Bazy danych i usługi sieciowe

Komputery Dużej Mocy
High Performance Computing (HPC)

Paweł Daniluk

Wydział Fizyki

Jesień 2014

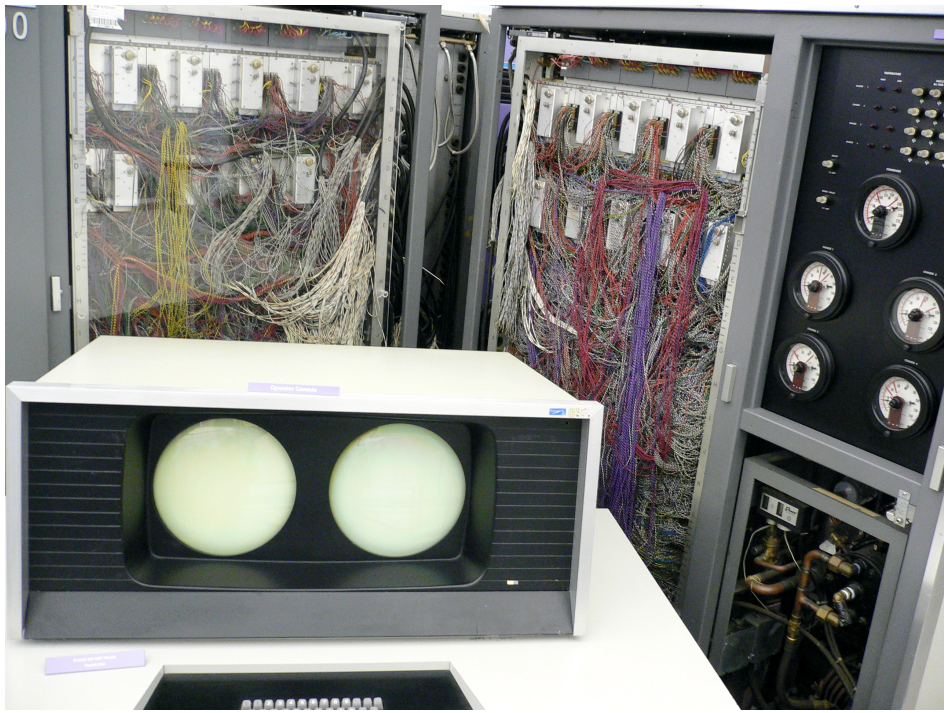


Superkomputery



Superkomputery

1. CDC 6600 – 1964 (500 kFLOPS)



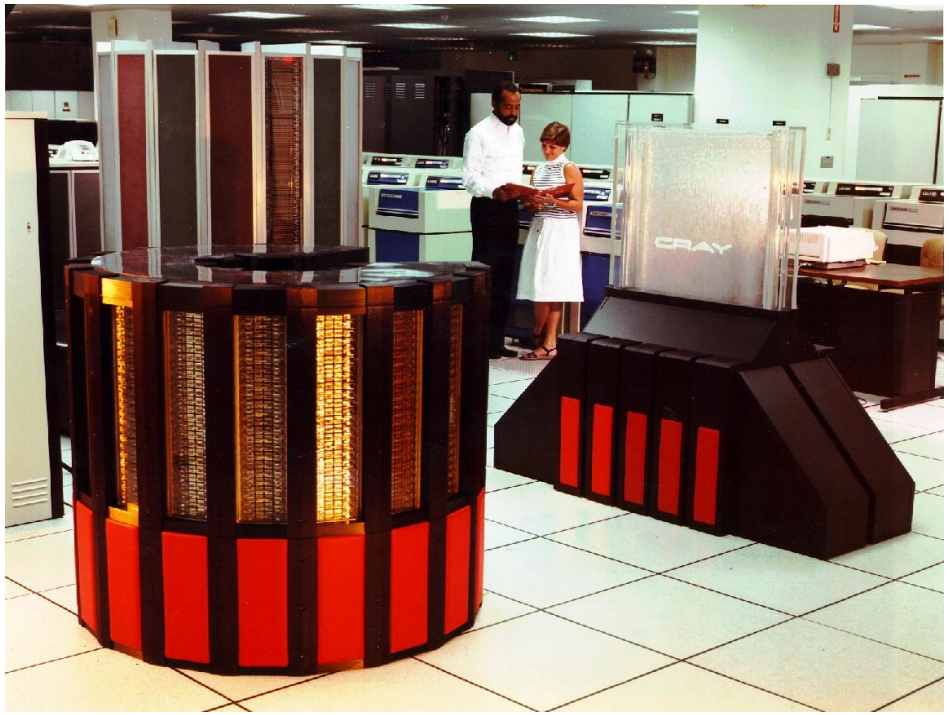
Superkomputery

1. CDC 6600 – 1964 (500 kFLOPS)
2. Cray-1 – 1976 (100 MFLOPS)



Superkomputery

1. CDC 6600 – 1964 (500 kFLOPS)
2. Cray-1 – 1976 (100 MFLOPS)
3. Cray-2 – 1985 (8 CPUs, 1.9 GFLOPS)



Superkomputery

1. CDC 6600 – 1964 (500 kFLOPS)
2. Cray-1 – 1976 (100 MFLOPS)
3. Cray-2 – 1985 (8 CPUs, 1.9 GFLOPS)
4. Fujitsu Numerical Wind Tunnel – 1985 (166 CPUs, 124.5 GFLOPS)



Superkomputery

1. CDC 6600 – 1964 (500 kFLOPS)
2. Cray-1 – 1976 (100 MFLOPS)
3. Cray-2 – 1985 (8 CPUs, 1.9 GFLOPS)
4. Fujitsu Numerical Wind Tunnel – 1985 (166 CPUs, 124.5 GFLOPS)
5. IBM Blue Gene/L – 2007 (106,496 CPUs, 478 TFLOPS)



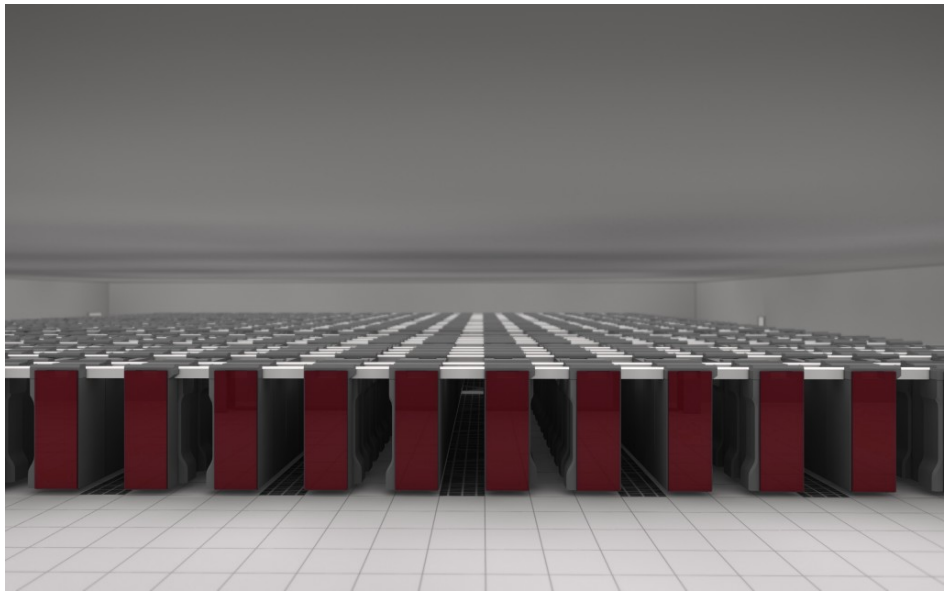
Superkomputery

1. CDC 6600 – 1964 (500 kFLOPS)
2. Cray-1 – 1976 (100 MFLOPS)
3. Cray-2 – 1985 (8 CPUs, 1.9 GFLOPS)
4. Fujitsu Numerical Wind Tunnel – 1985 (166 CPUs, 124.5 GFLOPS)
5. IBM Blue Gene/L – 2007 (106,496 CPUs, 478 TFLOPS)
6. Cray Jaguar – 2009 (224,256 CPUs, 1.75 PFLOPS)



