Akademickie Centrum Komputerowe

Cyfronet AGH

Komputery Dużej Mocy w Cyfronecie

Andrzej Oziębło

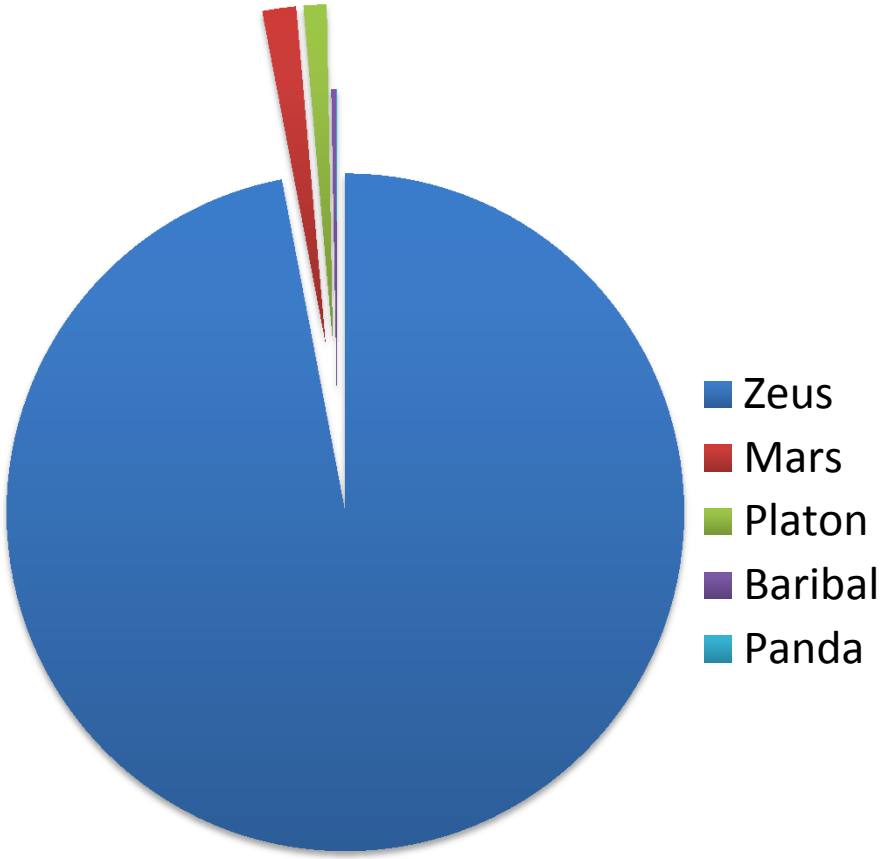
Patryk Lasoń, Łukasz Flis, Marek Magryś

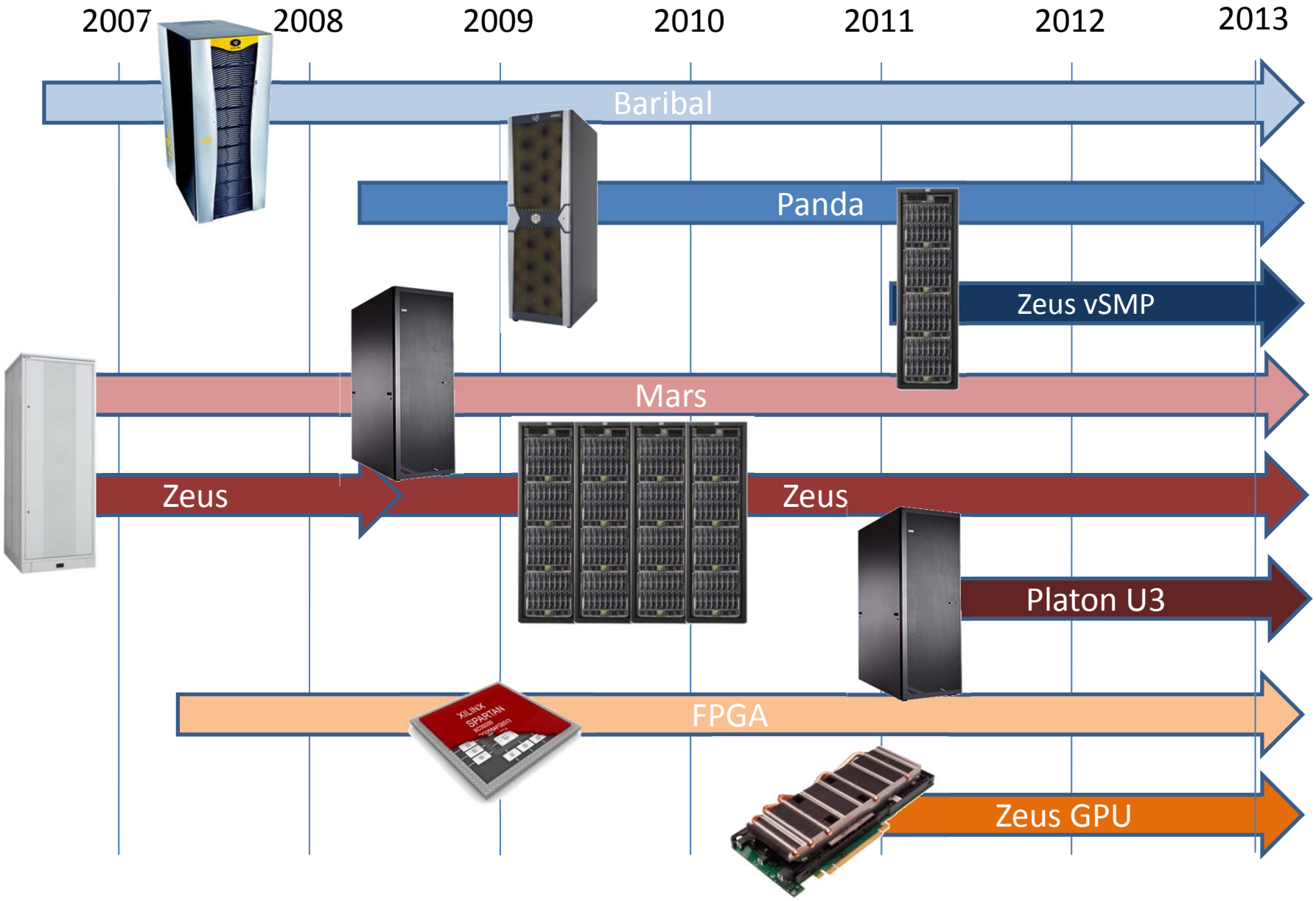
Administratorzy KDM

- Baribal, Mars, Panda, Platon U3:
  - **Stefan Świąć**
  - **Piotr Wyrostek**
- Zeus:
  - **Łukasz Flis**
  - **Patryk Lasoń**
  - **Marek Magryś**
  - **Jacek Budzowski**

