33rd TOP500 List

ISC'09, Hamburg

Agenda

- Welcome and Introduction (H.W. Meuer)
- TOP10 and Awards (H.D. Simon)
- Highlights of the 30th TOP500 (E. Strohmaier)
- HPC Power Consumption (J. Shalf)
- Multicore and Manycore and their Impact on HPC (J.J. Dongarra)
- Discussion

33rd List: The TOP10

Rank	Site	Manufacturer	Computer	Country	Cores	Rmax [Tflops]	Power [MW]
1	DOE/NNSA/LANL	IBM	Roadrunner BladeCenter QS22/LS21	USA	129,600	1,105.0	2.48
2	Oak Ridge National Laboratory	Cray Inc.	Jaguar Cray XT5 QC 2.3 GHz	USA	150,152	1,059.0	6.95
3	Forschungszentrum Juelich (FZJ)	IBM	Jugene Blue Gene/P Solution	Germany	294,912	825.50	2.26
4	NASA/Ames Research Center/NAS	SGI	Pleiades SGI Altix ICE 8200EX	USA	51,200	487.0	2.09
5	DOE/NNSA/LLNL	IBM	BlueGene/L eServer Blue Gene Solution	USA	212,992	478.2	2.32
6	University of Tennessee	Cray	Kraken Cray XT5 QC 2.3 GHz	USA	66,000	463.30	
7	Argonne National Laboratory	IBM	Intrepid Blue Gene/P Solution	USA	163,840	458.61	1.26
8	TACC/U. of Texas	Sun	Ranger SunBlade x6420	USA	62,976	433.2	2.0
9	DOE/NNSA/LANL	IBM	Dawn Blue Gene/P Solution	USA	147,456	415.70	1.13
10	Forschungszentrum Juelich (FZJ)	Sun/Bull SA	JUROPA NovaScale /Sun Blade	Germany	26,304	274.80	1.54

The TOP500 Project

- Listing the 500 most powerful computers in the world
- Yardstick: Rmax of Linpack
  – Solve Ax=b, dense problem, matrix is random
- Update twice a year:
  – ISC'xy in June in Germany • SCxy in November in the U.S.
- All information available from the TOP500 web site at: www.top500.org

TOP500 Status

- 1st Top500 List in June 1993 at ISC'93 in Mannheim
    - •
    - •
- 30th Top500 List on November 13, 2007 at SC07 in Reno
- 31st Top500 List on June 18, 2008 at ISC'08 in Dresden
- 32nd Top500 List on November 18, 2008 at SC08 in Austin
- 33rd Top500 List on June 24, 2009 at ISC'09 in Hamburg
- 34th Top500 List on November 18, 2009 at SC09 in Portland

- Acknowledged by HPC-users, manufacturers and media

33rd List: Highlights

- The Roadrunner system, which broke the petaflop/s barrier one year ago, held on to its No. 1 spot. It still is one of the most energy efficient systems on the TOP500.
- The most powerful system outside the U.S. is an IBM BlueGene/P system at the German Forschungszentrum Juelich (FZJ) at No. 3. FZJ has a second system in the TOP10, a mixed Bull and Sun system at No. 10.
- Intel dominates the high-end processor market with 79.8 percent of all systems and 87.7 percent of quad-core based systems.
- Intel Core i7 (Nehalem) makes its first appearance in the list with 33 systems
- Quad-core processors are used in 76.6 percent of the systems. Their use accelerates performance growth at all levels.
- Other notable systems are:
  - An IBM BlueGene/P system at the King Abdullah University of Science and Technology (KAUST) in Saudi Arabia at No. 14
  - The Chinese-built Dawning 5000A at the Shanghai Supercomputer Center at No 15. It is the largest system which can be operated with Windows HPC 2008.
- Hewlett-Packard kept a narrow lead in market share by total systems from IBM, but IBM still stays ahead by overall installed performance.
- Cray's XT system series is very popular for big customers 10 systems in the TOP50 (20 percent).

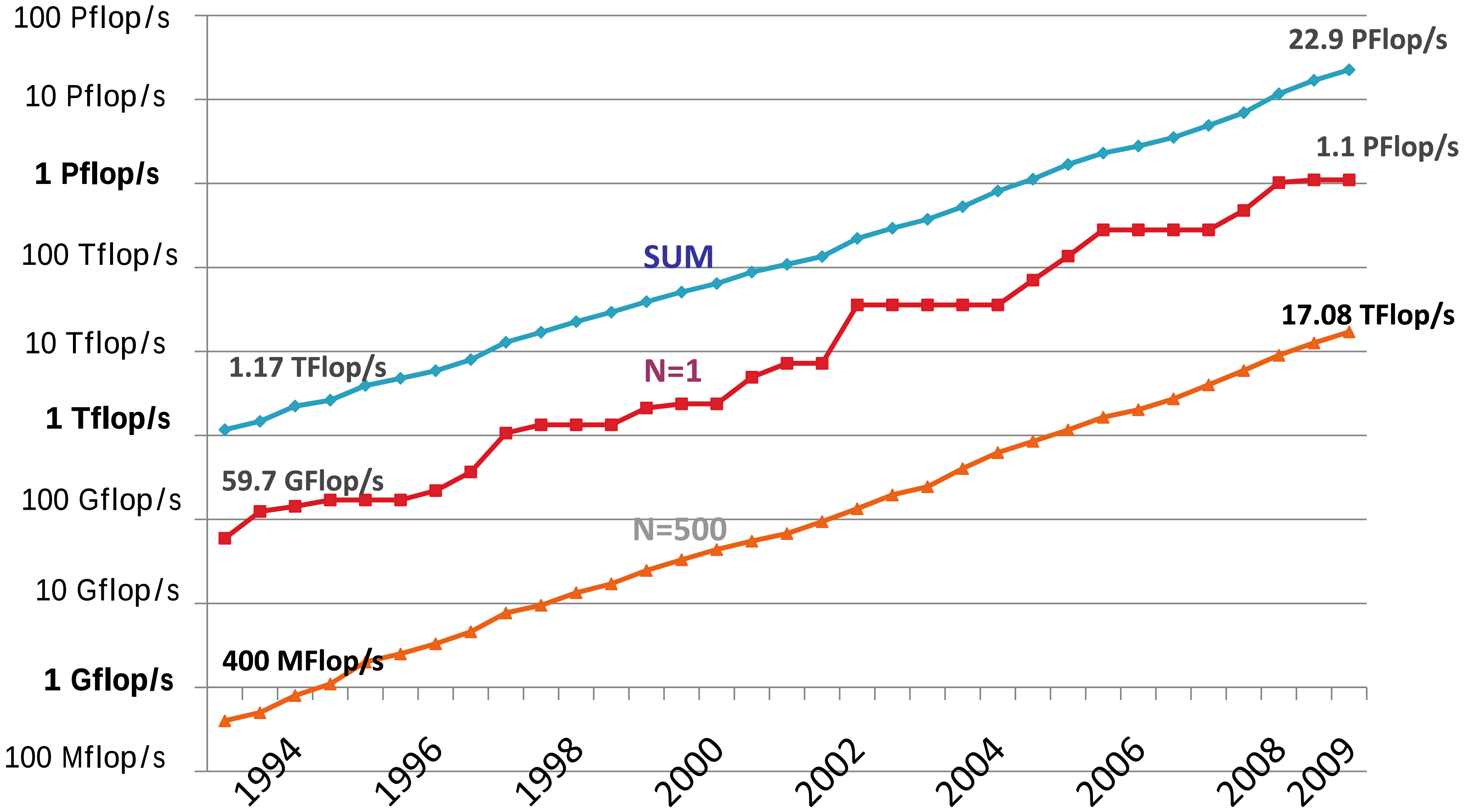
33rd List: Notable (New) Systems

- Shaheen, an IBM BlueGene/P system at the King Abdullah University of Science and Technology (KAUST) in Saudi Arabia at No. 14
- The Chinese-built Dawning 5000A at the Shanghai Supercomputer Center at No. 15. It is the largest system which can be operated with Windows HPC 2008.
- The (new) Earth Simulator at No. 22 (SX-9E based - only vector system) – it's back!
- Koi, a Cray internal system using Shanghai six-cores at #128.
- A Grape-DR system at the National Astronomical Observatory (NAO/CfCA) at # 259
  - target size 2PF/s peak !

