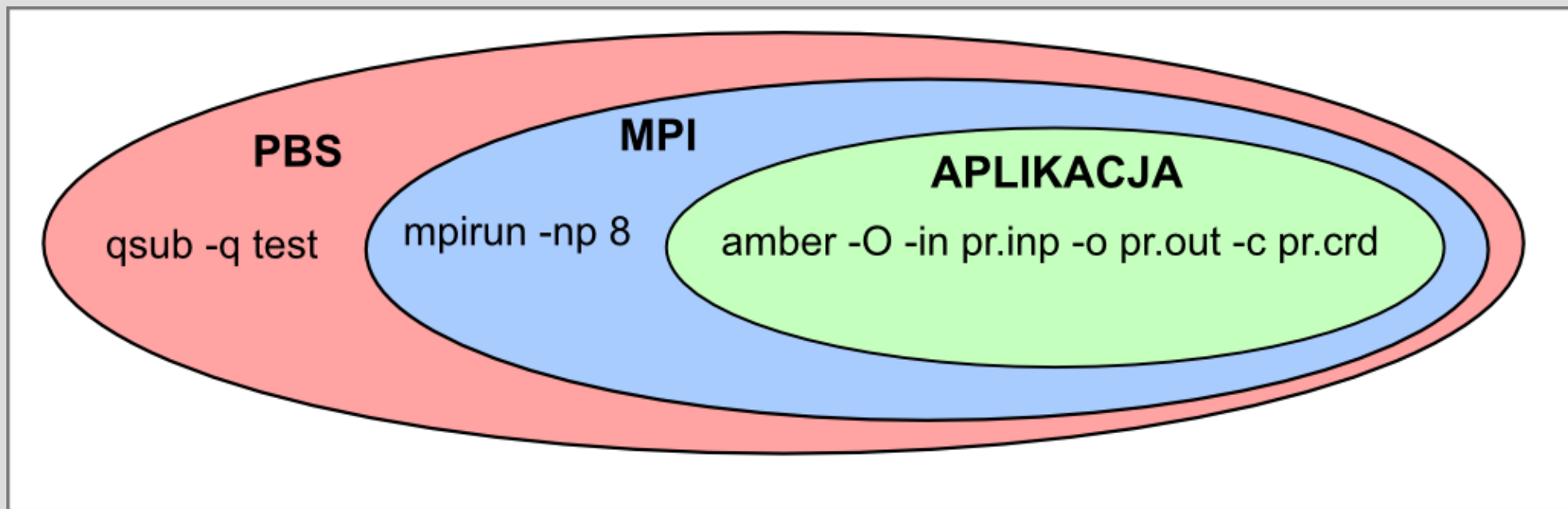




Konfigurowanie kolejek

- swobodę w dostępie do zasobów
- jak najlepsze wykorzystanie klastra
- sprawiedliwość
- uwzględnienie użytkowników lub zadań o specjalnych przywilejach

Użytkowanie



MPI a medium

- Sieć standardowa

- FastEthernet
- GigabitEthernet

Opóźnienie >20ms

Przepustowość <1Gb

- Sieć dedykowana

- Quadrics
- Mirynet
- SCI

Opóźnienie < 7 μ s

Przepustowość >1Gb

- **InfniBand**

IB - 4 μ s i 10Gb

Medium w MPI a aplikacja

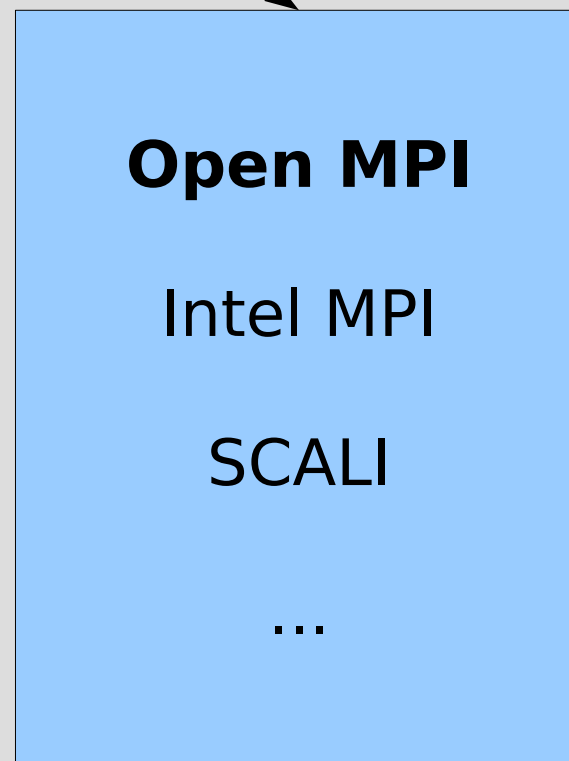
- Zależna od medium
 - w programie wbudowane są mechanizmy obsługujące dane medium
 - wydajność kosztem wygody
- Niezależna od medium
 - zależność od medium przeniesiona na warstwę systemu (biblioteki MPI)
 - wygoda kosztem wydajności (-10%)