Zasoby obliczeniowe

- Baribal, Mars, Panda, Platon U3
- Zeus:
  - Klaster
  - GPGPU
  - vSMP
  - BigMem

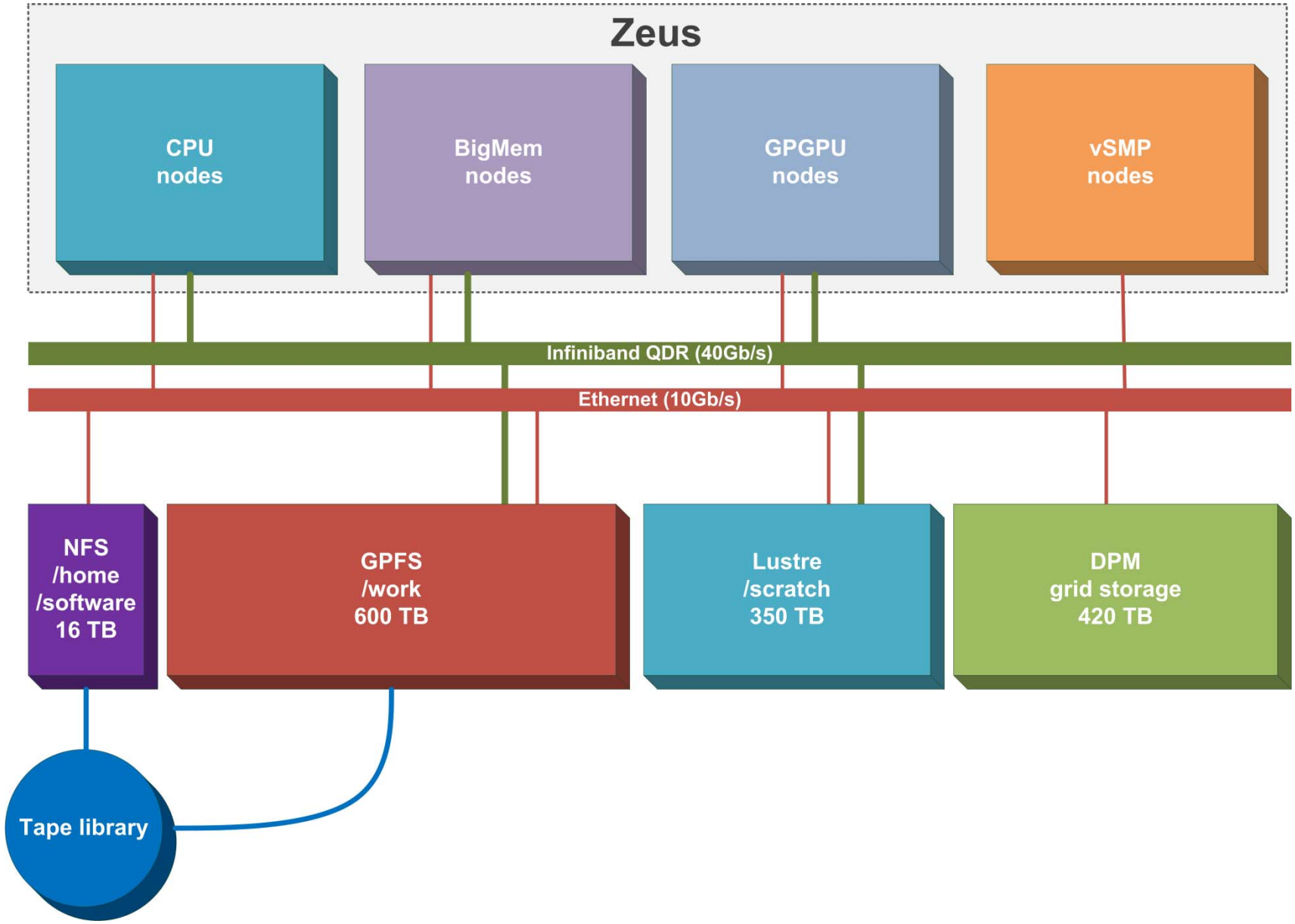


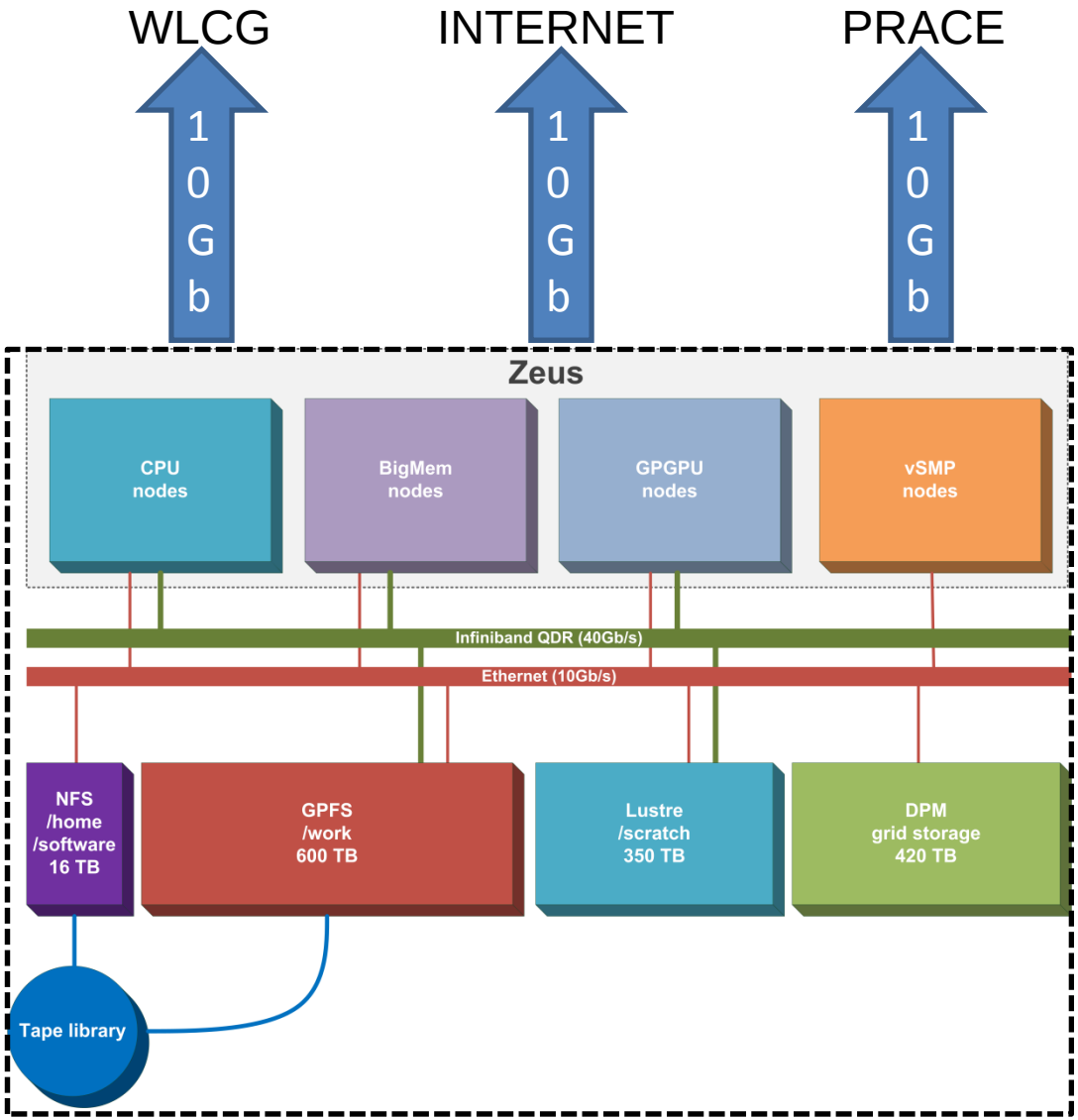

2007
2008
2009
2010
2011
2012
2013
Baribal
Panda
Zeus vSMP
Mars
Zeus
Zeus