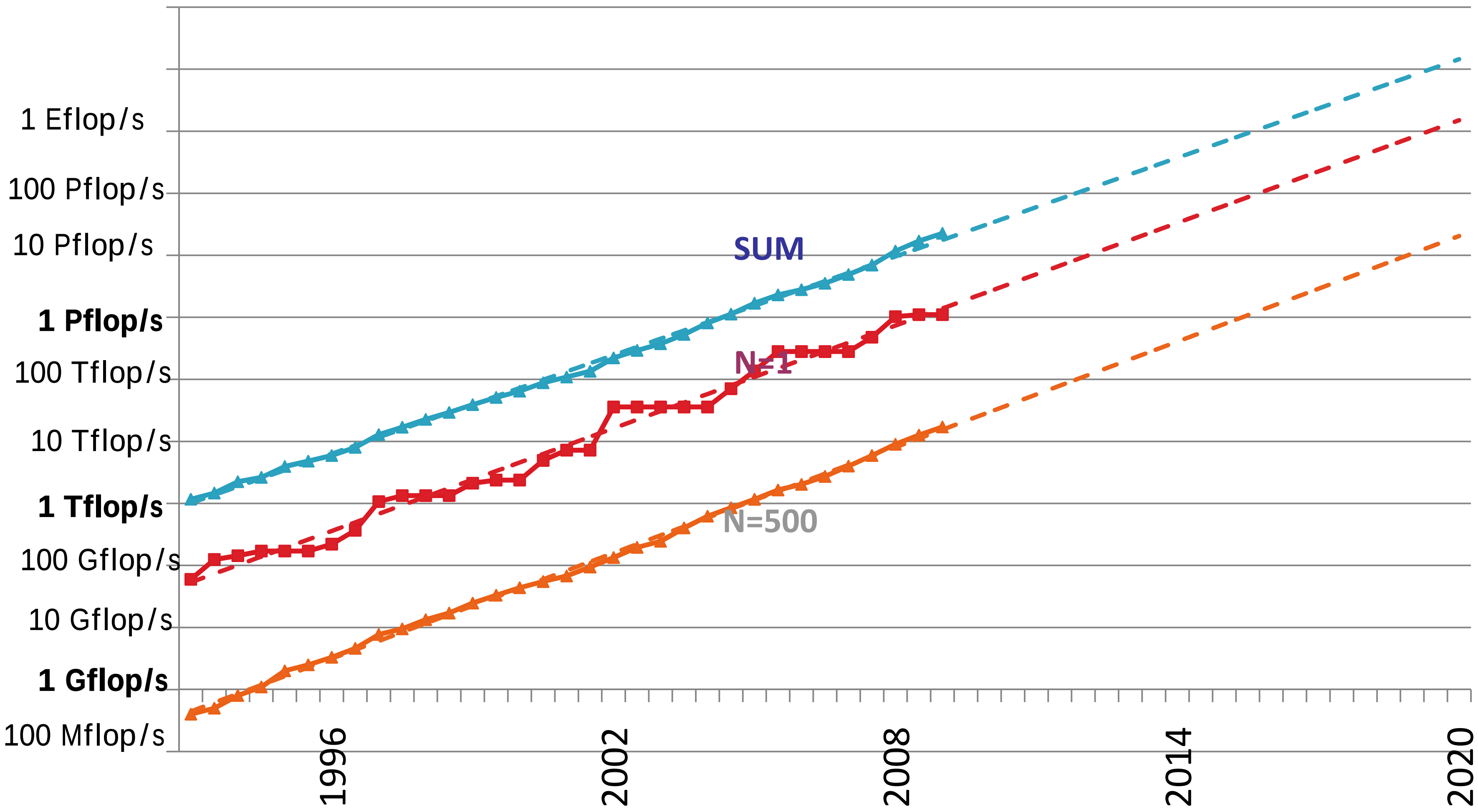
Multi-Core and Many-Core

- Power consumption of chips and systems has increased tremendously, because of ‘cheap’ exploitation of Moore’s Law.
  - Free lunch has ended
  - Stall of frequencies forces increasing concurrency levels, Multi-Cores
  - Optimal core sizes/power are smaller than current ‘rich cores’, which leads to Many-Cores
- Many-Cores, more (10-100x) but smaller cores:
  - Intel Polaris – 80 cores,
  - Clearspeed CSX600 – 96 cores,
  - nVidia G80 – 128 cores, or
  - CISCO Metro – 188 cores