Medium w MPI a aplikacja

Zależna od
medium

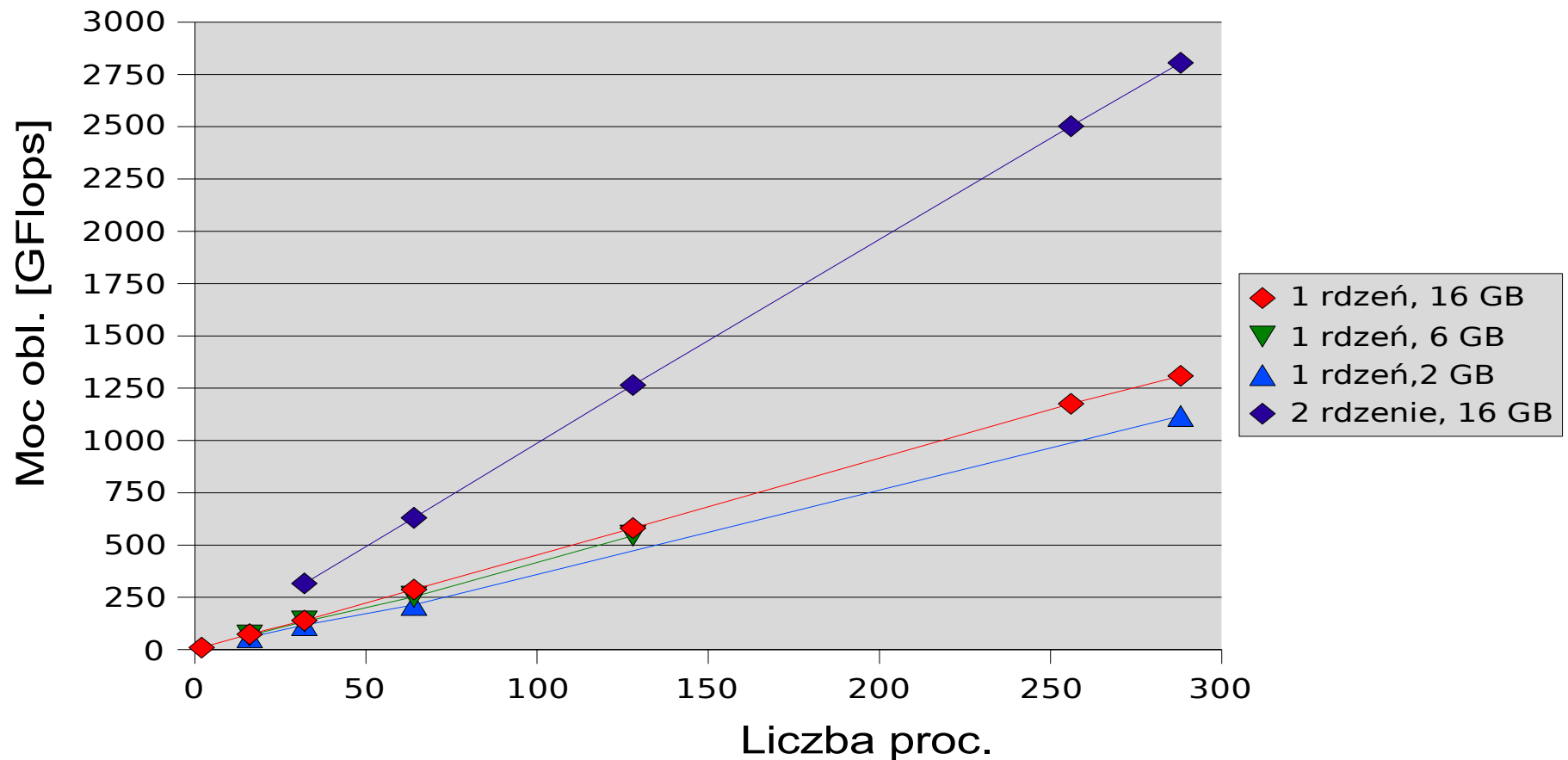


Niezależna od
medium



HPL w zależności od rozmiaru pamięci

Skalowalność testu HPL

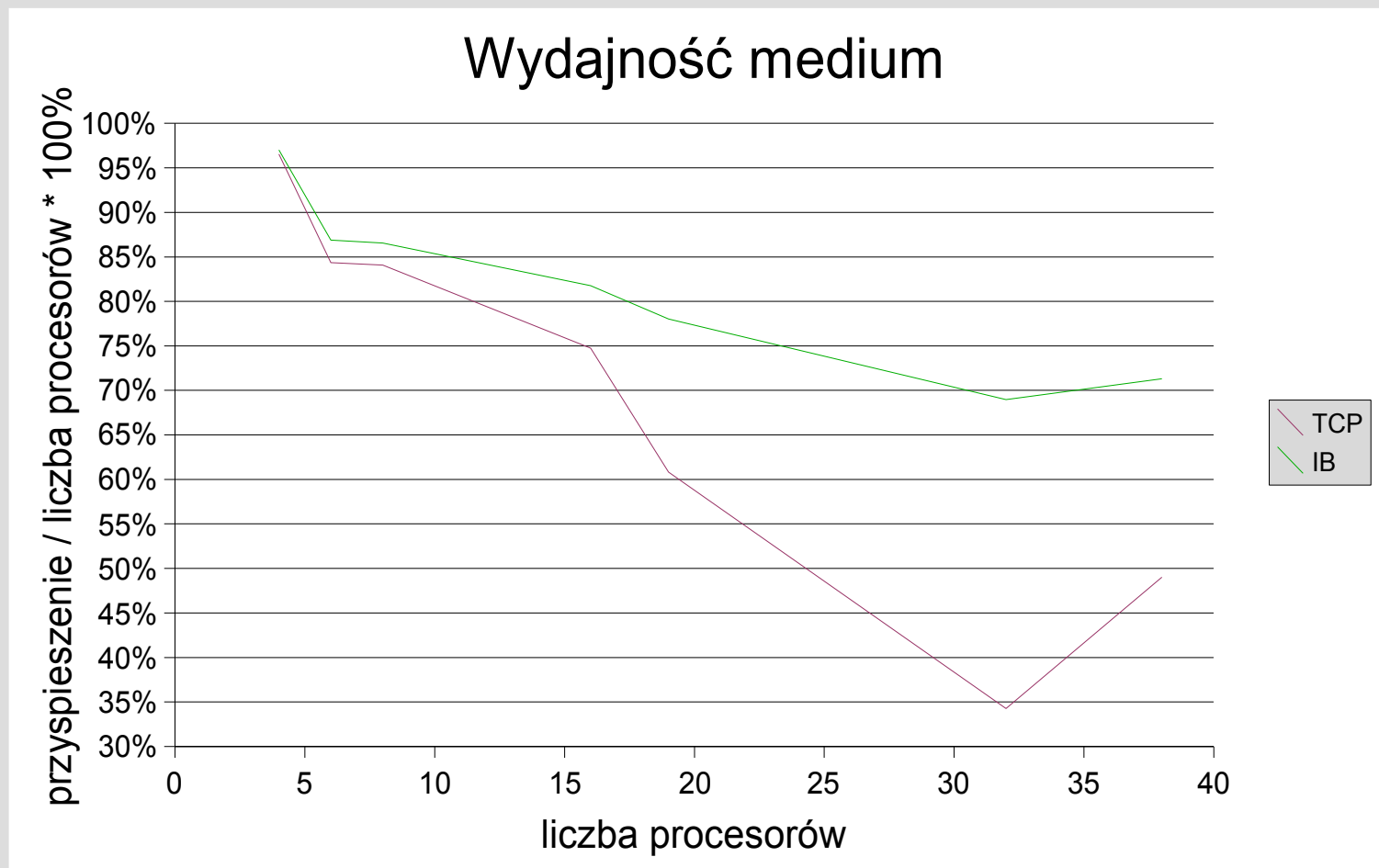


Źródło: raport wewnętrzny CI TASK

HPL w zależności od medium

<i>sieć</i>	<i>R[GFlop/s]</i>	<i>%Rpeak</i>
Gigabit Ethernet	641,6	48,2
InfiniBand	934,7	70,2

Wyniki nanoMD



Źródło: raport wewnętrzny CI TASK



„Tuningowanie”

- w zależności od potrzeb aplikacji może być konieczne dostosowanie parametrów systemu
- kernel - sysctl shmall, shmax, sem
- parametry interfejsów sieciowych
- wybór i parametry systemu plików np. bez księgowania na tymczasowy
- parametry NFS

Rozwiązania klastrowe

- SystemImager lub podobne
- System kolejkowy
- Ganglię
- MPI
- mechanizmy zarządzania

Gotowe rozwiązania klastrowe



- OSCAR Open Source Cluster Application Resources
- xCAT - Extreme Cluster Administration Toolkit, IBM
- ROCKS Cluster Distribution
- Scyld ClusterWare HPC

Źródła

- <http://top500.org/>
- <http://pliszka.net/hpc/>
- <http://www.klastry.org.pl/>
- <http://www.intel.com/>
- <http://www.beowulf.org/>
- <http://www.task.gda.pl/>
- <http://clusterix.pl/>
- <http://sisuite.org/>
- <http://oscar.openclustergroup.org/>
- <http://www.open-mpi.org/>
- <http://www.clusterresources.com/>