Platon U3
FPGA
Zeus GPU


CYFRONET
Akademickie
Centrum
Komputerowe
CYFRONET AGH
www.cyfronet.krakow.pl
ZEUS

Zeus

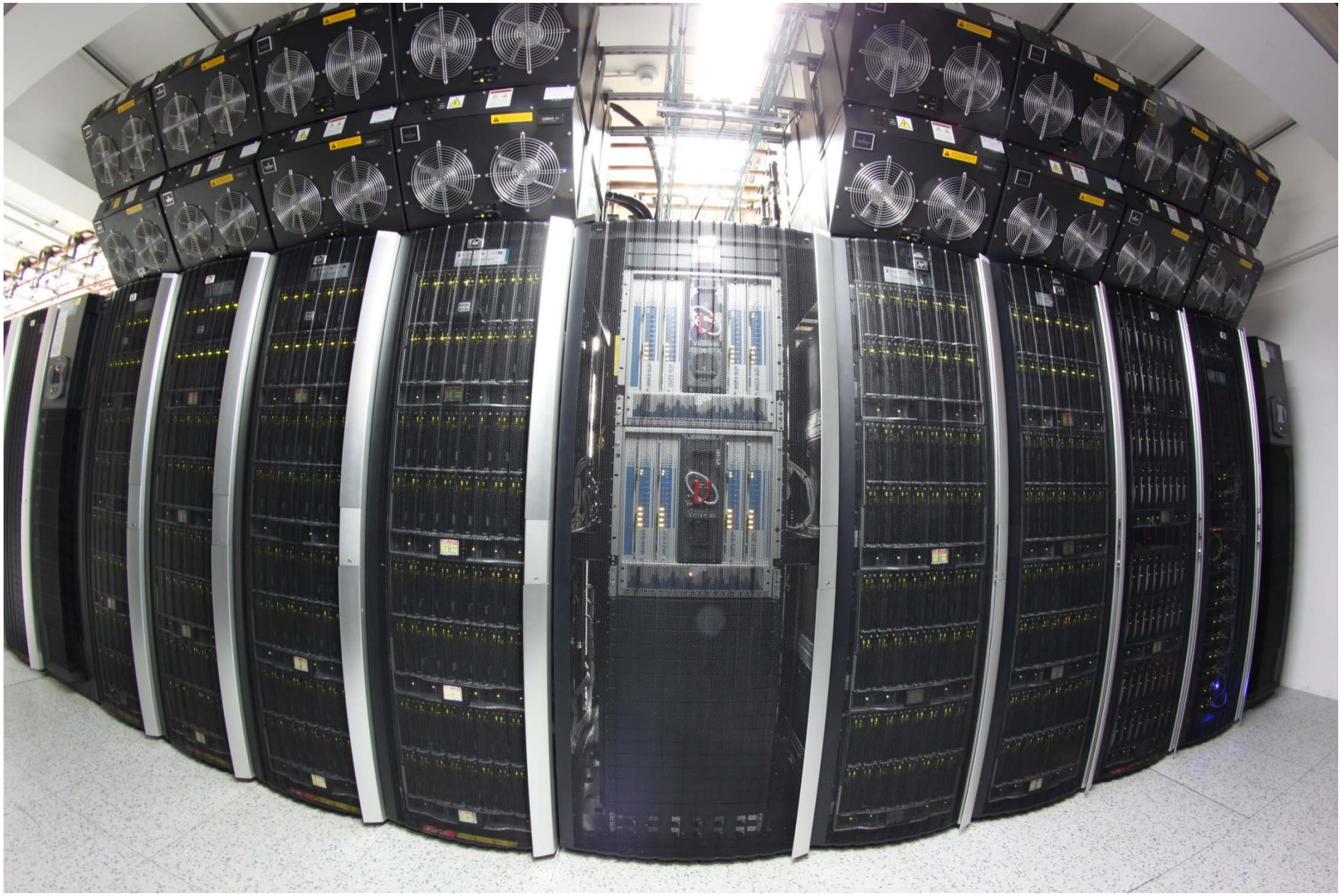
- 374 Tflops
- ponad 20 tys. rdzeni
- ponad 200 kart GPGPU
- 60 TB RAM
- **#1** w Polsce
- #145 na liście Top500