Superkomputery

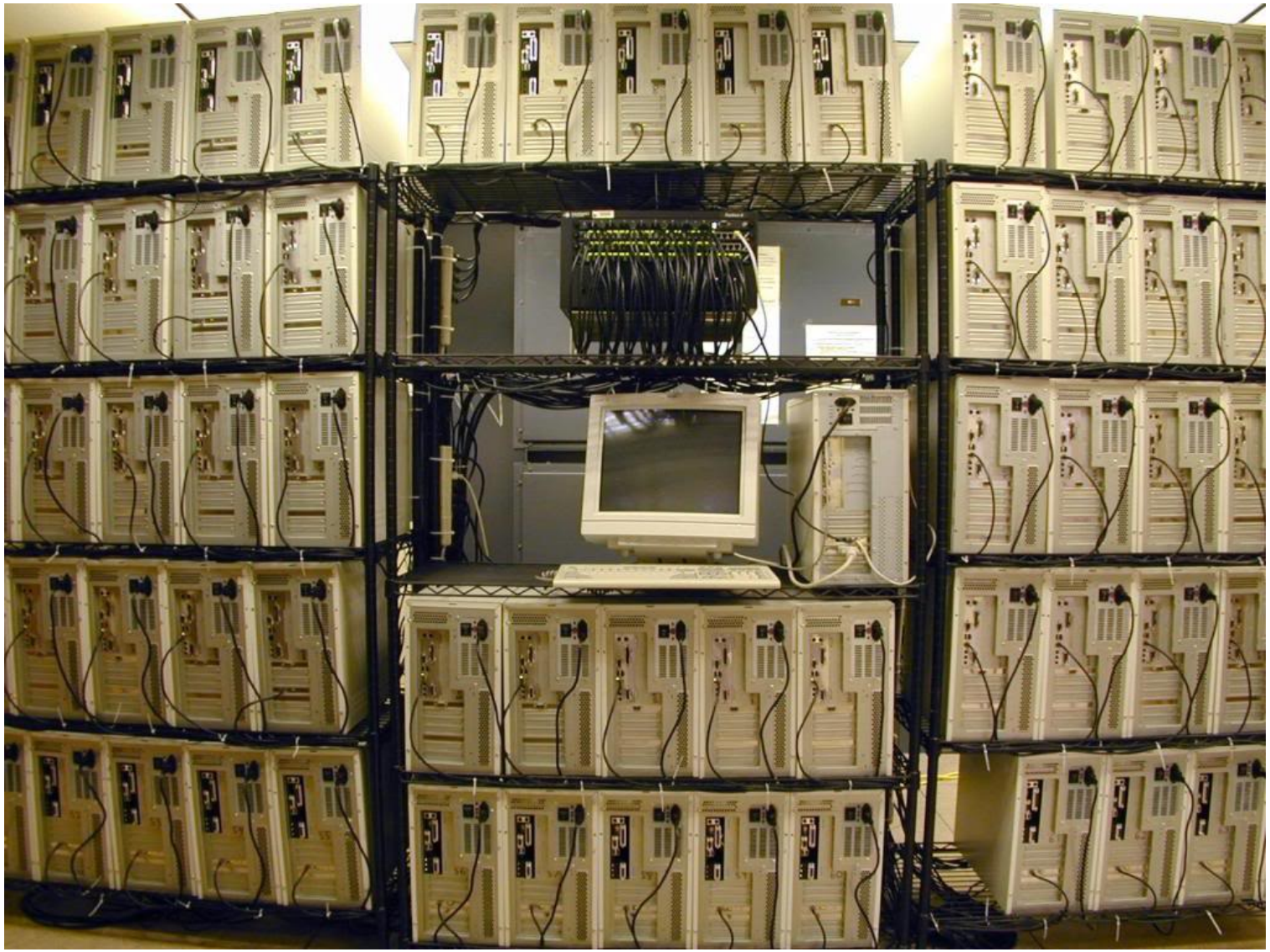
1. CDC 6600 – 1964 (500 kFLOPS)
2. Cray-1 – 1976 (100 MFLOPS)
3. Cray-2 – 1985 (8 CPUs, 1.9 GFLOPS)
4. Fujitsu Numerical Wind Tunnel – 1985 (166 CPUs, 124.5 GFLOPS)
5. IBM Blue Gene/L – 2007 (106,496 CPUs, 478 TFLOPS)
6. Cray Jaguar – 2009 (224,256 CPUs, 1.75 PFLOPS)
7. Fujitsu K – 2011 (88,128 CPUs, 10.51 PFLOPS)



Przełomy technologiczne

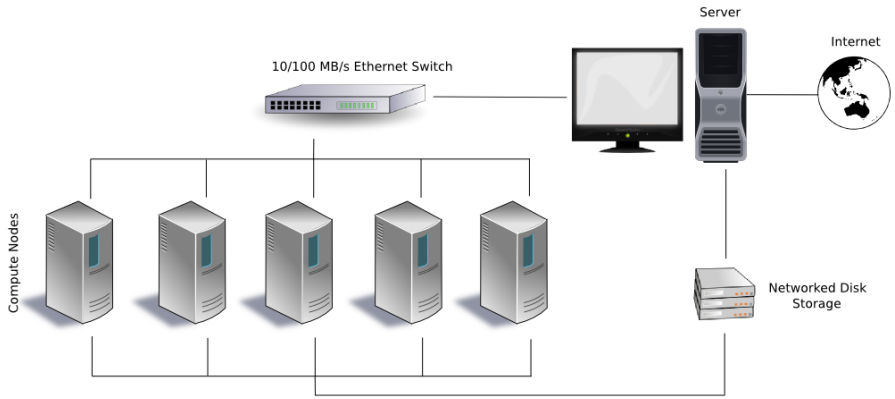
- Reduced Instruction Set Computing (RISC)
- Procesory wektorowe (Single Instruction Multiple Data – SIMD)
- Systemy wieloprocesorowe (Multiple Instruction Multiple Data – MIMD)
  - Symmetric multiprocessing – SMP
  - Non-Uniform Memory Access – NUMA
- Klastry komputerowe

Klastry komputerowe



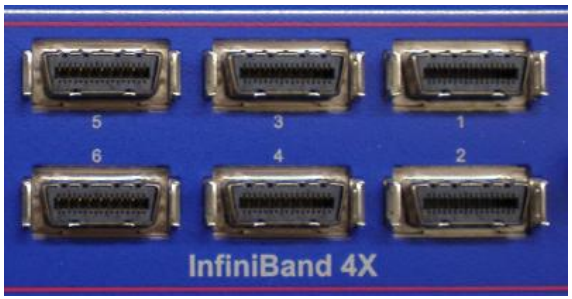
Klastry komputerowe

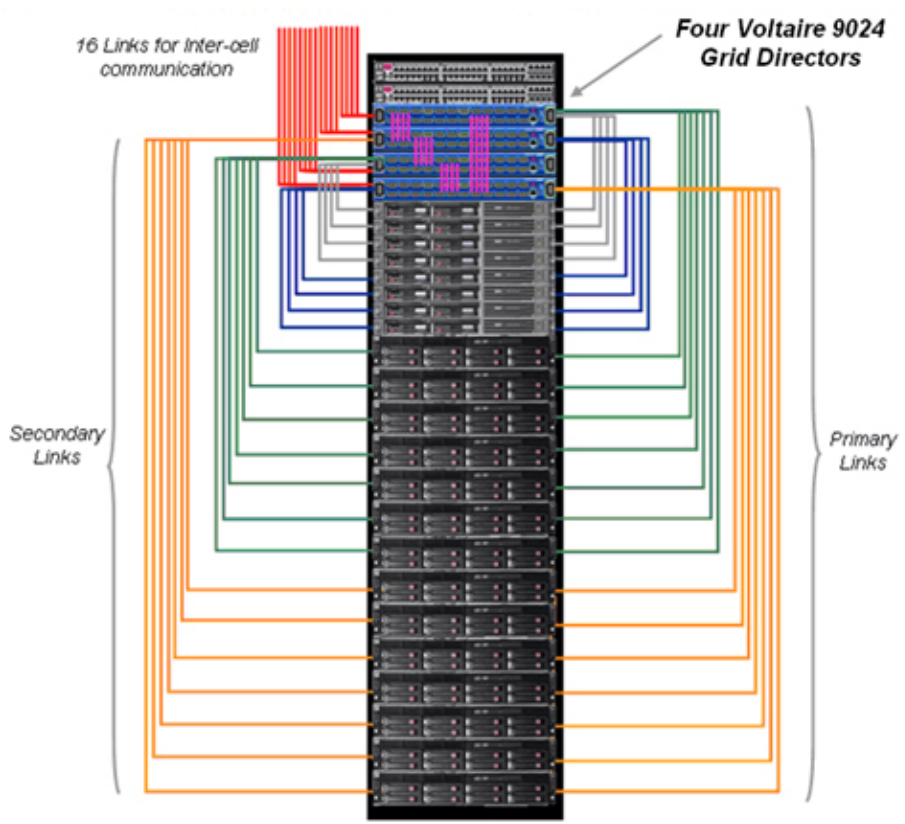
- Wiele serwerów połączonych szybkimi interfejsami sieciowymi (GigaBit Ethernet, Infiniband)
- Wspólna przestrzeń dyskowa
- Standardowe protokoły sieciowe
- Oprogramowanie do kolejkowania obliczeń (PBS)
- Biblioteki do obliczeń równoległych (PVM, MPI)



Infiniband

- Możliwość bezpośredniego zapisu danych do pamięci (DMA)
- Rozgłaszanie, transakcje, operacje atomowe
- Maksymalna przepustowość poj. łącza 25 Gbps
- Opóźnienie 1-2 μs (Gigabit Ethernet – ok. 20 μs)





Message Passing Interface – MPI

Aplikacja składa się z grupy procesów, które mogą być uruchomione na różnych komputerach.