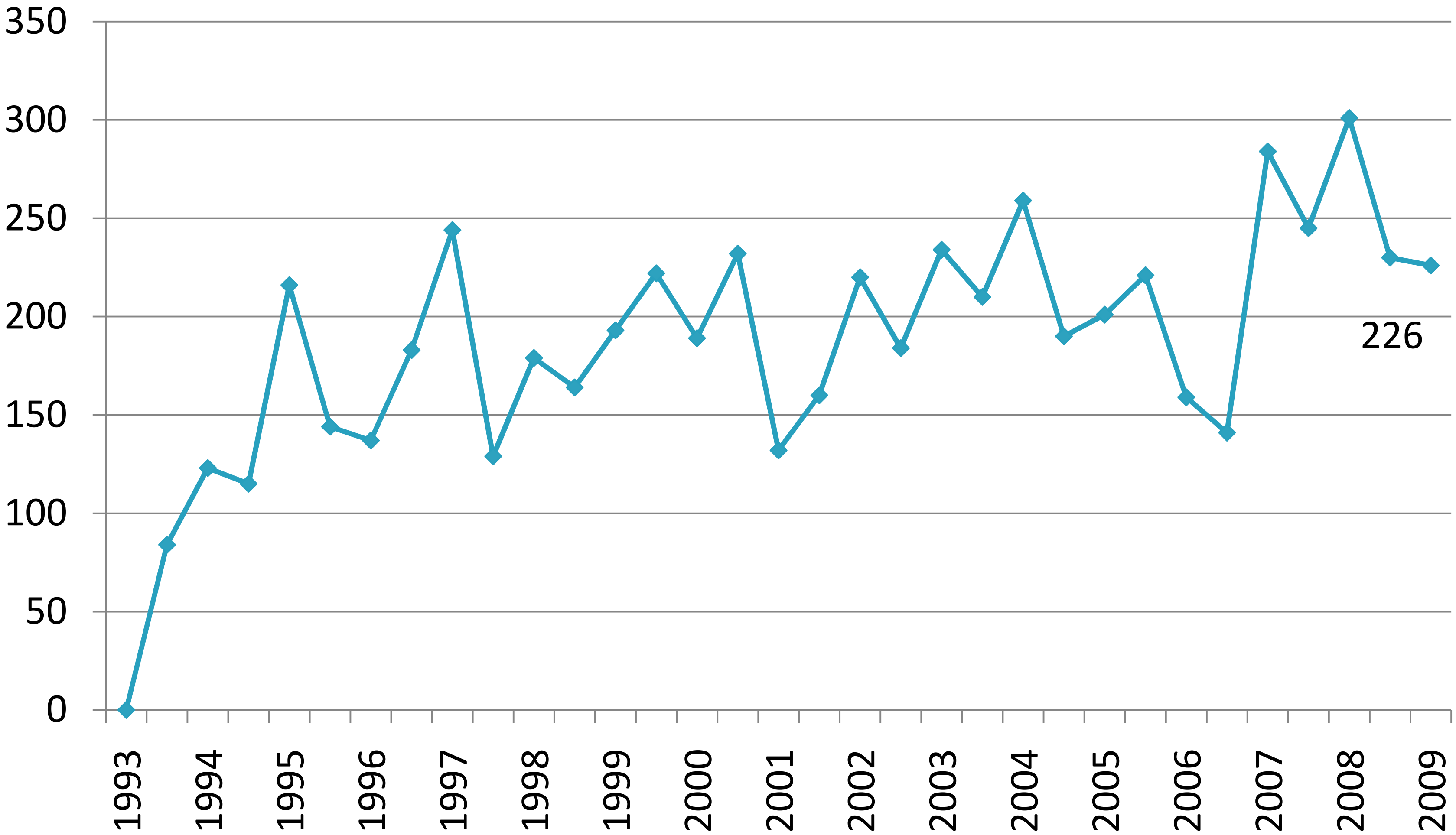
Performance Development



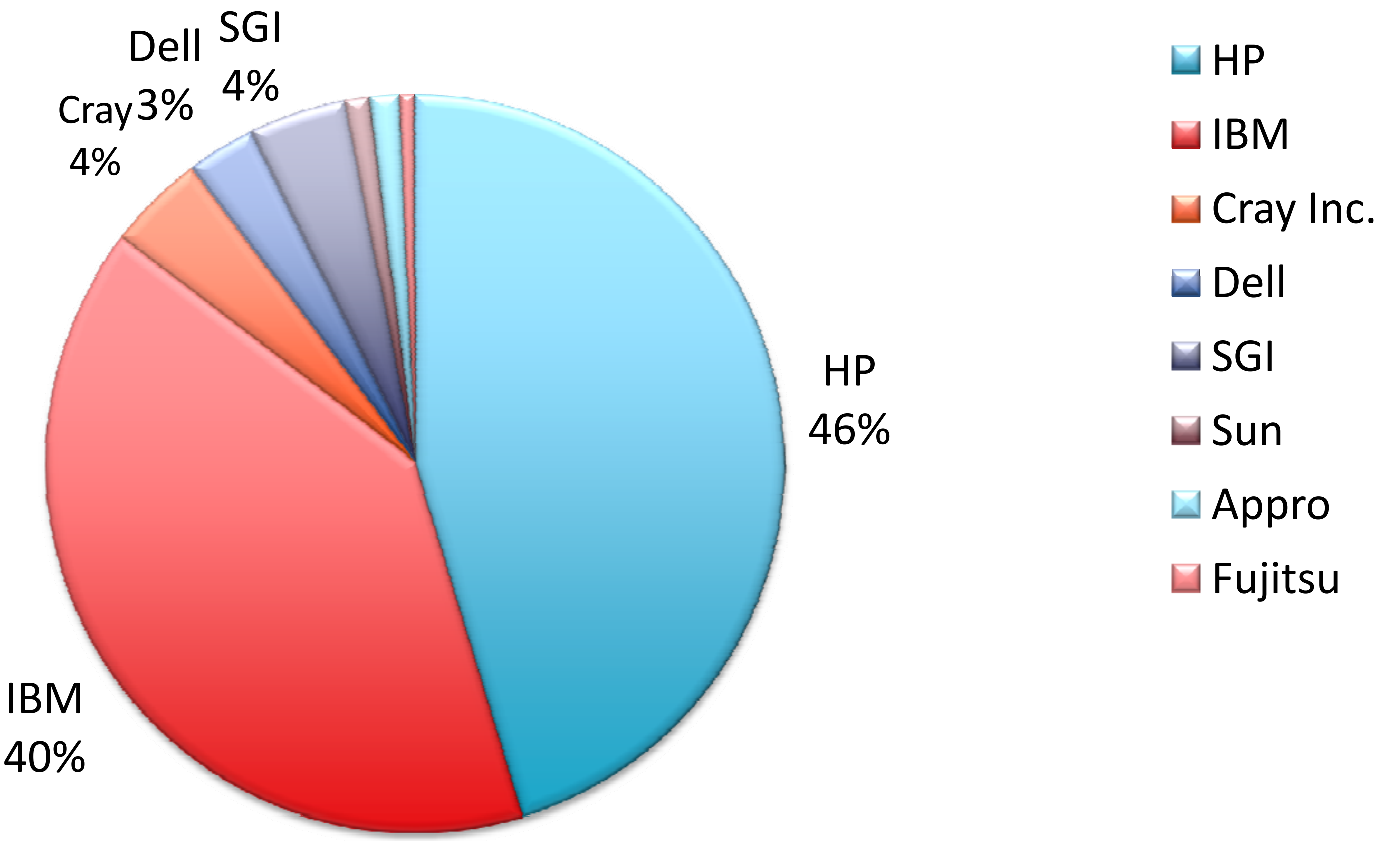
Performance Development



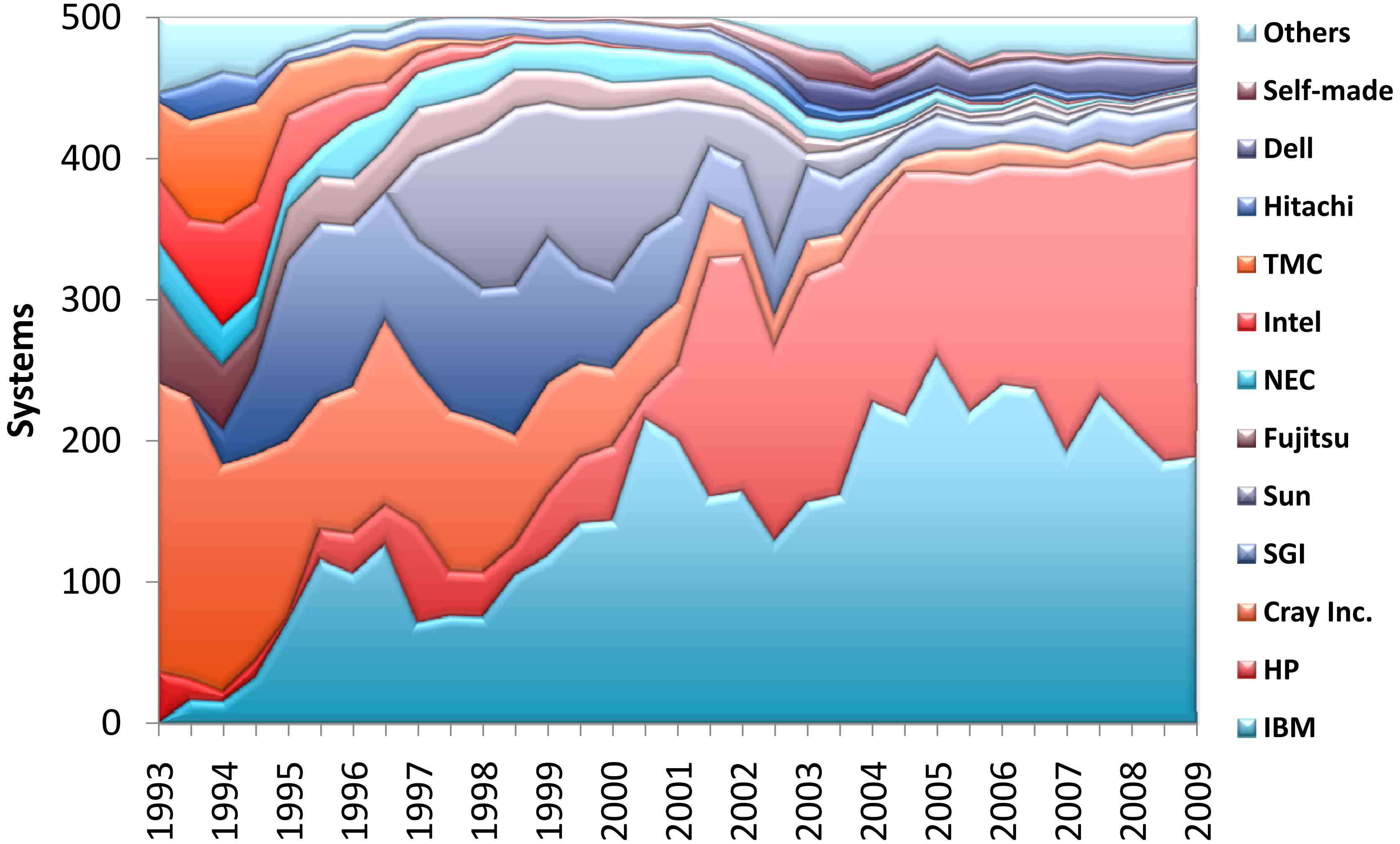
Replacement Rate



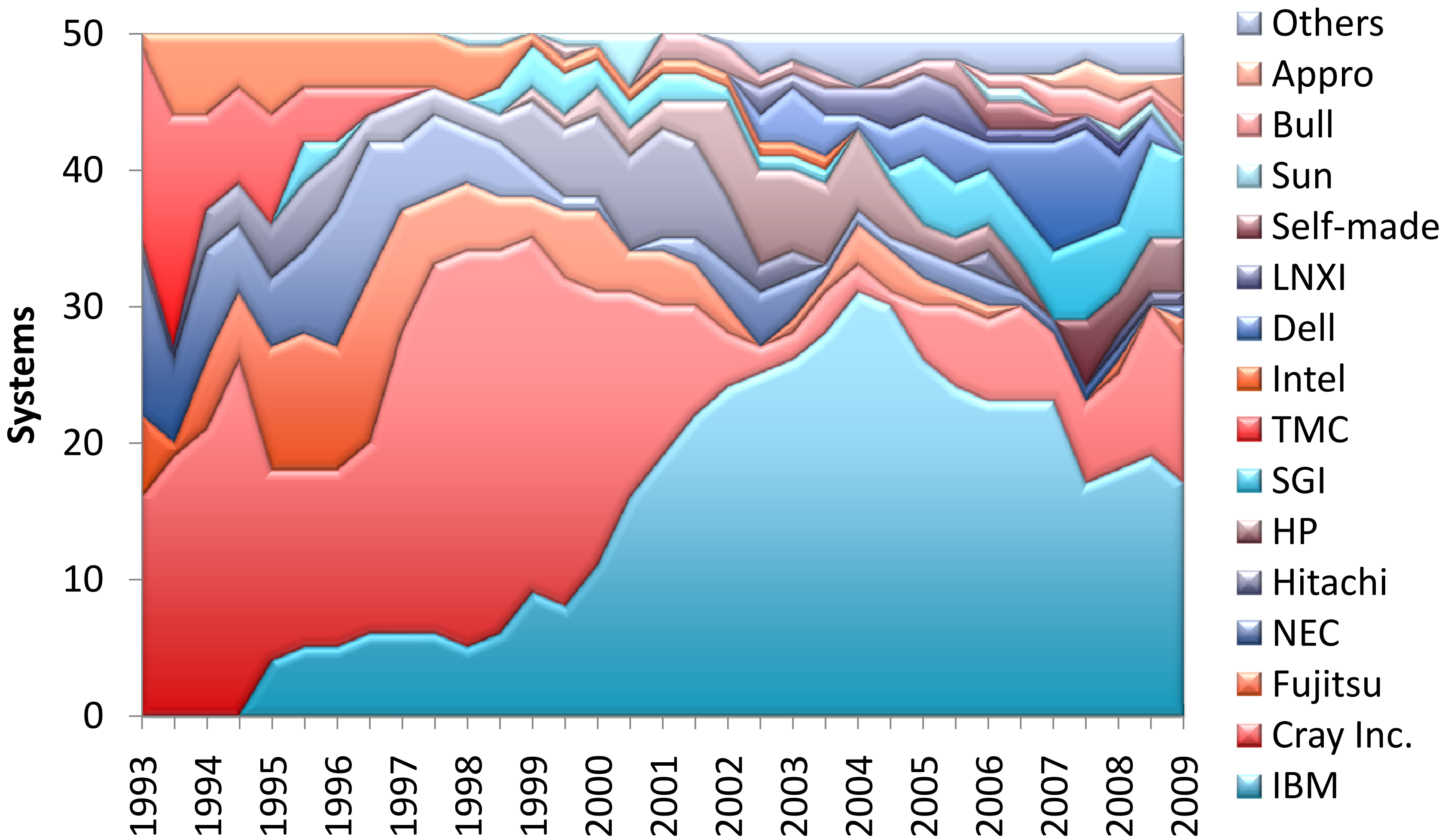
Vendors / System Share