Zeus
CPU nodes
BigMem nodes
GPGPU nodes
vSMP nodes
Infiniband QDR (40Gb/s)
Ethernet (10Gb/s)
NFS /home /software 16 TB
GPFS /work 600 TB
Lustre /scratch 350 TB
DPM grid storage 420 TB
Tape library


WLCG
INTERNET
PRACE
10Gb
10Gb
10Gb
Zeus
CPU nodes
BigMem nodes
GPGPU nodes
vSMP nodes
Infiniband QDR (40Gb/s)
Ethernet (10Gb/s)
NFS /home /software 16 TB
GPFS /work 600 TB
Lustre /scratch 350 TB
DPM grid storage 420 TB
Tape library

Zeus – klaster

- HP BL2x220c – ponad 1100 węzłów
- 12000 rdzeni Intel Xeon
- 22TB RAM
- Interconnect:
  - 1GbE per węzeł, 10GbE switch-switch
  - Infiniband 4x QDR (40Gbps)
- Pamięć dyskowa: 1.8PB
- 120 Tflops

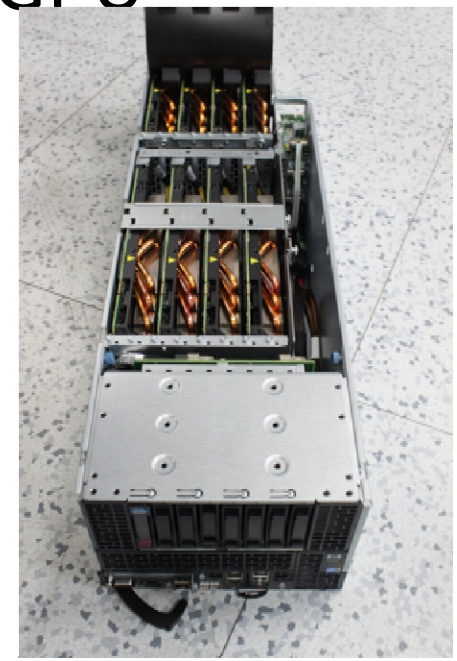


Zeus – BigMem

- HP BL685c G7 - **104 węzły**
- **6656** rdzeni AMD Opteron
- **26 TB** RAM
- 64 rdzeni, 256GB RAM **per węzeł**
- Interconnect:
  - 10GbE per węzeł
  - Infiniband: 4x QDR (**40Gbps**)
- 61 Tflops