Funkcjonalność

- komunikacja 1 do 1 i 1 do wielu
- synchronizacja procesów
- możliwość definiowania topologii procesów
- biblioteki dla C/C++ i Fortrana

Istnieją implementacje dedykowane do konkretnych rozwiązań sprzętowych.

Systemy kolejkowe

Systemy kolejkowe służą do przydzielania zasobów zadaniom oczekującym na wykonanie.

Funkcjonalność

- kontrola dostępności zasobów (procesorów, oprogramowania, pamięci, przestrzeni dyskowej etc.)
- opis wymagań zadania
- uprawnienia użytkowników i priorytetyzacja zadań
- rejestrowanie wykorzystania zasobów

Przechowywanie danych

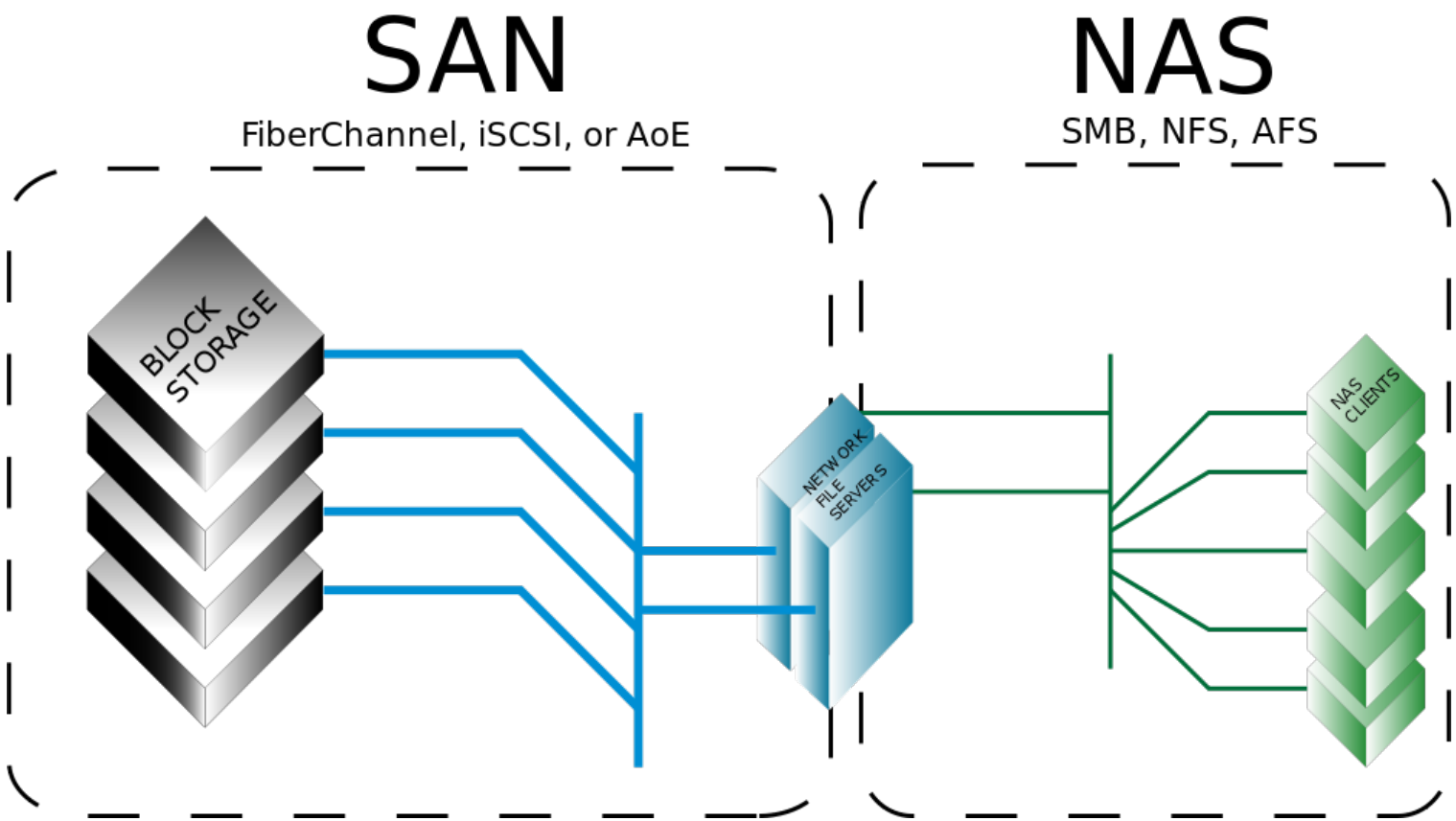
Network Attached Storage – NAS

- Urządzenie pozwalające na udostępnienie zasobów dyskowych w sieci komputerowej.
- Dostęp na poziomie plików
- Protokoły: NFS, AFP, SMB/CIFS

Storage Area Network – SAN

- Dedykowana sieć do łączenia urządzeń dyskowych z serwerami
- Dostęp na poziomie urządzeń blokowych (woluminów)
- Typy sieci: FibreChannel, iSCSI

Przechowywanie danych – c.d.



Przechowywanie danych – c.d.