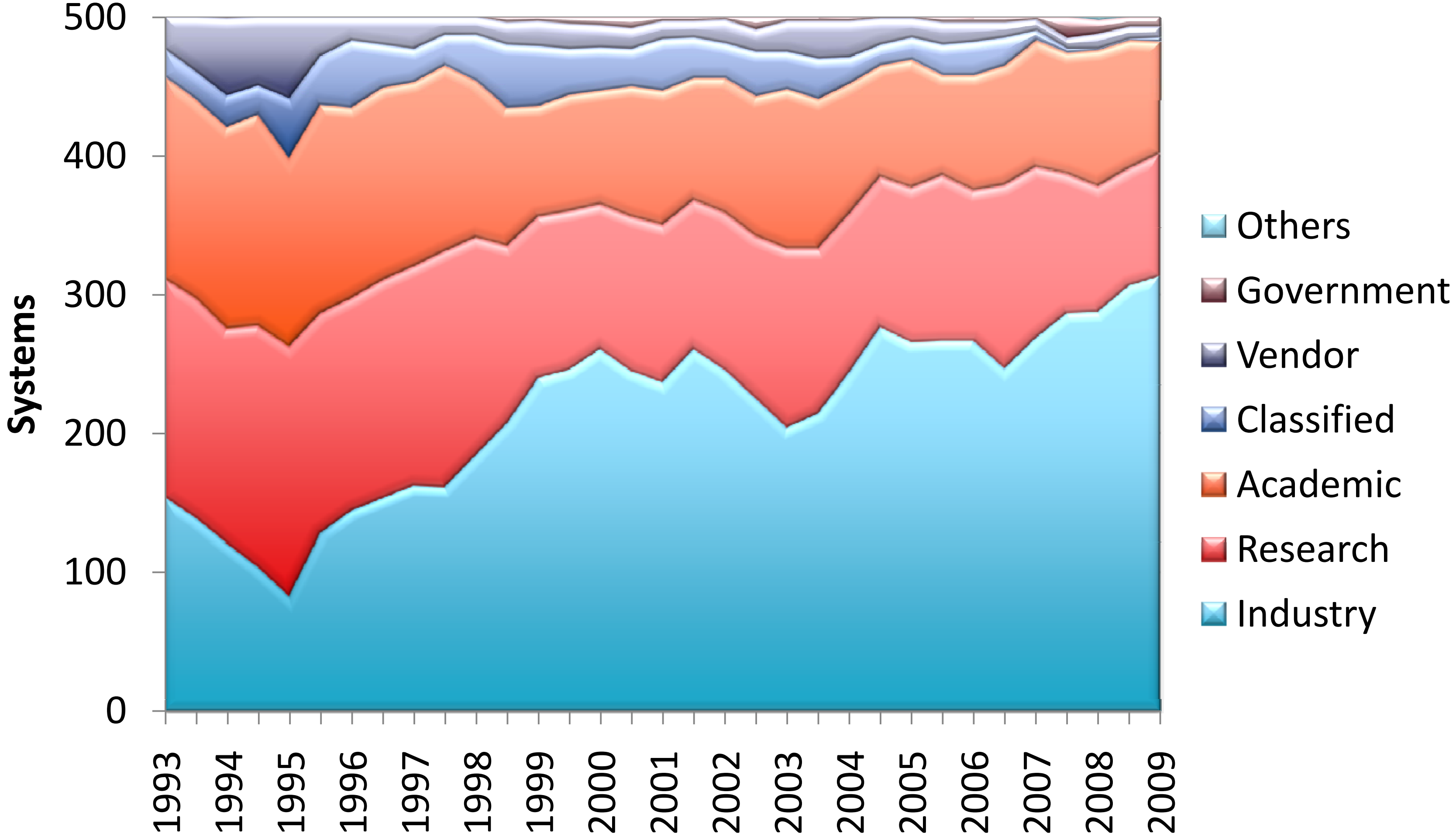
Vendors


Systems
500
400
300
200
100
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Others
Self-made
Dell
Hitachi
TMC
Intel
NEC
Fujitsu
Sun
SGI
Cray Inc.
HP
IBM

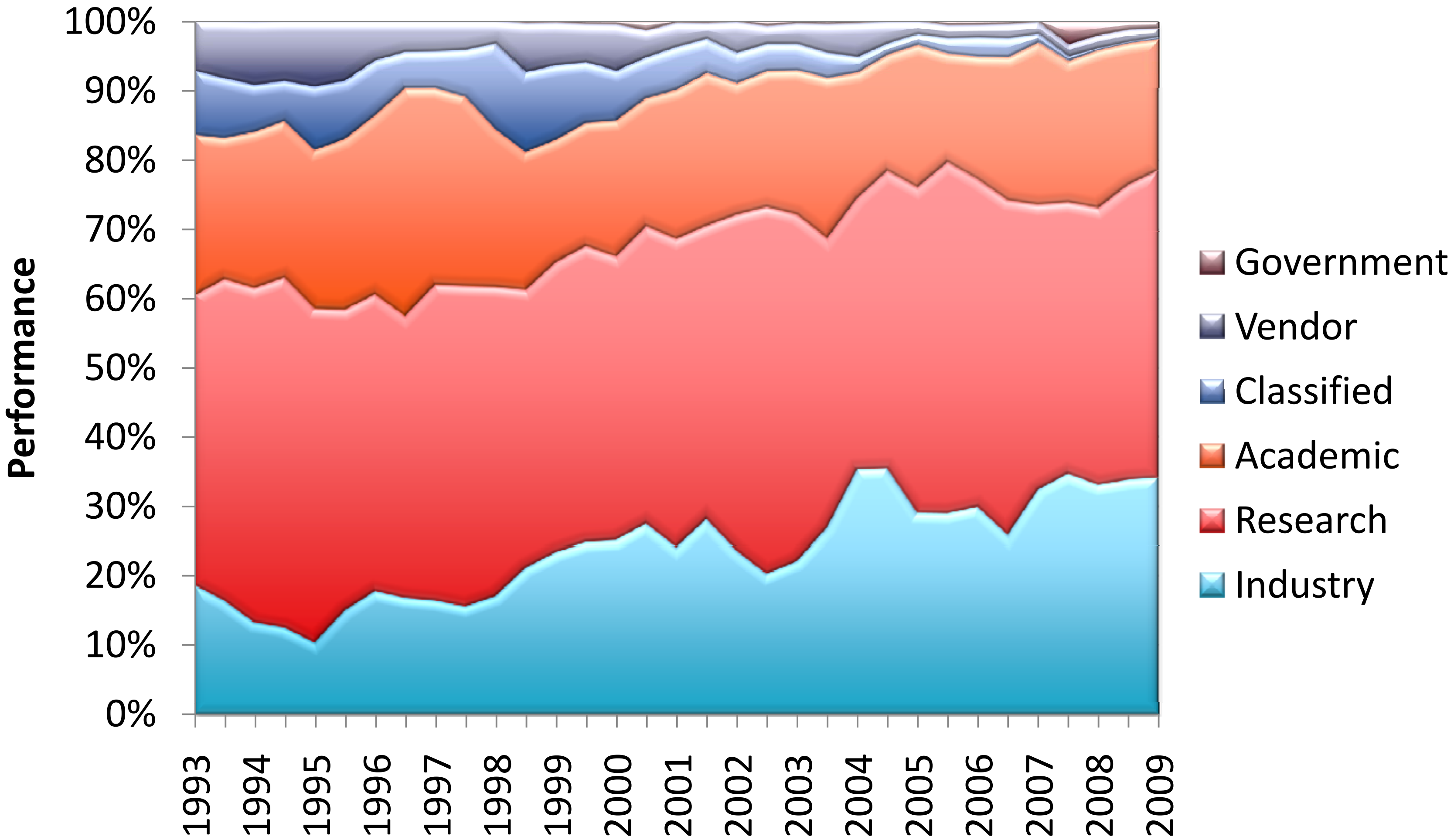
Vendors (TOP50)