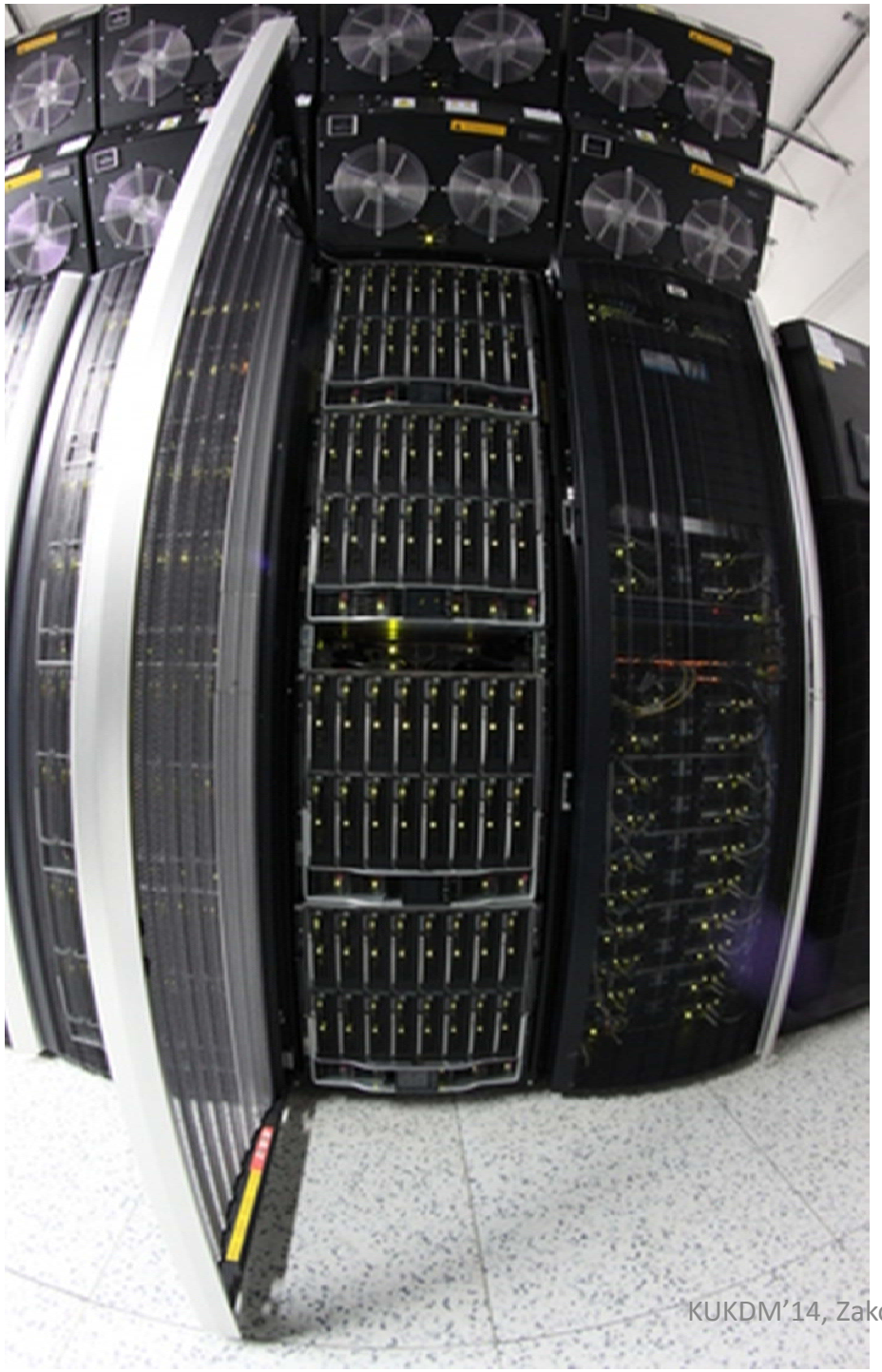
Zeus – GPGPU

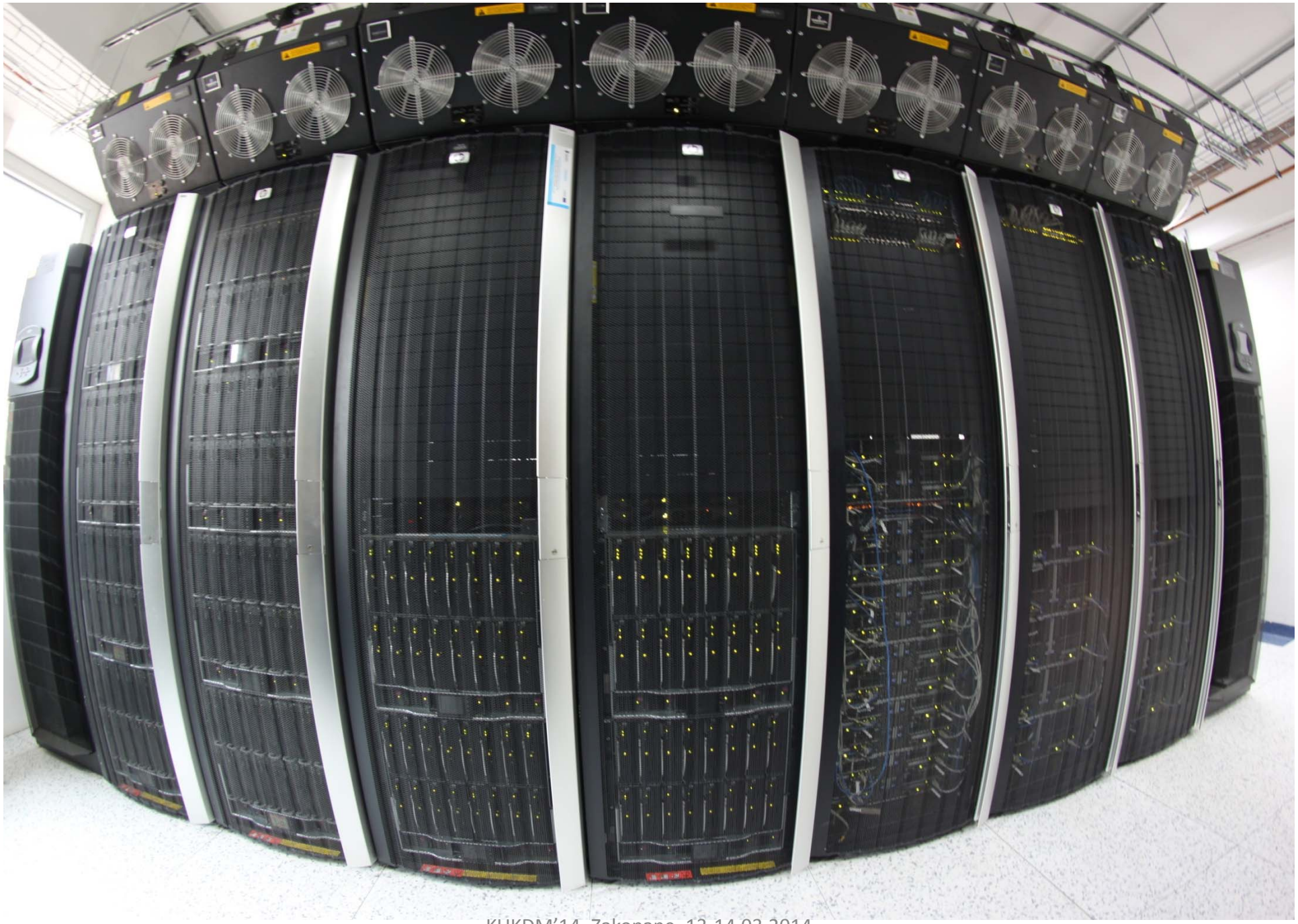
- HP SL390s - **44 węzły**
- **528** rdzeni Intel Xeon
- **3,6 TB** RAM
- **208** kart NVIDIA M2050/M2090 GPGPU
- Interconnect:
  - 1GbE per węzeł
  - Infiniband: 4x QDR (**40Gbps**)
- 136 Tflops