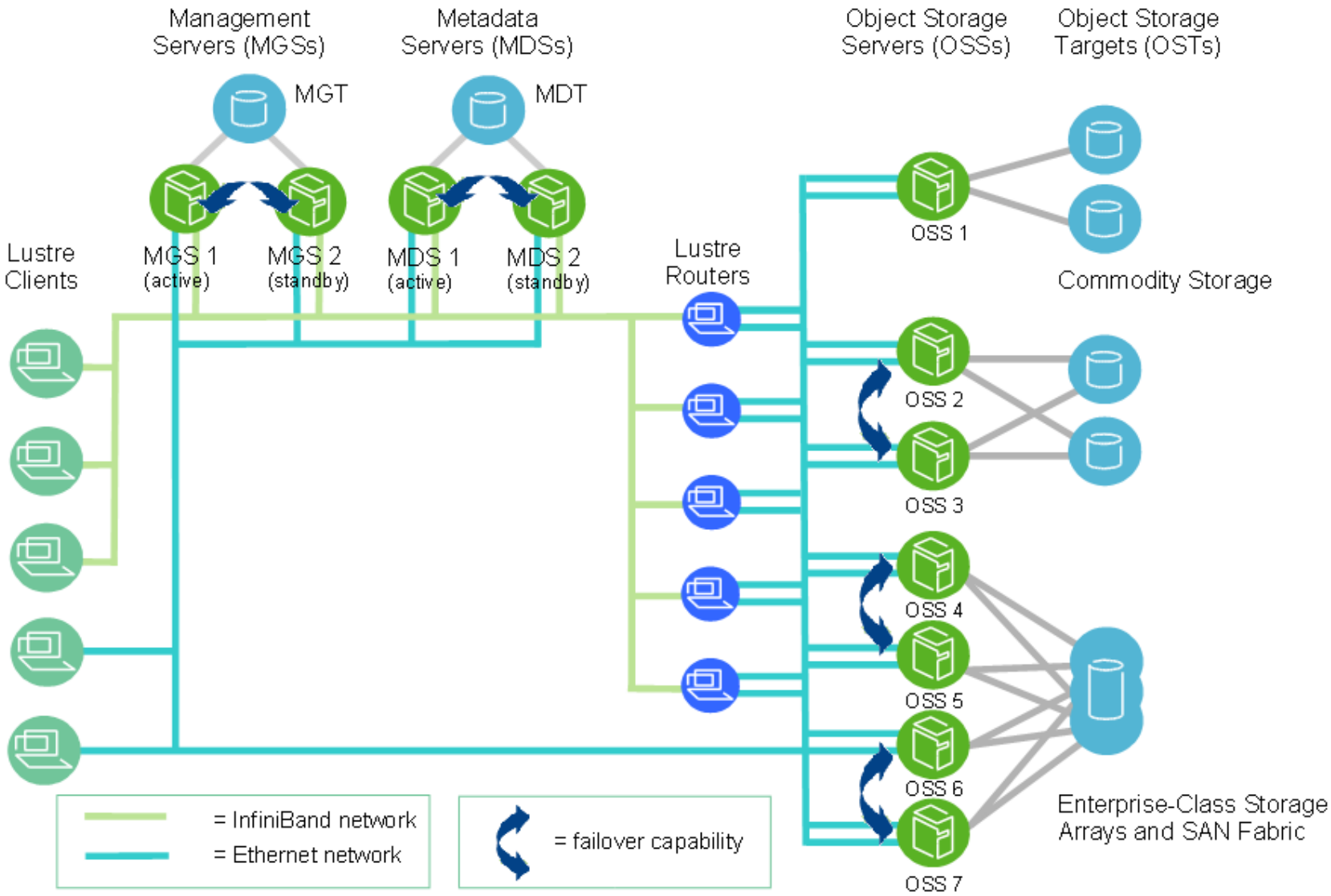
NAS



SAN



Rozproszone systemy plików – Lustre

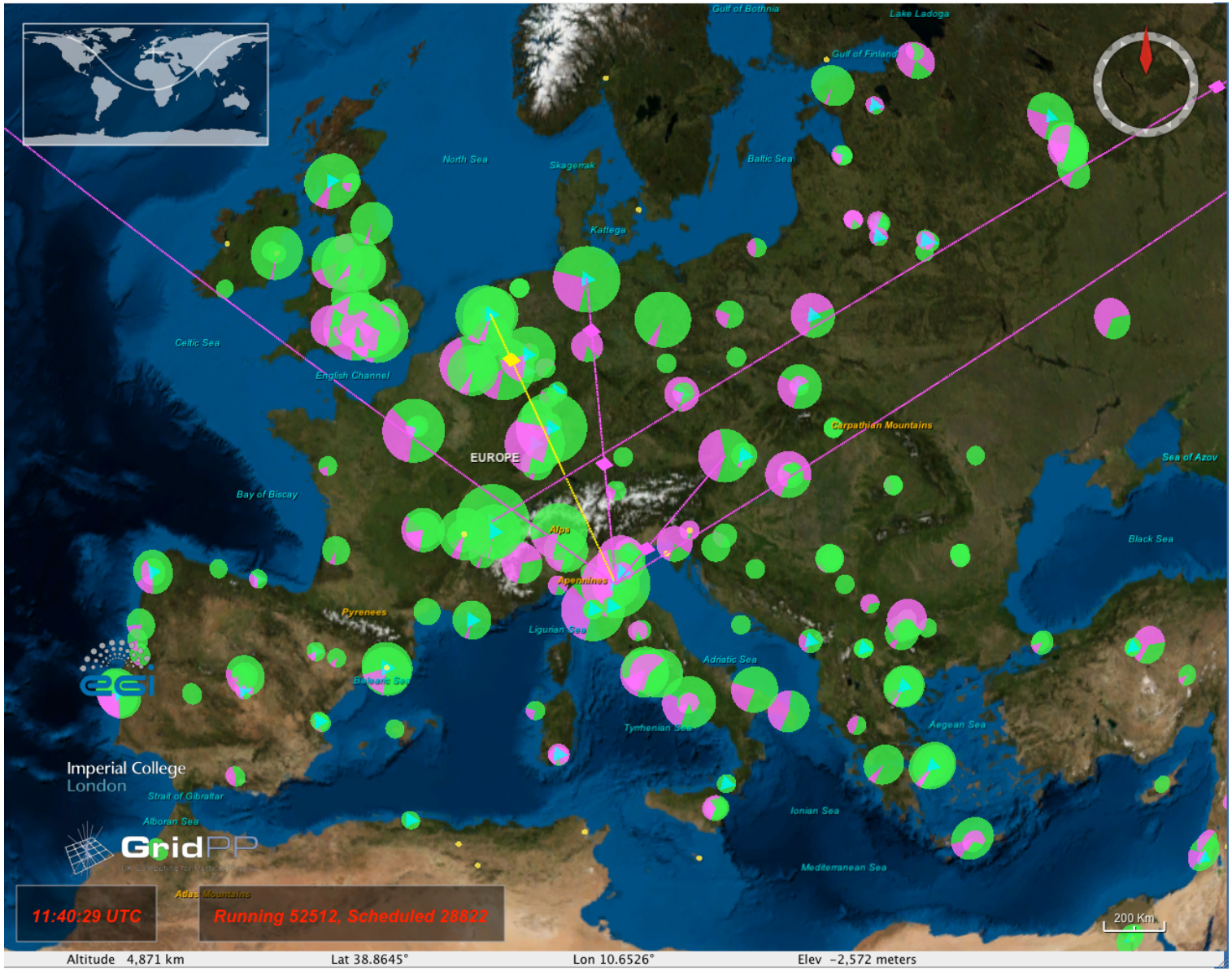


Opportunistic supercomputing

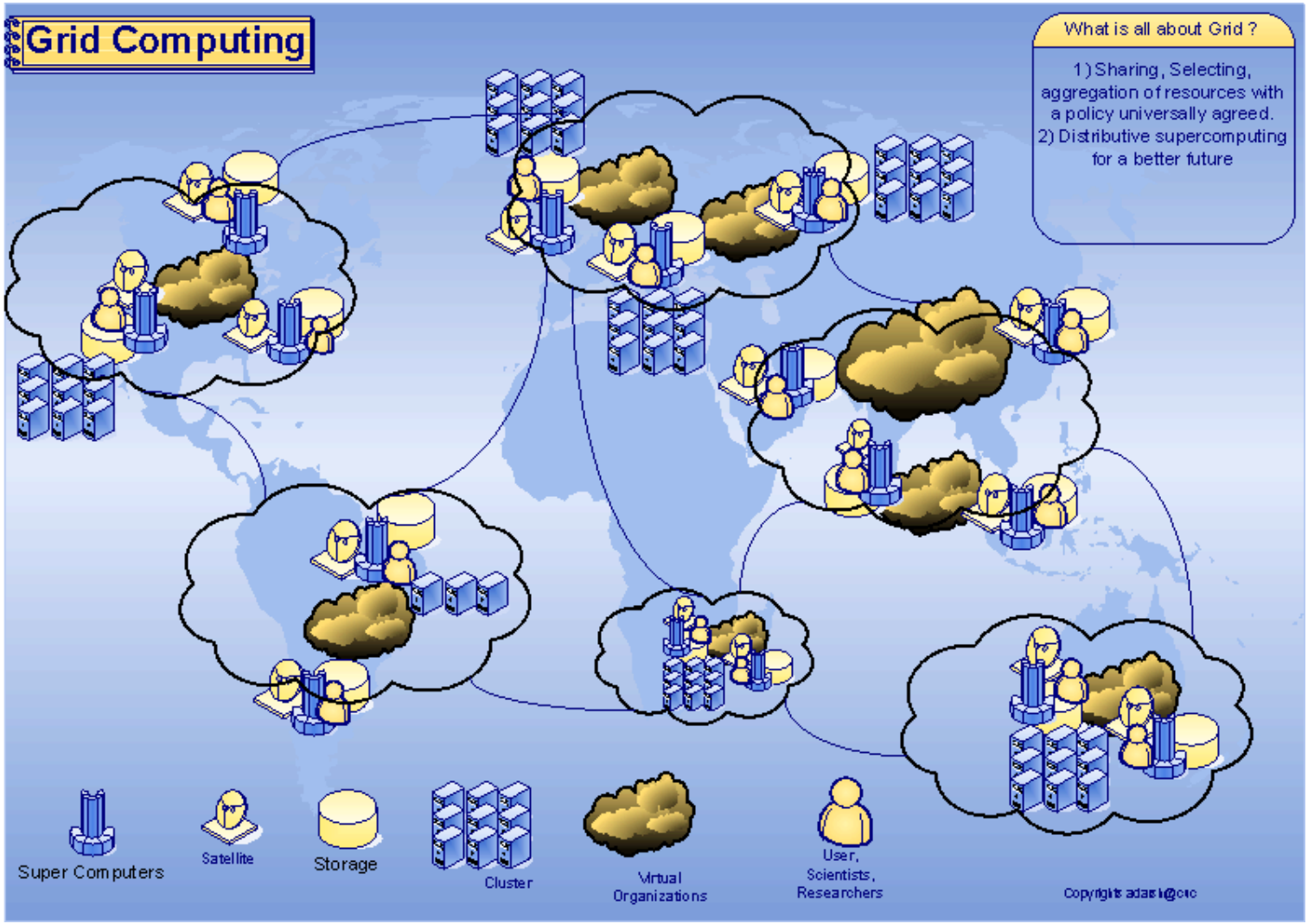
Możliwe jest wykorzystanie mocy obliczeniowej komputerów rozproszonych w Internecie (np. niewykorzystanych cykli darowanych przez ochotników).