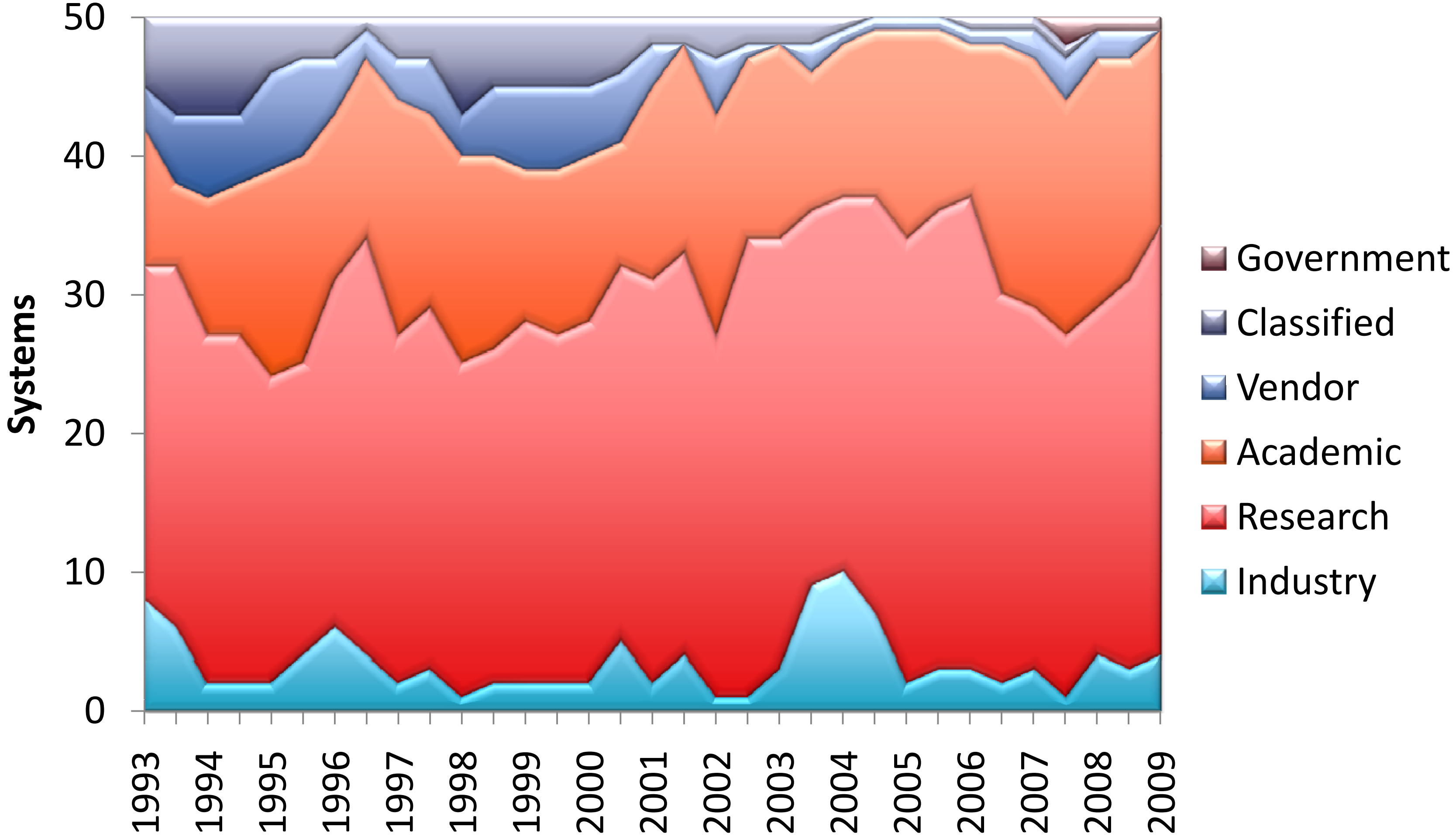
Customer Segments



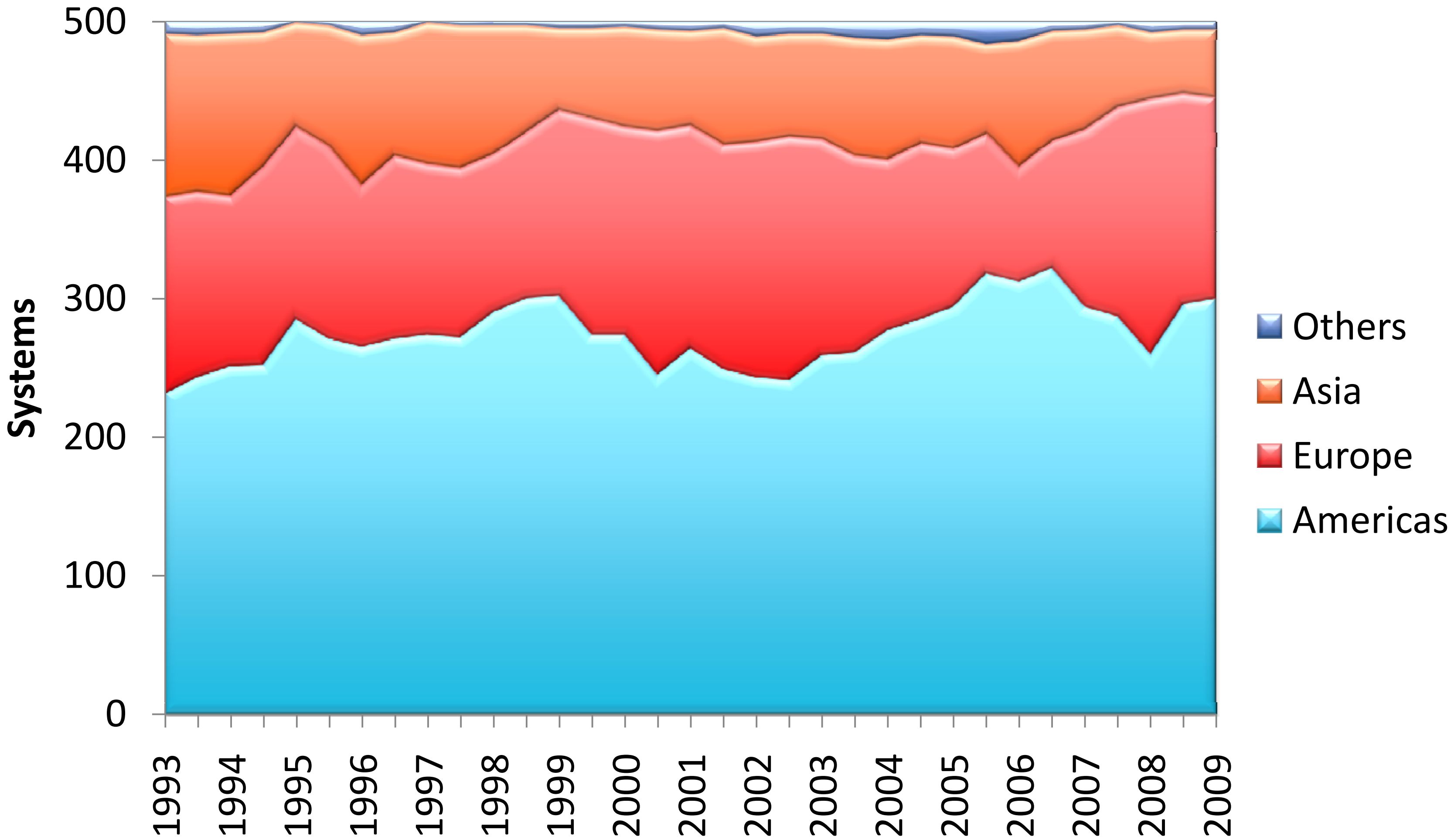
Customer Segments



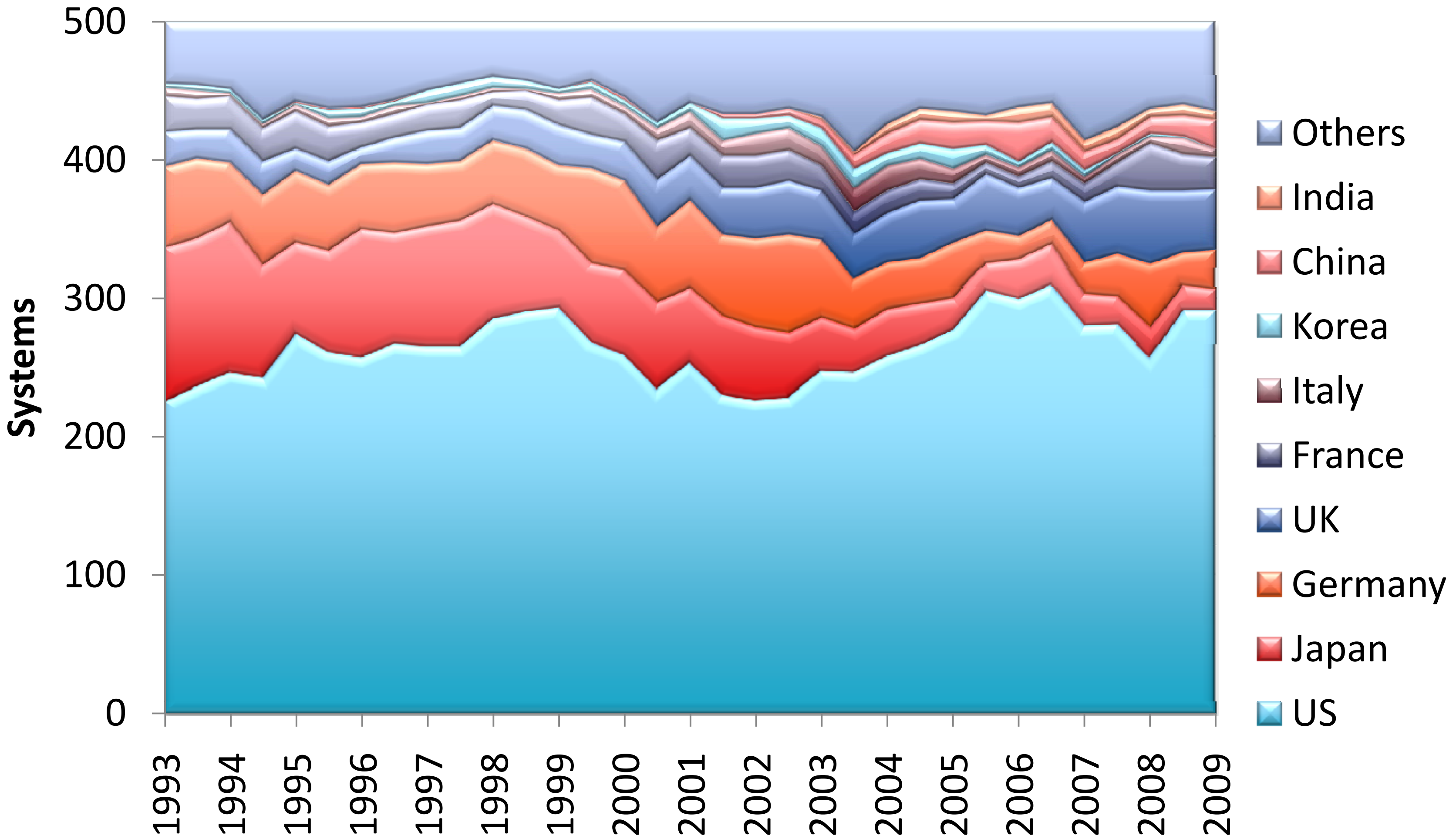
Customer Segments (TOP50)