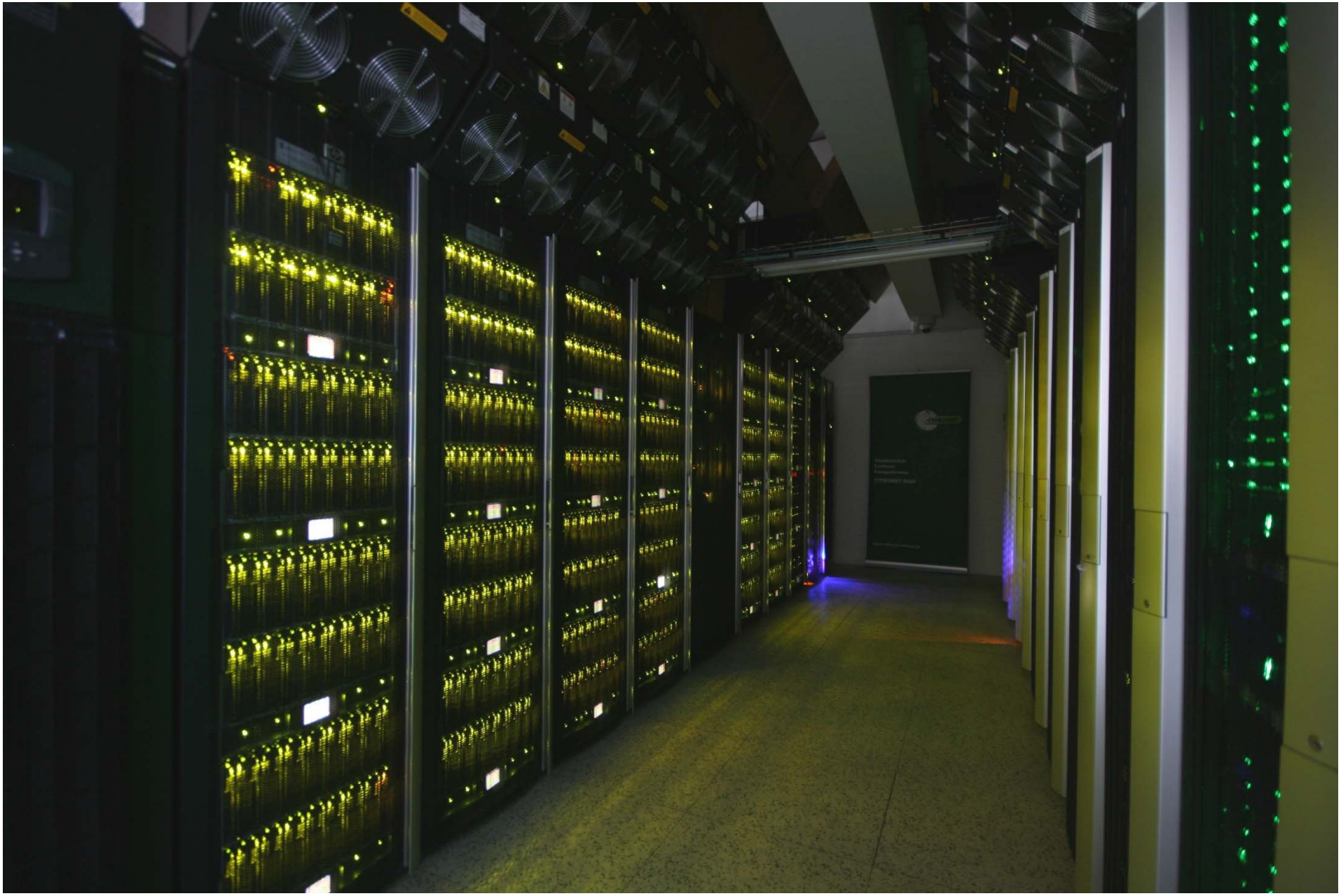


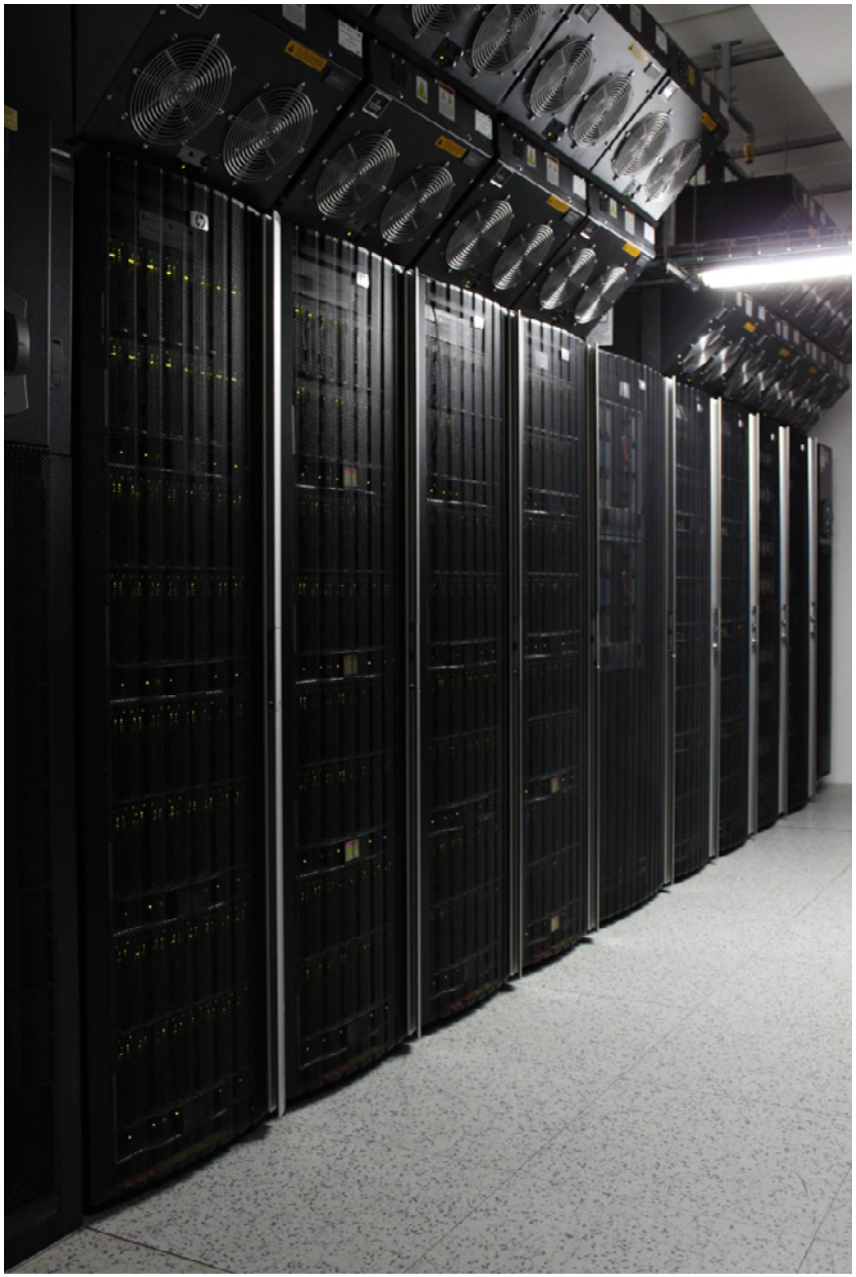
Zeus – vSMP

- HP BL490c - **64 serwery**
- **ScaleMP vSMP**
- **768 rdzeni** Intel Xeon
- **6 TB** RAM
- Interconnect:
  - Infiniband: dual-rail Infiniband QDR 40Gbps
- **Single System Image**
- 8 Tflops