- Folding@Home – 8.8 PFLOPS
- BOINC (Berkeley Open Infrastructure for Network Computing) – 5.5 PFLOPS

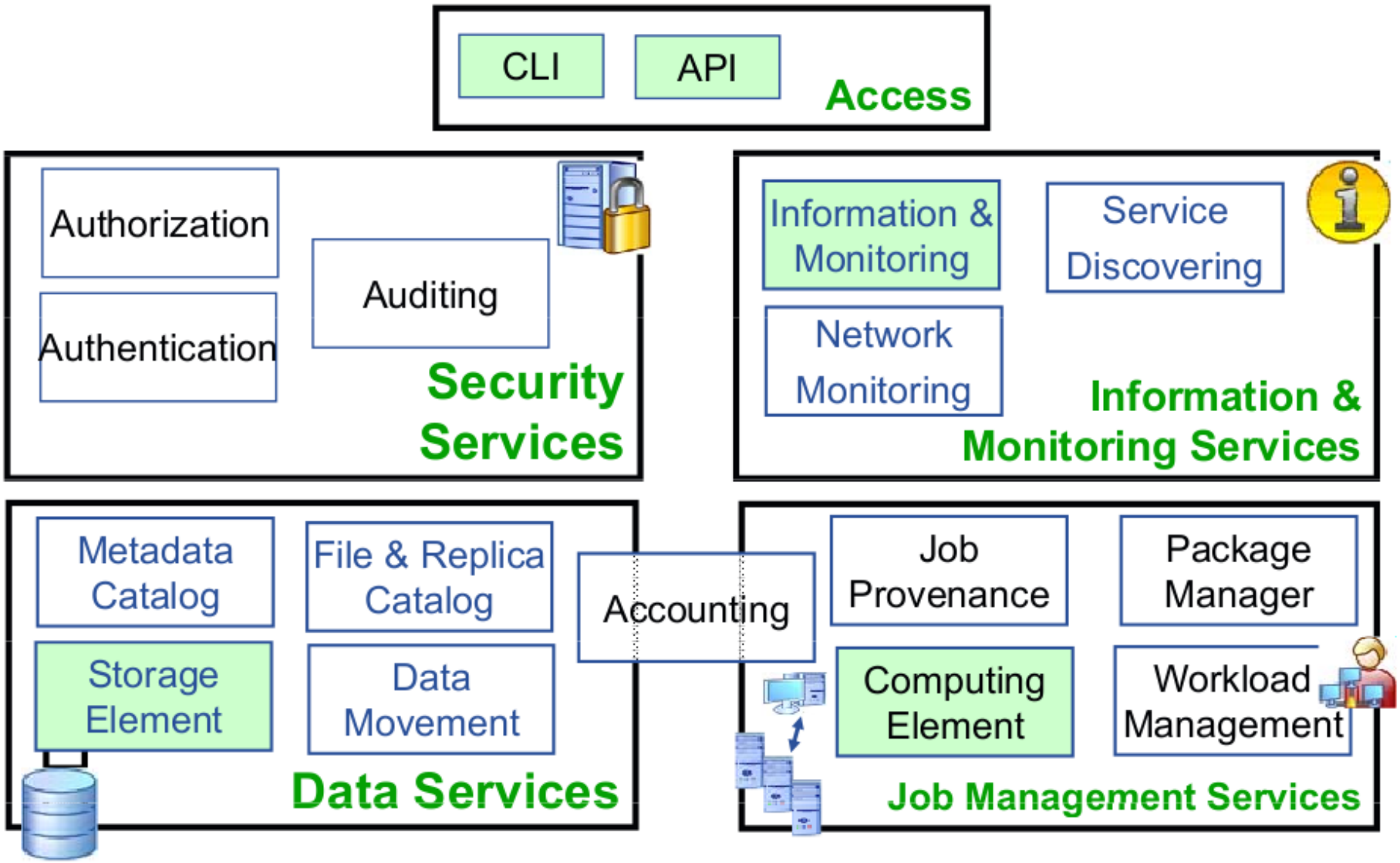
Gridy obliczeniowe



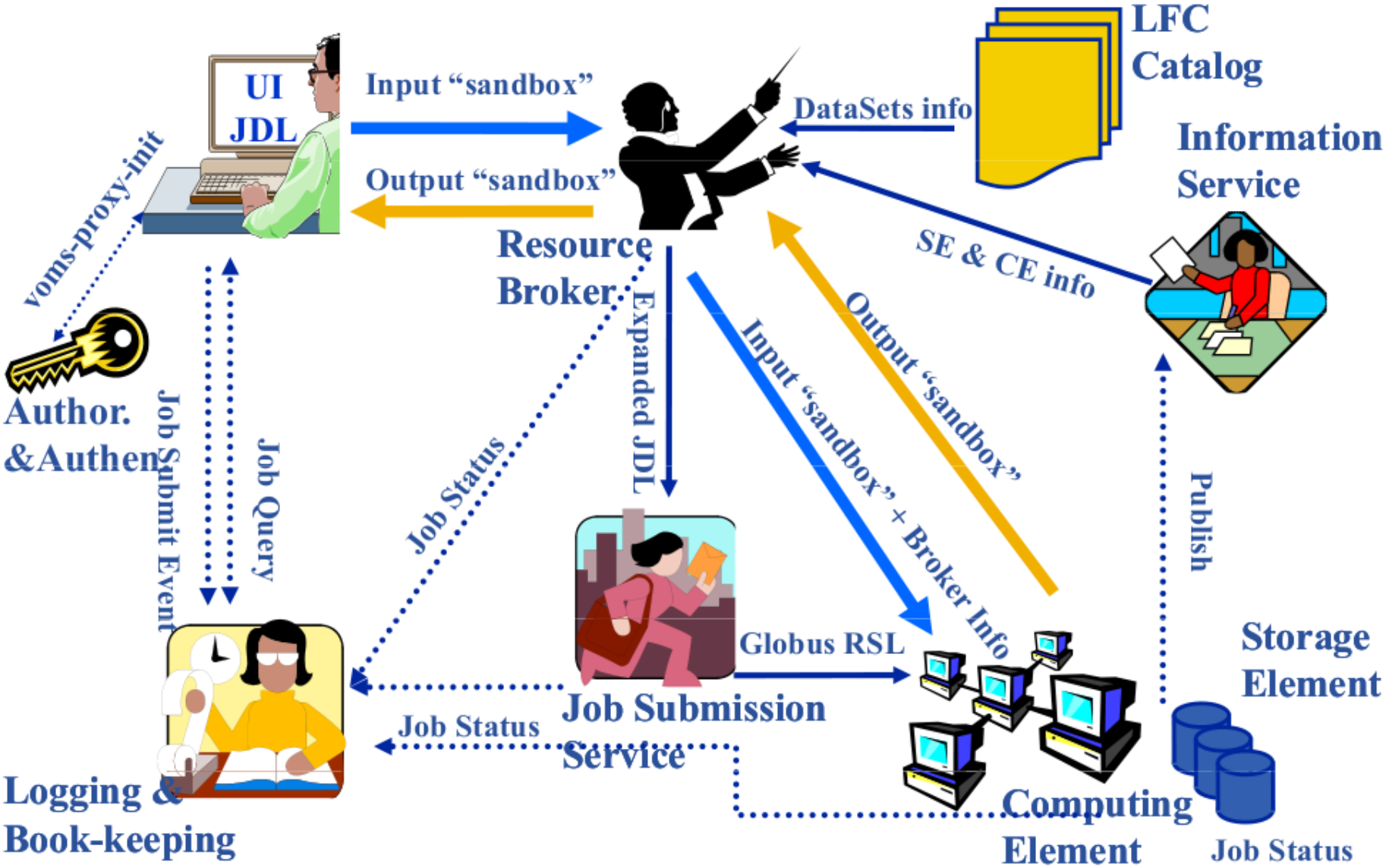
Gridy obliczeniowe c.d.