Continents

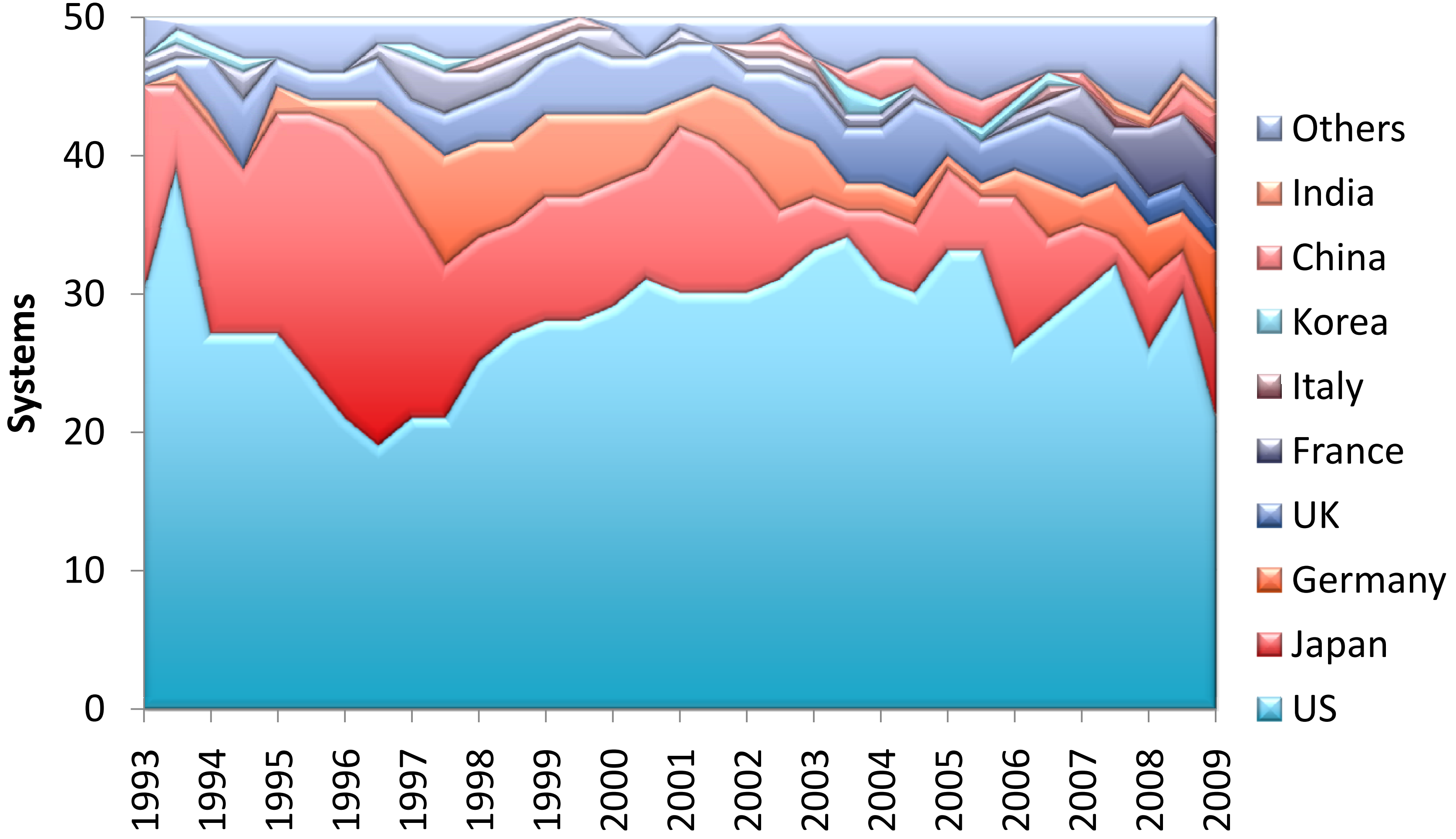

Systems
500
400
300
200
100
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Others
Asia
Europe
Americas

Countries

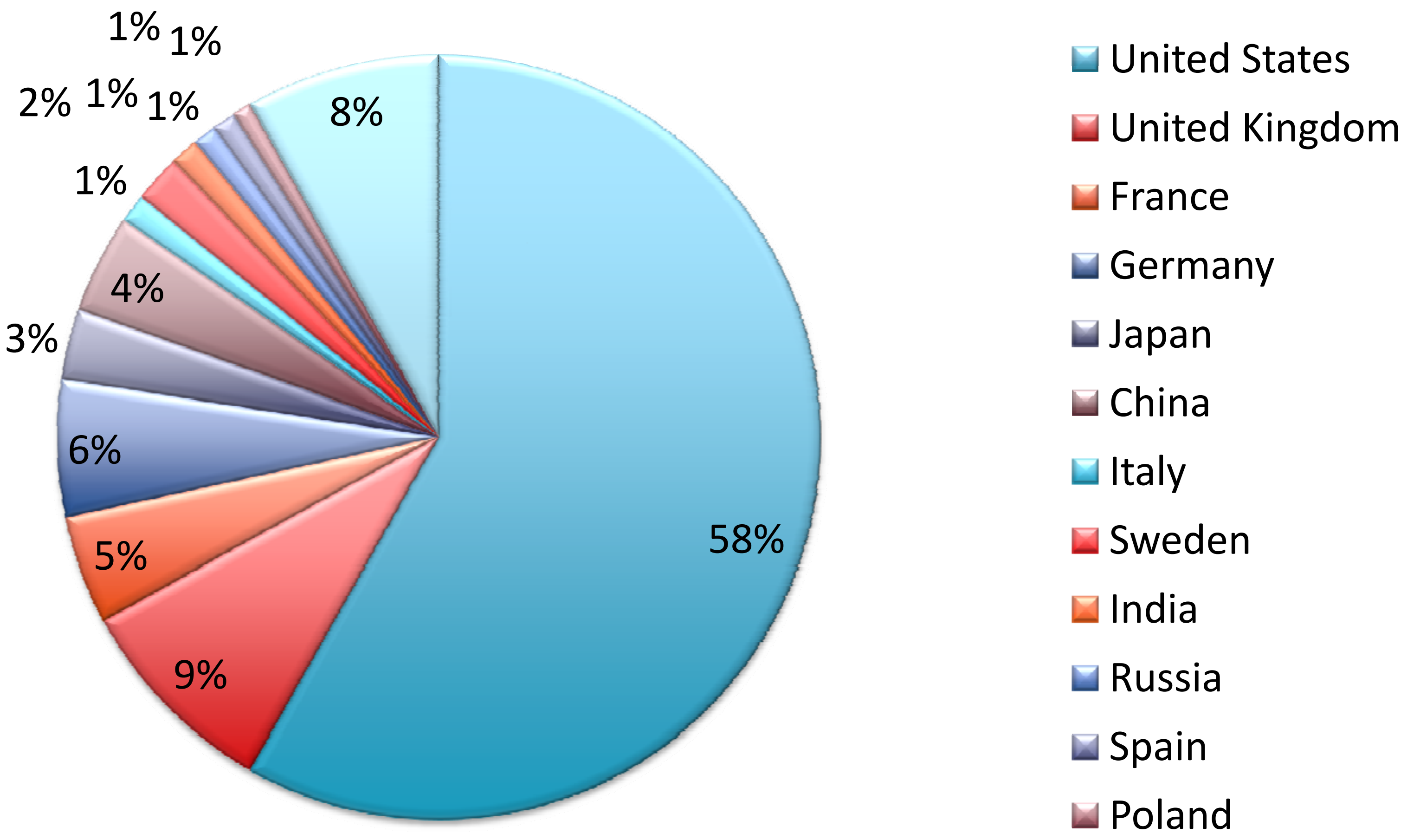


Systems
500
400
300
200
100
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Others
India
China
Korea
Italy
France
UK
Germany
Japan
US

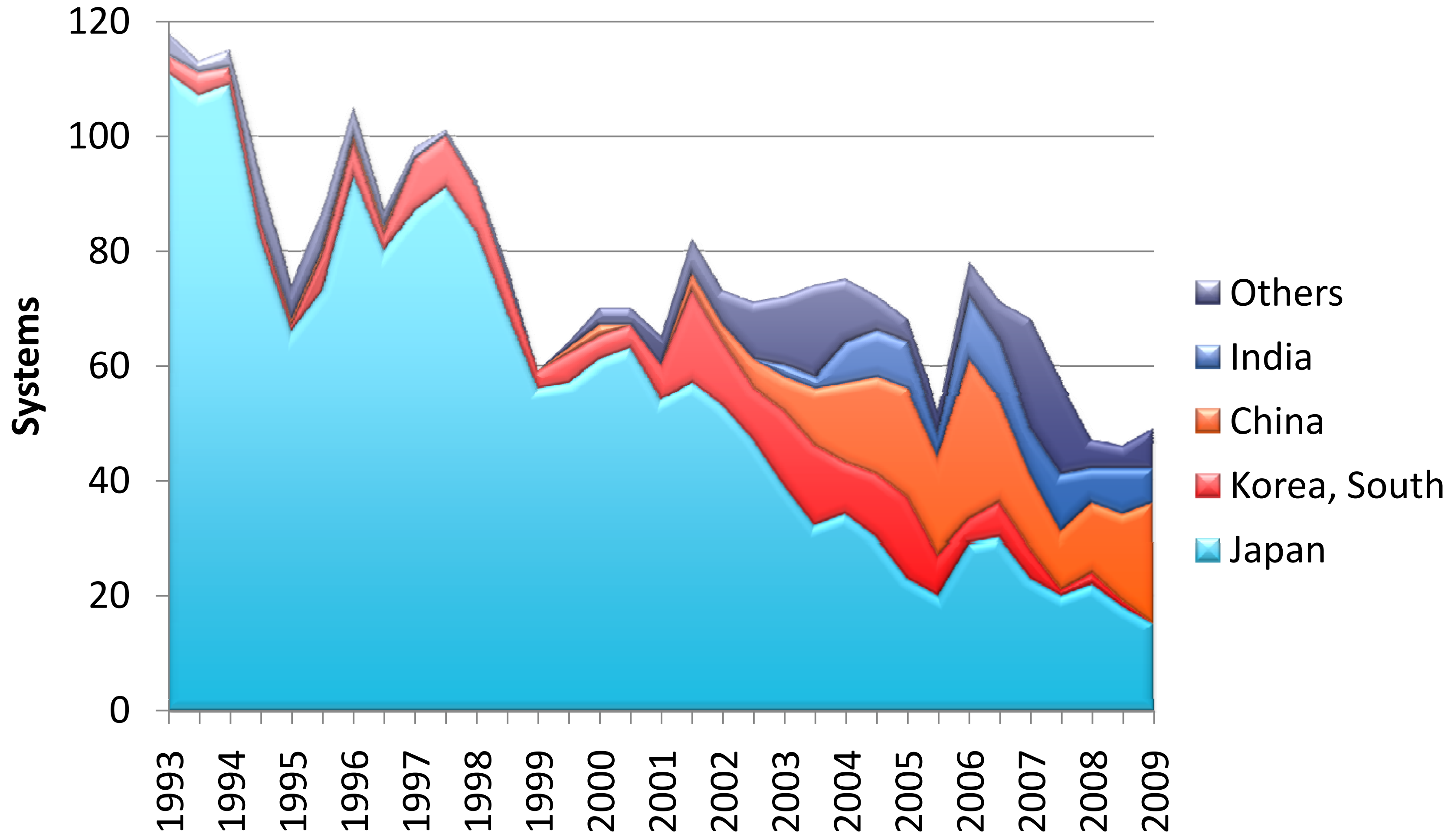
Countries (TOP50)