VOLTAIRE
VOLTAIRE

Ludzie

- Administracja:
  - Patryk Lasoń
  - Łukasz Flis
  - Marek Magryś
- Programowanie:
  - Jacek Budzowski
  - Maciej Golik
  - Maciej Pawlik
- Oprogramowanie:
  - Maciej Czuchry
  - Mariusz Sterzel
  - Klemens Noga
- 1st line support
- I inni

Zeus w 2013

- Niemal 8 mln zadań -> ponad 21 tys. dziennie
- 96 mln godzin CPU
- Ponad 1100 aktywnych użytkowników
- Ponad 100 PB ruchu na /scratch
- Najdłuższe zadanie: 90 dni
- Największe zadanie: 4096 rdzeni
- Ponad 70% czasu CPU to zadania wieloprocesorowe
  - w tym prawie 80% to zadania wielowęzłowe

Pierwsza pomoc

- KDM wiki http://kdm.cyfronet.pl
- Podręcznik Użytkownika PL-Grid
- Helpdesk
- zeus@cyfronet.pl
- +48 12 632 33 55 w. 107
- +48 12 632 33 55
- Ul. Nawojki 11

Zeus-toolkit

Skrypty:

- zeus-jobs
- zeus-jobs-history
- zeus-gpus
- zeus-fs
- zeus-*grant* (aliasy dla plg-*grant*)

zeus-jobs

Note	ID	Queue	Name	State	Nodes	Cores	GPUs	Decl. mem	Mem usage	Efficiency	Walltime
-	26573086	l_exclusive	y1.sh	R	1	12	0	18.0G	0.0%	100.0%	00:00:00
-	26573087	l_exclusive	y2.sh	R	1	12	0	18.0G	76.7%	97.6%	202:22:01
-	26573088	l_exclusive	y3.sh	R	1	12	0	18.0G	77.1%	91.9%	202:22:01
-	26573089	l_exclusive	y4.sh	R	1	12	0	18.0G	76.7%	97.7%	202:22:01
-	26573090	l_exclusive	y5.sh	R	1	12	0	18.0G	76.7%	95.9%	202:22:01
-	26573091	l_exclusive	y6.sh	R	1	12	0	18.0G	76.7%	95.7%	202:21:57
-	26573113	l_exclusive	y.sh	R	1	12	0	18.0G	75.0%	96.4%	202:21:57
C	26615992	l_exclusive	x1837.04.sh	R	1	12	0	18.0G	103.5%	45.1%	178:30:57
C	26615993	l_exclusive	x1837.06.sh	R	1	12	0	18.0G	103.0%	43.9%	178:30:44
C	26615994	l_exclusive	x1837.08.sh	R	1	12	0	18.0G	103.1%	39.4%	178:29:46
C	26615995	l_exclusive	x1837.10.sh	R	1	12	0	18.0G	102.8%	41.0%	178:30:36
C	26615996	l_exclusive	x1837.12.sh	R	1	12	0	18.0G	102.4%	37.1%	178:30:57
-	26665187	l_exclusive	x1837.00.sh	R	1	12	0	18.0G	0.0%	100.0%	00:00:00
C	26665189	l_exclusive	x1837.14.sh	R	1	12	0	18.0G	90.9%	1.0%	105:17:09
C	26665190	l_exclusive	x1837.16.sh	R	1	12	0	18.0G	101.5%	68.2%	99:19:19
C	26665191	l_exclusive	x1837.18.sh	R	1	12	0	18.0G	89.9%	1.5%	98:09:31
C	26665192	l_exclusive	x1837.20.sh	R	1	12	0	18.0G	89.3%	2.1%	98:09:41
C	26665193	l_exclusive	x1837.22.sh	R	1	12	0	18.0G	88.7%	2.0%	97:28:16
-	26665194	l_exclusive	x1837.24.sh	R	1	12	0	18.0G	99.0%	67.6%	96:00:21
-	26665195	l_exclusive	x1837.26.sh	R	1	12	0	18.0G	98.2%	67.7%	95:43:37
C	26665196	l_exclusive	x1837.28.sh	R	1	12	0	18.0G	86.0%	3.4%	93:59:17
C	26665197	l_exclusive	x1837.30.sh	R	1	12	0	18.0G	96.6%	4.3%	91:34:52
C	26665198	l_exclusive	x1837.32.sh	R	1	12	0	18.0G	95.7%	7.8%	91:15:17
C	26665199	l_exclusive	x1837.34.sh	R	1	12	0	18.0G	94.7%	8.7%	91:14:08
C	26665200	l_exclusive	x1837.36.sh	R	1	12	0	18.0G	93.7%	11.2%	90:49:16
C	26665201	l_exclusive	x1837.38.sh	R	1	12	0	18.0G	92.6%	11.4%	90:47:43
C	26695701	l_exclusive	x1837.02.sh	R	1	12	0	18.0G	92.4%	1.3%	44:02:03
-	26715566	l_exclusive	x1837.40.sh	R	1	12	0	18.0G	91.5%	76.5%	15:03:49