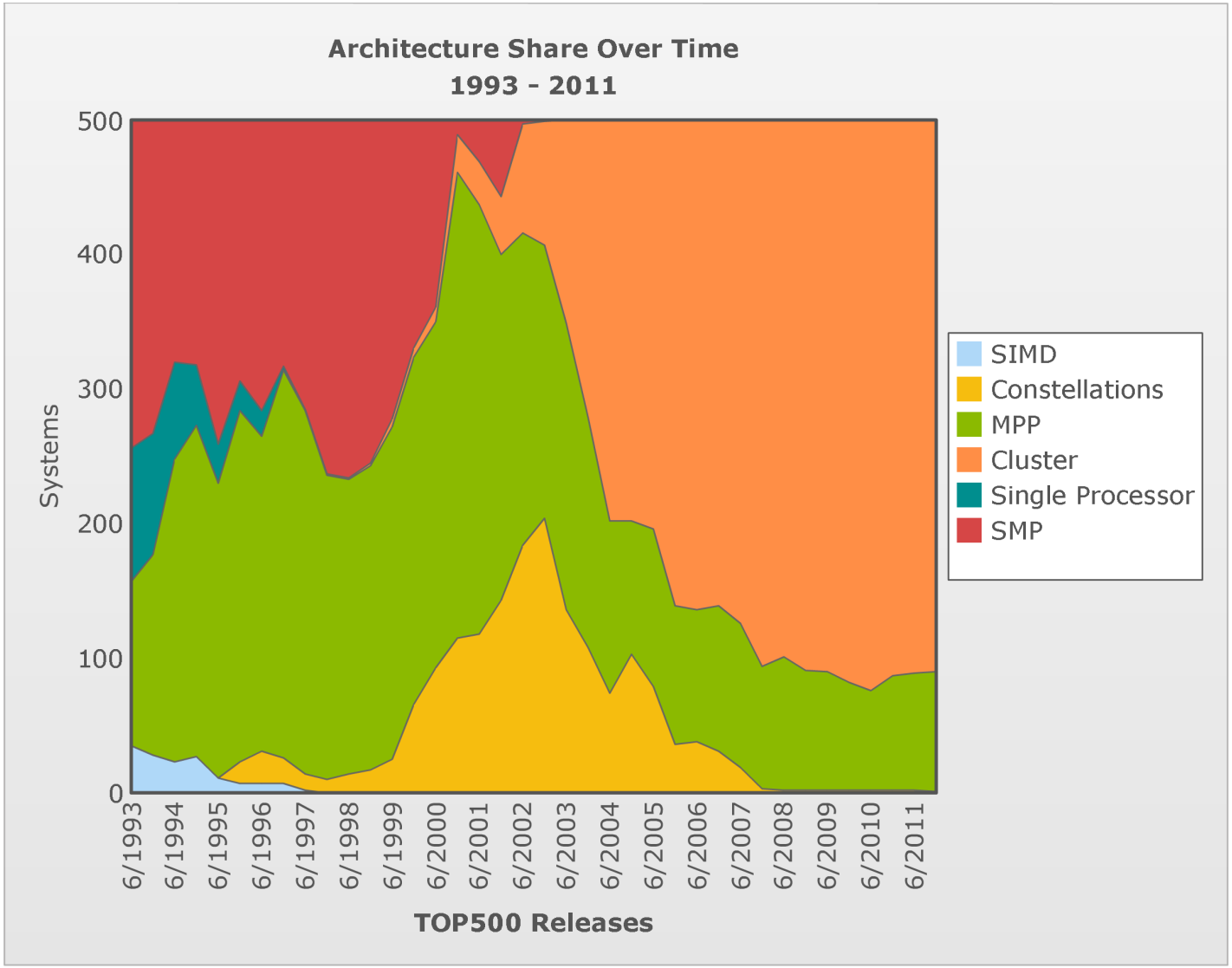
Przykład gLite, EGEE, PL-GRID



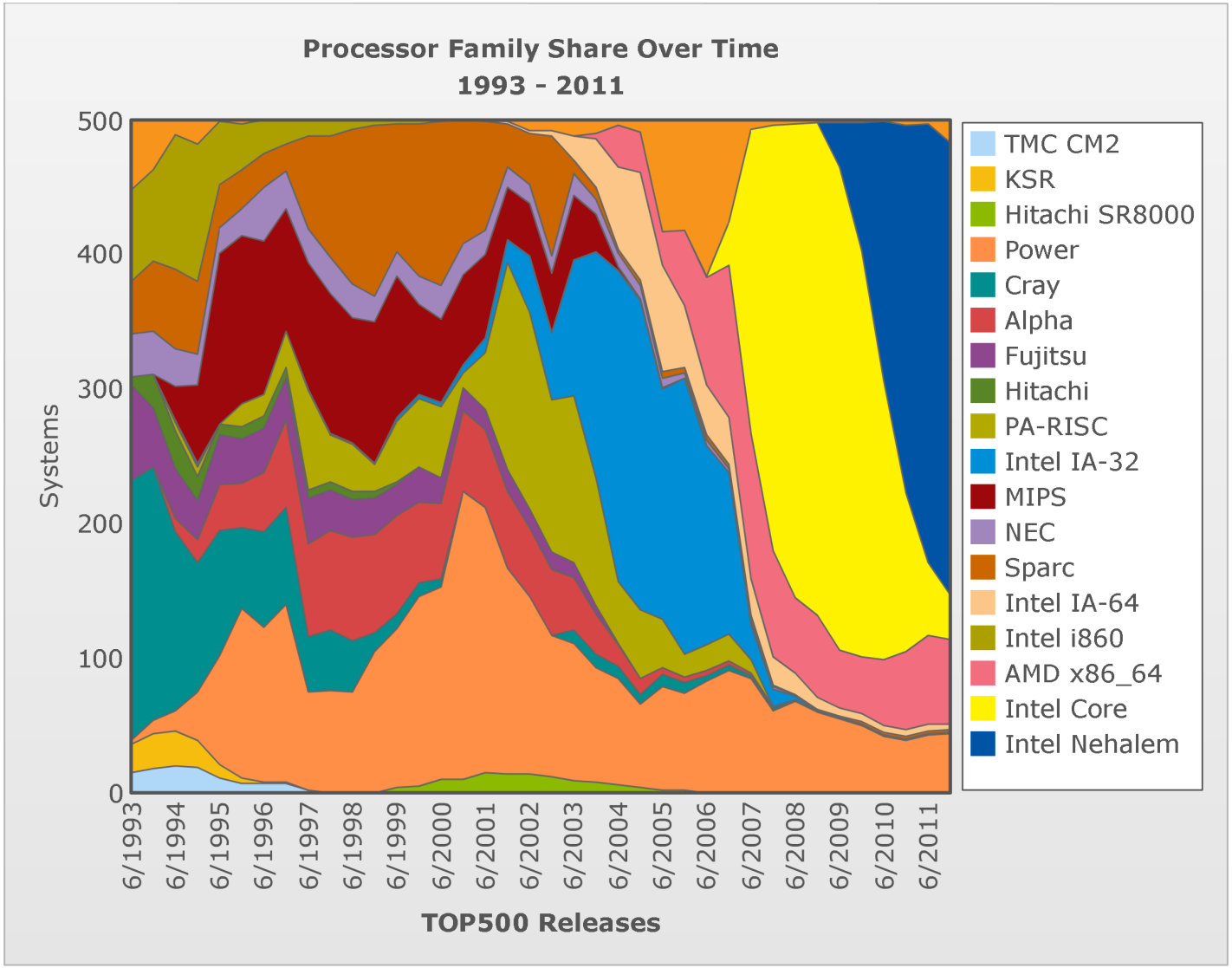
Przykład gLite, EGEE, PL-GRID c.d.