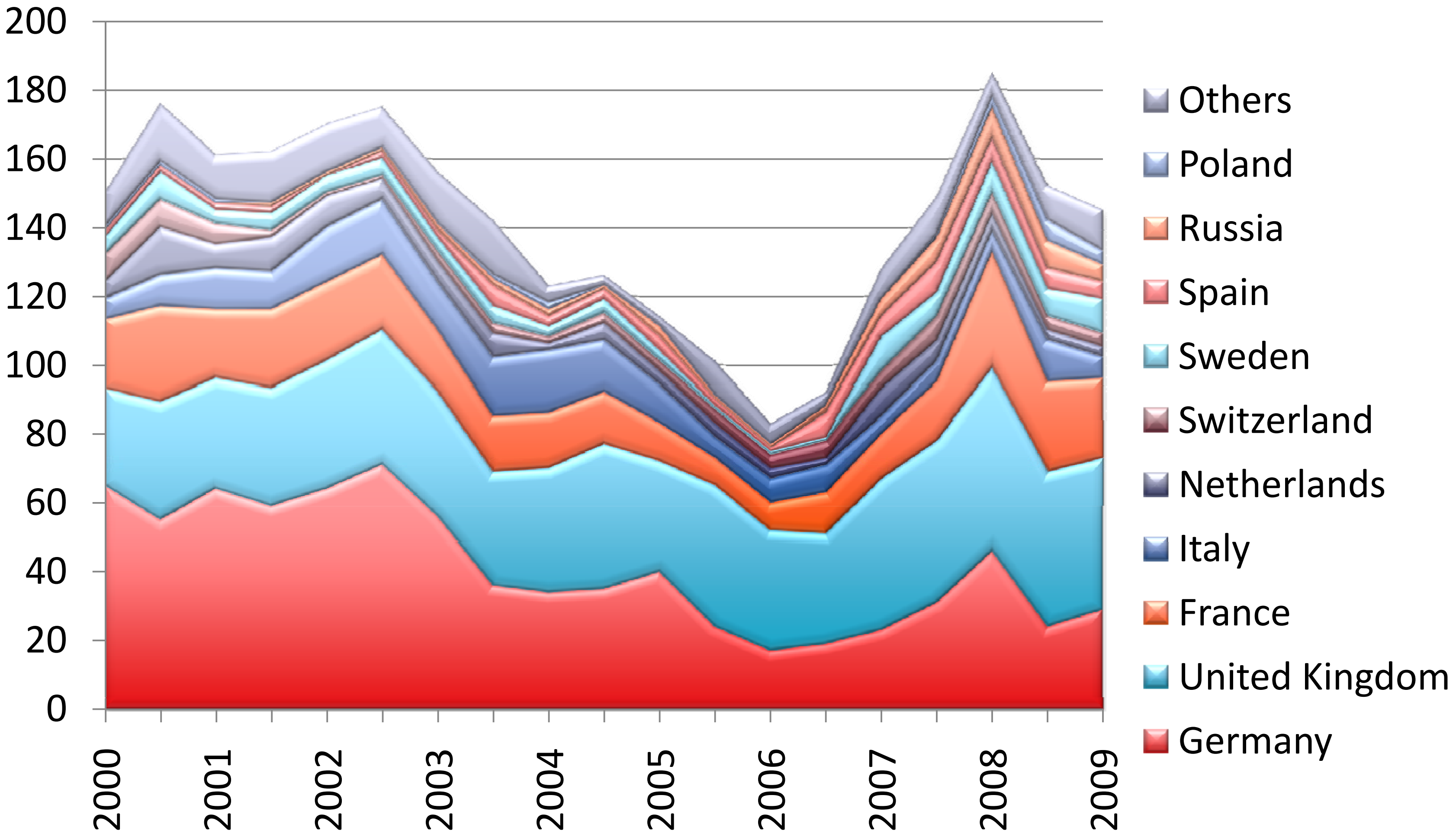
Countries / System Share



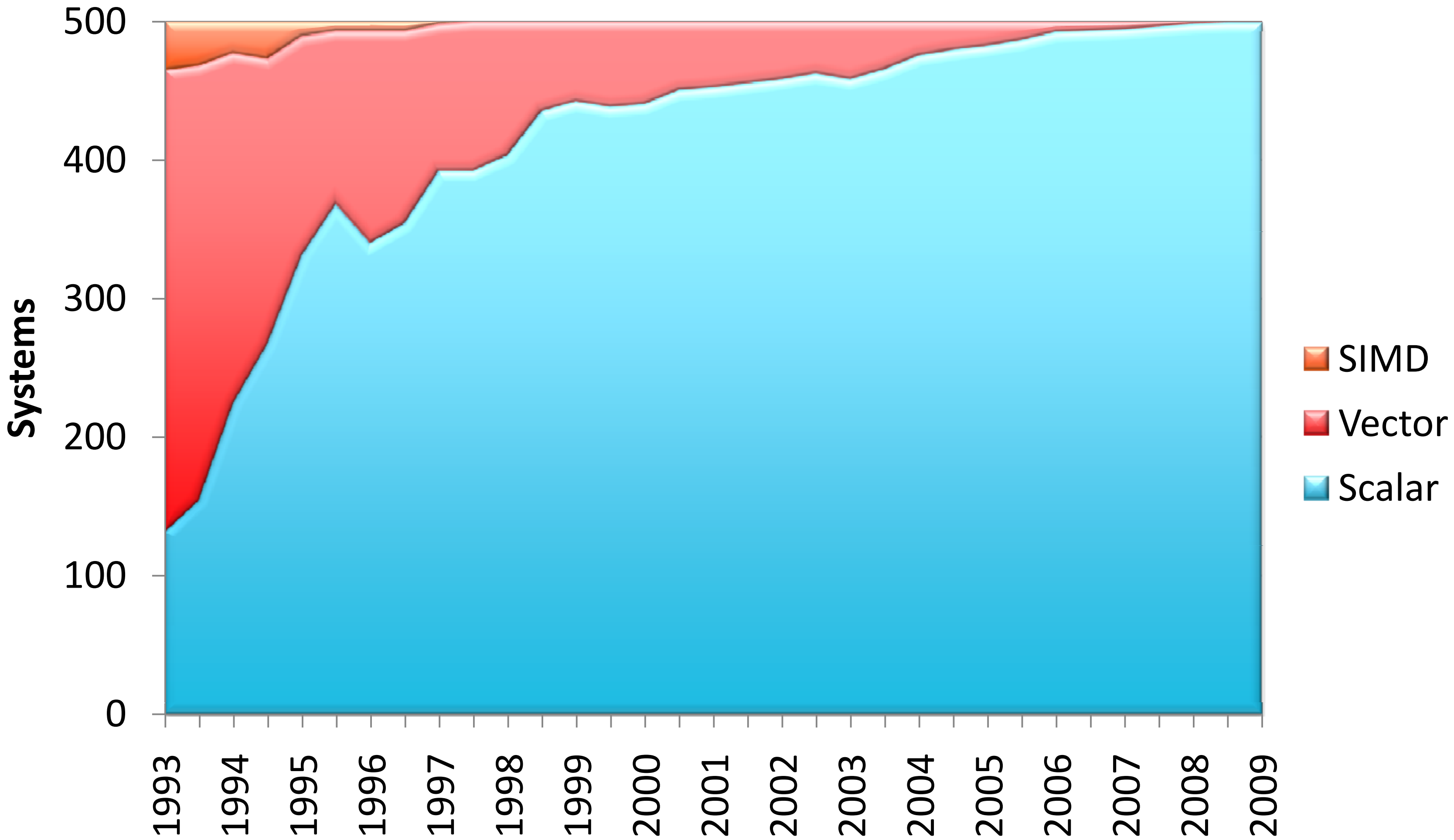
Asian Countries



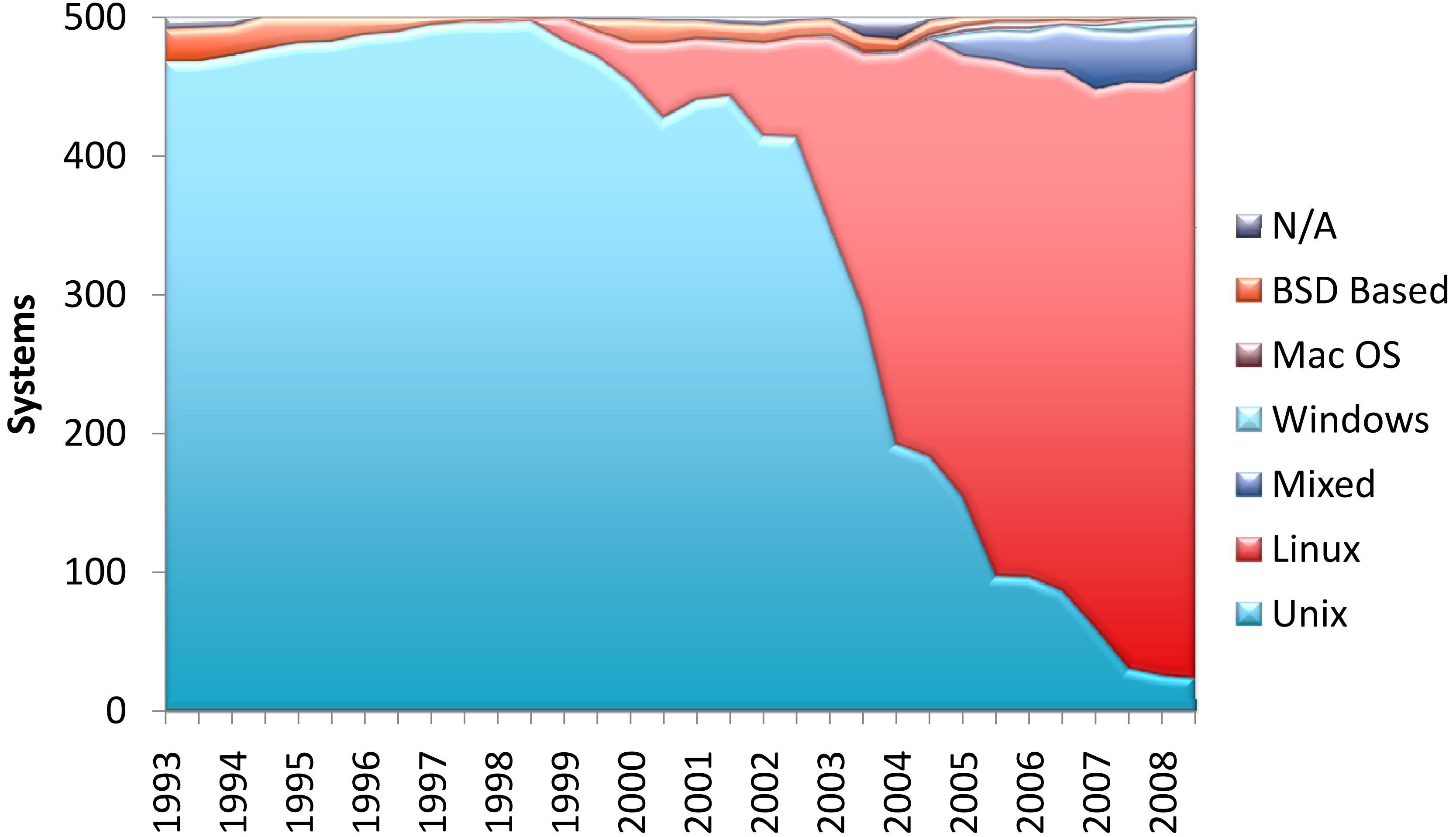