zeus-jobs-history

ID	Queue	Name	Nodes	Cores	Decl. mem	Mem. usage	Efficiency	Walltime Used	Walltime Req.	Exit Status	End Time
26665202	l_exclusive	x1837.40.sh	1	12	18.0GiB	16.5GiB	81.0%	13:49:45	336:00:00	38	2013-02-25 07:36:26
26664880	l_exclusive	x1239.00.sh	1	12	18.0GiB	3.6GiB	85.7%	00:40:32	336:00:00	271	2013-02-24 02:11:00
26664813	l_exclusive	x1239.00.sh	1	12	18.0GiB	3.6GiB	68.7%	00:01:13	336:00:00	271	2013-02-24 01:24:28
26597712	l_exclusive	x1226.24.sh	1	12	18.0GiB	2.2GiB	79.9%	49:57:02	336:00:00	0	2013-02-22 15:16:45
26597706	l_exclusive	x1226.06.sh	1	12	18.0GiB	2.6GiB	80.3%	43:58:20	336:00:00	0	2013-02-22 09:18:03
26597709	l_exclusive	x1226.10.sh	1	12	18.0GiB	2.5GiB	82.9%	37:20:48	336:00:00	0	2013-02-22 02:40:31
26597679	l_exclusive	x1027.06.sh	1	12	18.0GiB	1.9GiB	80.7%	34:28:09	336:00:00	0	2013-02-21 23:46:04
26597710	l_exclusive	x1226.14.sh	1	12	18.0GiB	2.5GiB	78.5%	30:26:53	336:00:00	0	2013-02-21 19:46:36
26597777	l_exclusive	x1027.04.sh	1	12	18.0GiB	1.9GiB	80.5%	26:38:31	336:00:00	0	2013-02-21 16:01:59
26597683	l_exclusive	x1027.10.sh	1	12	18.0GiB	1.9GiB	80.3%	20:49:31	336:00:00	0	2013-02-21 10:07:26
26597707	l_exclusive	x1226.08.sh	1	12	18.0GiB	2.6GiB	85.4%	16:11:04	336:00:00	0	2013-02-21 05:30:47
26597705	l_exclusive	x1226.04.sh	1	12	18.0GiB	2.6GiB	89.4%	11:51:50	336:00:00	0	2013-02-21 01:11:31
26573348	l_exclusive	x1027.36.sh	1	12	18.0GiB	1.4GiB	87.7%	20:21:11	336:00:00	0	2013-02-20 22:29:00
26573346	l_exclusive	x1027.18.sh	1	12	18.0GiB	1.7GiB	84.5%	18:19:16	336:00:00	0	2013-02-20 20:27:05
26597681	l_exclusive	x1027.08.sh	1	12	18.0GiB	1.9GiB	84.1%	04:10:10	336:00:00	0	2013-02-20 17:28:05
26599753	l_exclusive	x1530.24.sh	1	12	18.0GiB	497.5MiB	6.8%	00:52:10	336:00:00	0	2013-02-20 15:10:30
26573344	l_exclusive	x1027.14.sh	1	12	18.0GiB	1.8GiB	82.2%	14:23:39	336:00:00	0	2013-02-20 14:21:05
26599751	l_exclusive	x1530.16.sh	1	12	18.0GiB	506.8MiB	23.6%	00:00:54	336:00:00	39	2013-02-20 14:19:14
26599749	l_exclusive	x1530.12.sh	1	12	18.0GiB	0B	38.9%	00:00:51	336:00:00	39	2013-02-20 14:19:11
26599755	l_exclusive	x1530.30.sh	1	12	18.0GiB	453.8MiB	19.6%	00:00:43	336:00:00	39	2013-02-20 14:19:03

zeus-gpus

```
[ymlason@zeus ~]$ zeus-gpus
Total GPUs in cluster:    150
Total GPUs available:     114

GPUs used:     36
GPUs queued:   0
```

zeus-fs

Name	Location	Filesystem	Limit	Used space	%QuotaUsage
$HOME	/people/ymlason	NFS	7.00GB	5.39GB	77.1%
$STORAGE	/mnt/gpfs/work/people/ymlason	GPFS	100.00GB	32.99GB	33.0%

zeus-*grant*

```
[plglason@zeus ~]$ zeus-show-grants

Your active PL-Grid grants on THIS site:
+-------------------+------------+------------+--------------------+--------------------+
| GrantID           | Start Date |  End Date  | Total Walltime [h] | Total Storage [GB] |
+-------------------+------------+------------+--------------------+--------------------+
| plglason2014a (*) | 2014-02-01 | 2014-08-01 |               1000 |                 40 |
+-------------------+------------+------------+--------------------+--------------------+
* default grant

[plglason@zeus ~]$ zeus-show-default-grant

Your default grant information:

Grant ID : plglason2014a
Status   : ACTIVE (on THIS site)

[plglason@zeus ~]$ zeus-show-grant-details plglason2014a

Checking details of grant with id: plglason2014a

name      : plglason2014a
start     : 2014-02-01
end       : 2014-08-01
walltime  : 1000
storage   : 40
group     : plglason
users     : PERSONAL_GRANT

[plglason@zeus ~]$
```

Nowości

- Pierwsza (!) duża przerwa serwisowa (aktualizacja sprzętowych serwerów NFS)
- Migracja do Scientific Linux 6 (OS, rekompilacja aplikacji, tuning OS)
- Uruchomienie GPFS
  - Przeniesienie katalogów zespołów na GPFS
- Uruchomienie BigMem
- Elementy vSMP jako serwery obliczeniowe
- Zmiana nazwy maszyny dostępowej: ui -> zeus

Nowości

- Zapobieganie fragmentacji zasobów – zadania pełnowęzłowe
  - Zmniejszenie czasu oczekiwania na start dużych zadań
- Automatyczne czyszczenie przestrzeni „scratch”
- Zwiększenie przepustowości „na świat”
  1->10 Gb/s

Nowości w Zeus-toolkit

- Zwiększenie szybkości działania
- Sortowanie ilości CPU, walltime
- Filtrowanie statusu zakończenia zadań
- Defniowanie wyświetlanych kolumn
- Opcja „--changelog”

Modules

- „Automagiczne” załadowanie modułu w zależności od serwera obliczeniowego
- Sprawdzanie dostępu do licencjonowanych aplikacji
- Porządek w gałęziach (apps, libs, compilers, tools)

Plany z KUKDM'13

- Scientific Linux 6 (limity pamięciowe)
- Uruchomienie BigMem
- GPFS-HSM
- G(UI)
- Więcej filtrów w systemie kolejkowym
- Wdrożenie knem dla MPI
- Szybszy 'ls'

Plany z KUKDM'13

- Wykrywanie 'martwych' zadań
- Konieczność specyfikacji walltime i pamięci dla każdego zadania
- Redukcja liczby kolejek lokalnych
- Ograniczenie dostępu do l_infinite

Plany dla Zeusa na 2014

- Migracja na nowy (większy) system Lustre + włączenie kwot
- Rozwiązanie problemu z kwotami na GPFS
- Usunięcie problemu z blokowaniem zakolejkowanych zadań – sprawdzanie przed uruchomieniem
- Nowe zakupy

Zalecenia ogólne

- Nie liczyć na maszynie dostępowej!
- Zapisywać wyniki cząstkowe
- Nie trzymać popularnego oprogramowania tylko dla siebie – stworzymy moduł
- Korzystać z modułów (replikacja problemów)
- Szczegółowy opis problemu (numer zadania, jak administrator może zreplikować)

Dyskusja