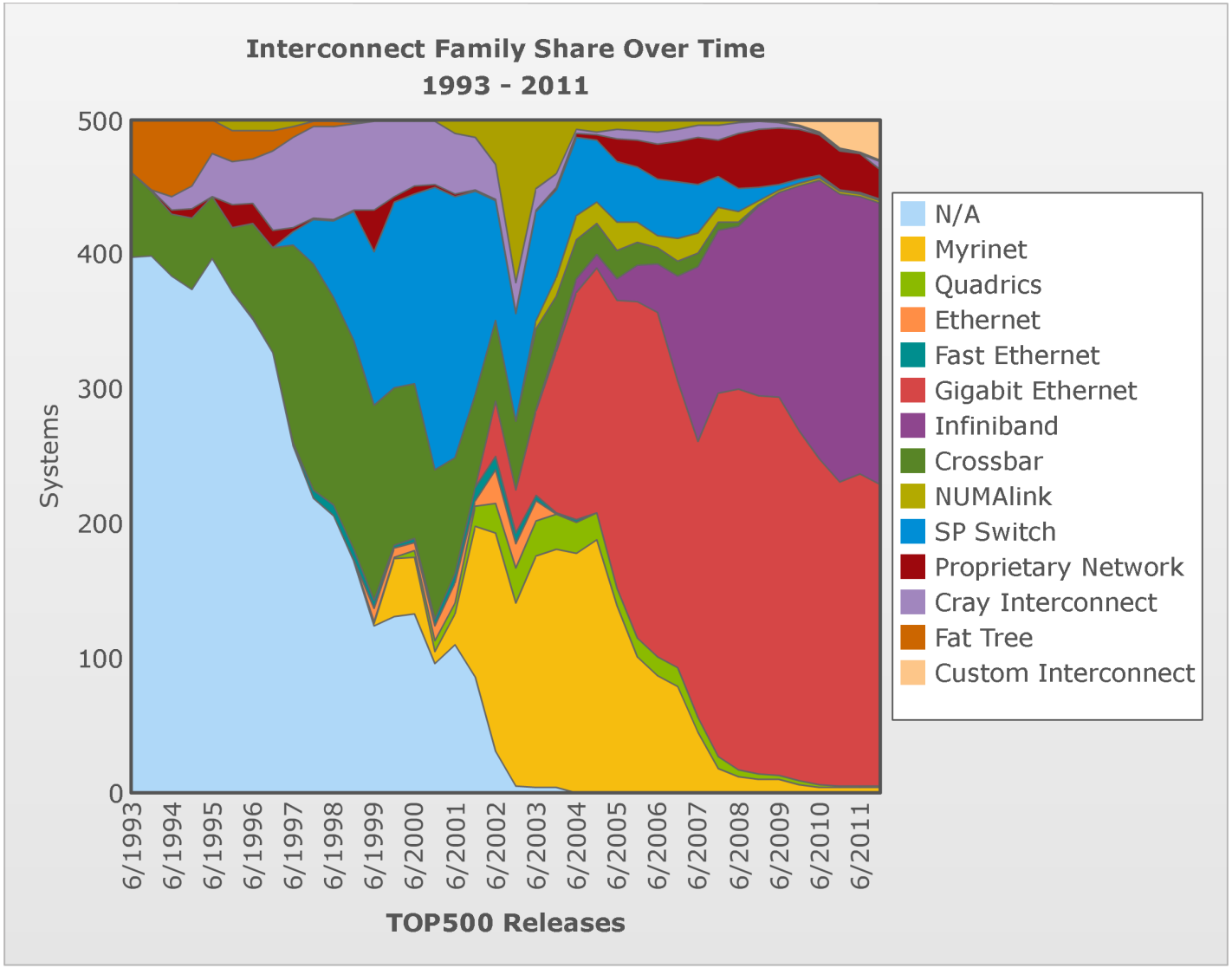
Top500


Architecture Share Over Time
1993 - 2011
500
400
300
200
100
0
Systems
6/1993
6/1994
6/1995
6/1996
6/1997
6/1998
6/1999
6/2000
6/2001
6/2002
6/2003
6/2004
6/2005
6/2006
6/2007
6/2008
6/2009
6/2010
6/2011
TOP500 Releases
SIMD
Constellations
MPP
Cluster
Single Processor
SMP

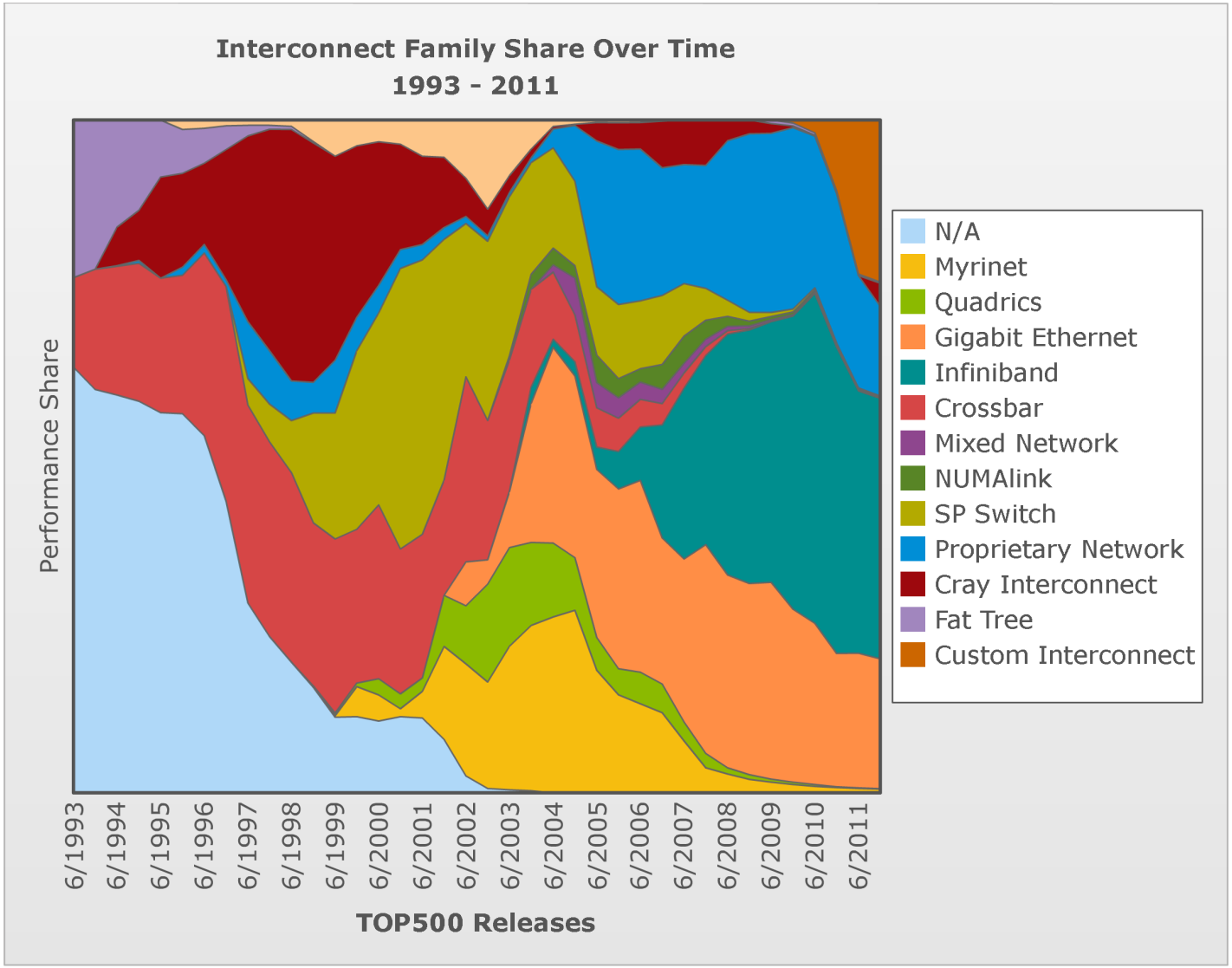
Top500


Processor Family Share Over Time
1993 - 2011
Systems
500
400
300
200
100
0
6/1993
6/1994
6/1995
6/1996
6/1997
6/1998
6/1999
6/2000
6/2001
6/2002
6/2003
6/2004
6/2005
6/2006
6/2007
6/2008
6/2009
6/2010
6/2011
TOP500 Releases
TMC CM2
KSR
Hitachi SR8000
Power
Cray
Alpha
Fujitsu
Hitachi
PA-RISC
Intel IA-32
MIPS
NEC
Sparc
Intel IA-64
Intel i860
AMD x86_64
Intel Core
Intel Nehalem

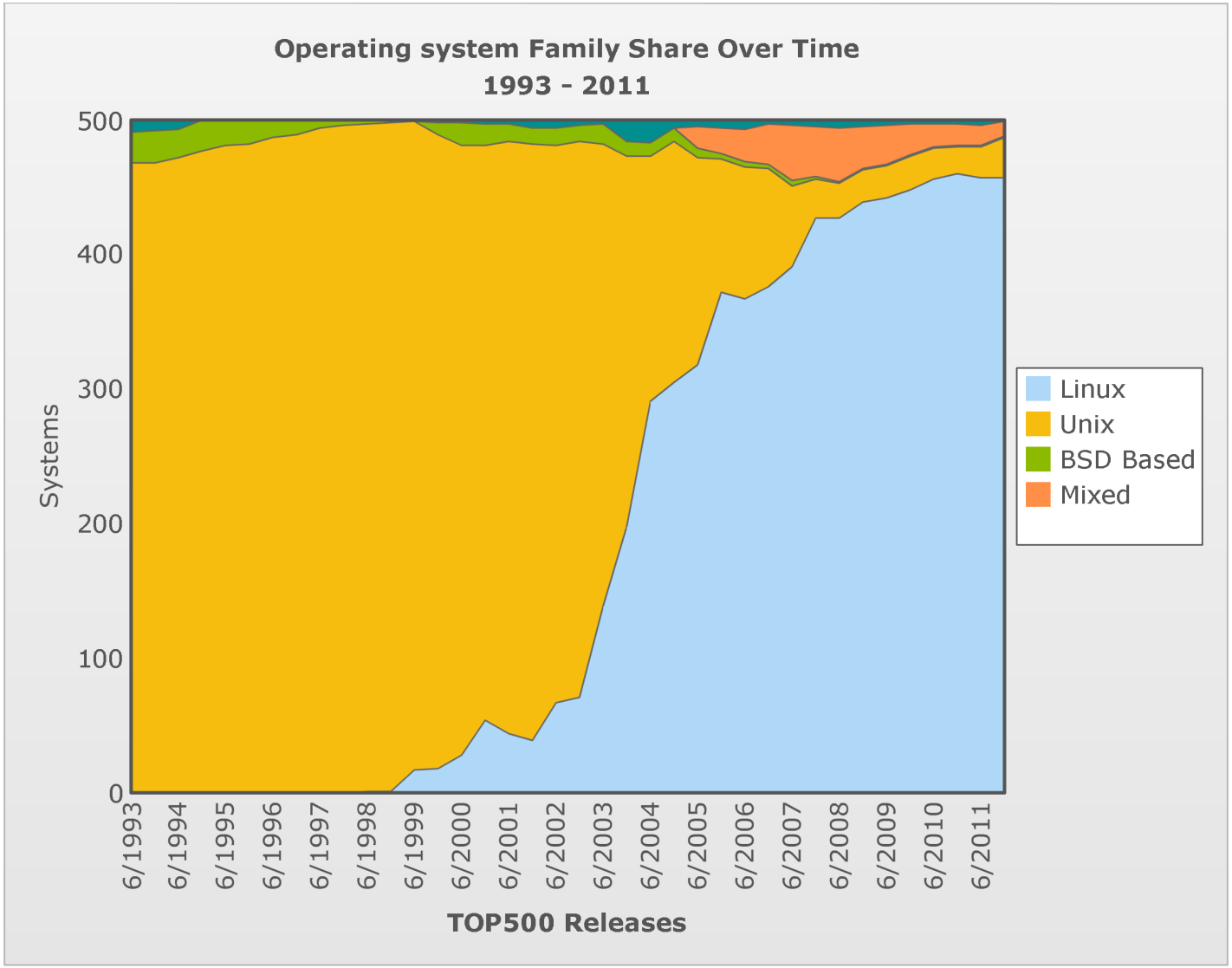
Top500


Interconnect Family Share Over Time
1993 - 2011
Systems
500
400
300
200
100
0
6/1993
6/1994
6/1995
6/1996
6/1997
6/1998
6/1999
6/2000
6/2001
6/2002
6/2003
6/2004
6/2005
6/2006
6/2007
6/2008
6/2009
6/2010
6/2011
TOP500 Releases
N/A
Myrinet
Quadrics
Ethernet
Fast Ethernet
Gigabit Ethernet
Infiniband
Crossbar
NUMAlink
SP Switch
Proprietary Network
Cray Interconnect
Fat Tree
Custom Interconnect

Top500


Interconnect Family Share Over Time
1993 - 2011
Performance Share
6/1993
6/1994
6/1995
6/1996
6/1997
6/1998
6/1999
6/2000
6/2001
6/2002
6/2003
6/2004
6/2005
6/2006
6/2007
6/2008
6/2009
6/2010
6/2011
TOP500 Releases
N/A
Myrinet
Quadrics
Gigabit Ethernet
Infiniband
Crossbar
Mixed Network
NUMAlink
SP Switch
Proprietary Network
Cray Interconnect
Fat Tree
Custom Interconnect

Top500


Operating system Family Share Over Time
1993 - 2011
Systems
500
400
300
200
100
0
6/1993
6/1994
6/1995
6/1996
6/1997
6/1998
6/1999
6/2000
6/2001
6/2002
6/2003
6/2004
6/2005
6/2006
6/2007
6/2008
6/2009
6/2010
6/2011
TOP500 Releases
Linux
Unix
BSD Based
Mixed

Top500 c.d.

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (Kw)
1	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12659.9
2	National Supercomputing Center in Tianjin China	**Tianhe-1A** - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 NUDT	186368	2566.0	4701.0	4040
3	DOE/SC/Oak Ridge National Laboratory United States	**Jaguar** - Cray XT5-HE Opteron 6-core 2.6 GHz Cray Inc.	224162	1759.0	2331.0	6950
4	National Supercomputing Centre in Shenzhen (NSCS) China	**Nebulae** - Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050 Dawning	120640	1271.0	2984.3	2580
5	GSIC Center, Tokyo Institute of Technology Japan	**TSUBAME 2.0** - HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows NEC/HP	73278	1192.0	2287.6	1398.6
6	DOE/NNSA/LANL/SNL United States	**Cielo** - Cray XE6, Opteron 6136 8C 2.40GHz, Custom Cray Inc.	142272	1110.0	1365.8	3980
7	NASA/Ames Research Center/NAS United States	**Pleiades** - SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband SGI	111104	1088.0	1315.3	4102
8	DOE/SC/LBNL/NERSC United States	**Hopper** - Cray XE6, Opteron 6172 12C 2.10GHz, Custom Cray Inc.	153408	1054.0	1288.6	2910
9	Commissariat a l'Energie Atomique (CEA) France	**Tera-100** - Bull bullx super-node S6010/S6030 Bull SA	138368	1050.0	1254.5	4590
10	DOE/NNSA/LANL United States	**Roadrunner** - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband IBM	122400	1042.0	1375.8	2345

Top500 c.d.

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (Kw)
88	Cyfronet Poland	**Zeus** - Cluster Platform 3000 BL 2x220, Xeon X5650 6C 2.66 GHz, Infiniband Hewlett-Packard	15264	128.8	162.4	
279	Gdansk University of Technology, CI Task Poland	**Galera Plus** - ACTION Xeon HP BL2x220/BL490 E5345/L5640 Infiniband ACTION	10384	65.6	97.8	
296	Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw Poland	**Boreas** - Power 775, POWER7 8C 3.84 GHz, Custom IBM	2560	64.3	78.6	156.7
298	Poznan Supercomputing and Networking Center, Institute of Bioorganic Chemistry Poland	Rackable C1103-G15, Opteron 6234 12C 2.40 GHz, Infiniband QDR SGI	5640	63.9	136.4	
348	Grupa Allegro Poland	Cluster Platform 3000 BL 2x220, Xeon L5420 4C 2.50 GHz, Gigabit Ethernet Hewlett-Packard	10748	59.1	107.5	
360	Wroclaw Centre for Networking and Supercomputing Poland	**Supernova** - Cluster Platform 3000 BL2x220, X56xx 2.66 Ghz, Infiniband Hewlett-Packard	6348	57.4	67.5	

Koszt

Date	Approximate cost per GFLOPS	Technology	Comments
1961	US $1,100,000,000,000 ($1.1 trillion)	About 17 million IBM 1620 units costing $64,000 each	The 1620's multiplication operation takes 17.7 ms.[39]
1984	$15,000,000	Cray X-MP	
1997	$30,000	Two 16-processor Beowulf clusters with Pentium Pro microprocessors[40]	
April 2000	$1,000	Bunyip Beowulf cluster	Bunyip was the first sub-US$1/MFLOPS computing technology. It won the Gordon Bell Prize in 2000.
May 2000	$640	KLAT2	KLAT2 was the first computing technology which scaled to large applications while staying under US$1/MFLOPS.[41]
August 2003	$82	KASY0	KASY0 was the first sub-US$100/GFLOPS computing technology.[42]
August 2007	$48	Microwulf	As of August 2007, this 26.25 GFLOPS "personal" Beowulf cluster can be built for $1256.[43]
March 2011	$1.80	HPU4Science	This $30,000 cluster was built using only commercially available "gamer" grade hardware.[44]

Strona wykładu

http://bioexploratorium.pl/wiki/
Bazy_Danych_i_Usługi_Sieciowe_-_2014z