European Countries



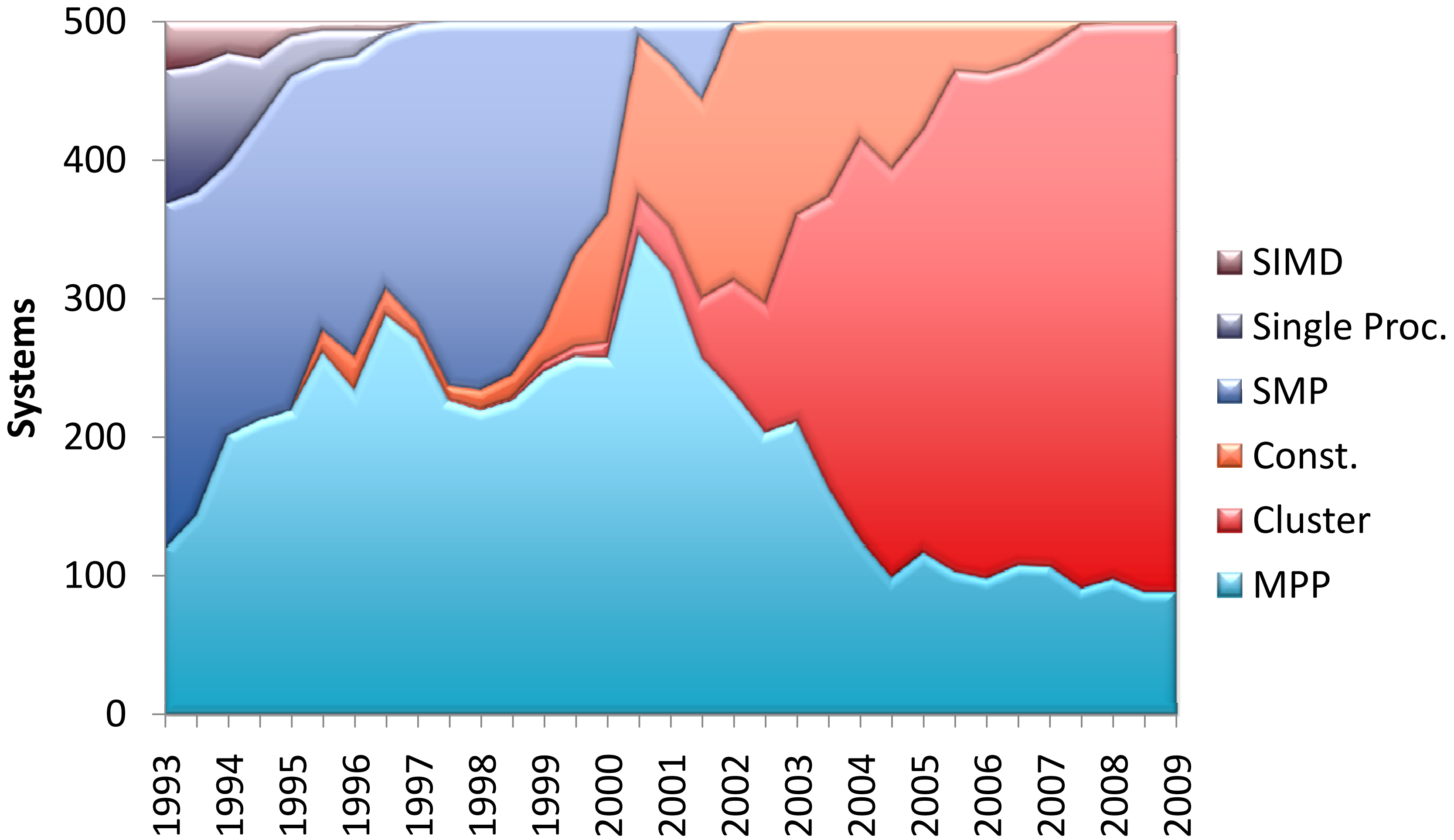
Processor Architecture / Systems



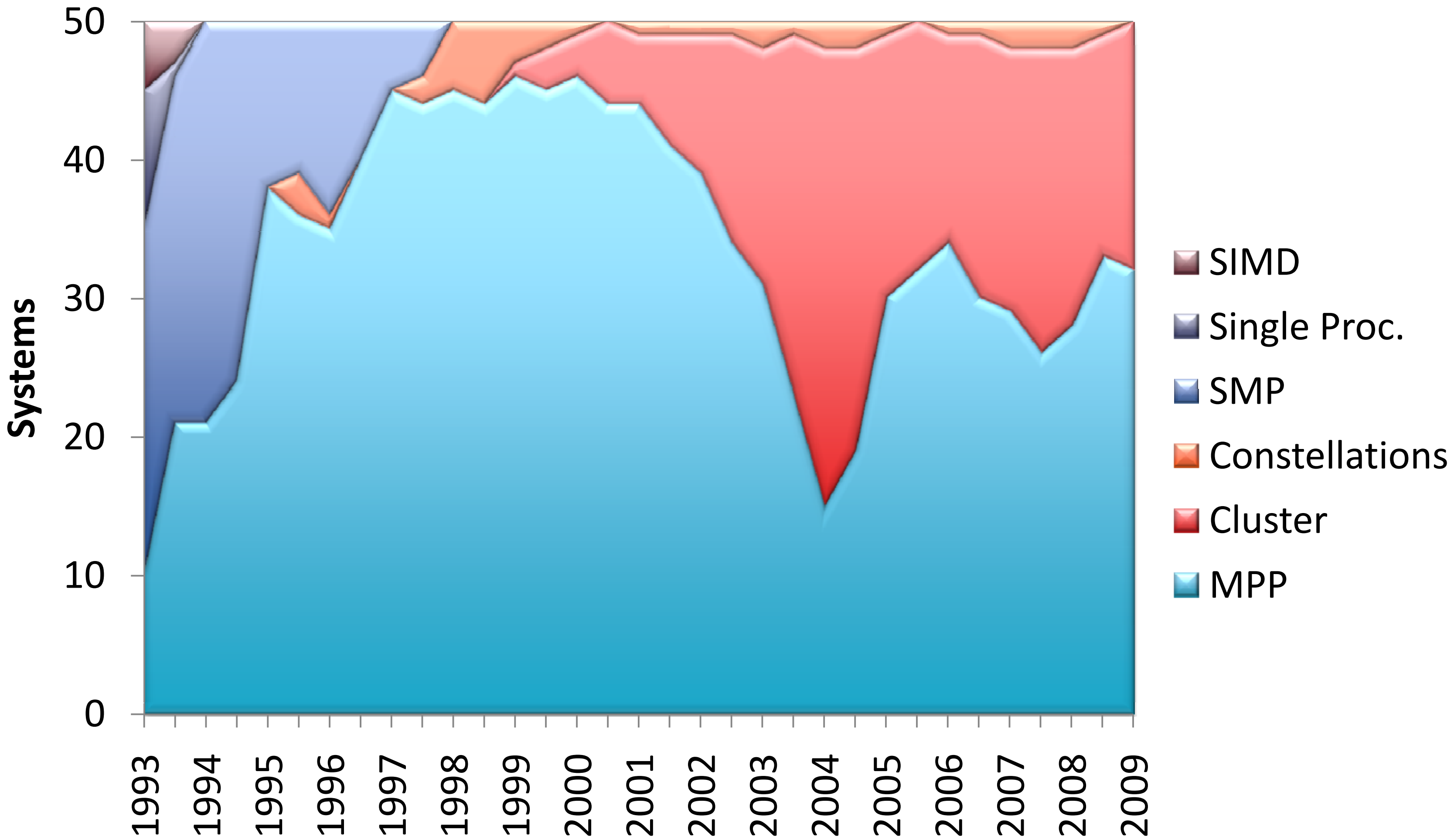
Operating Systems



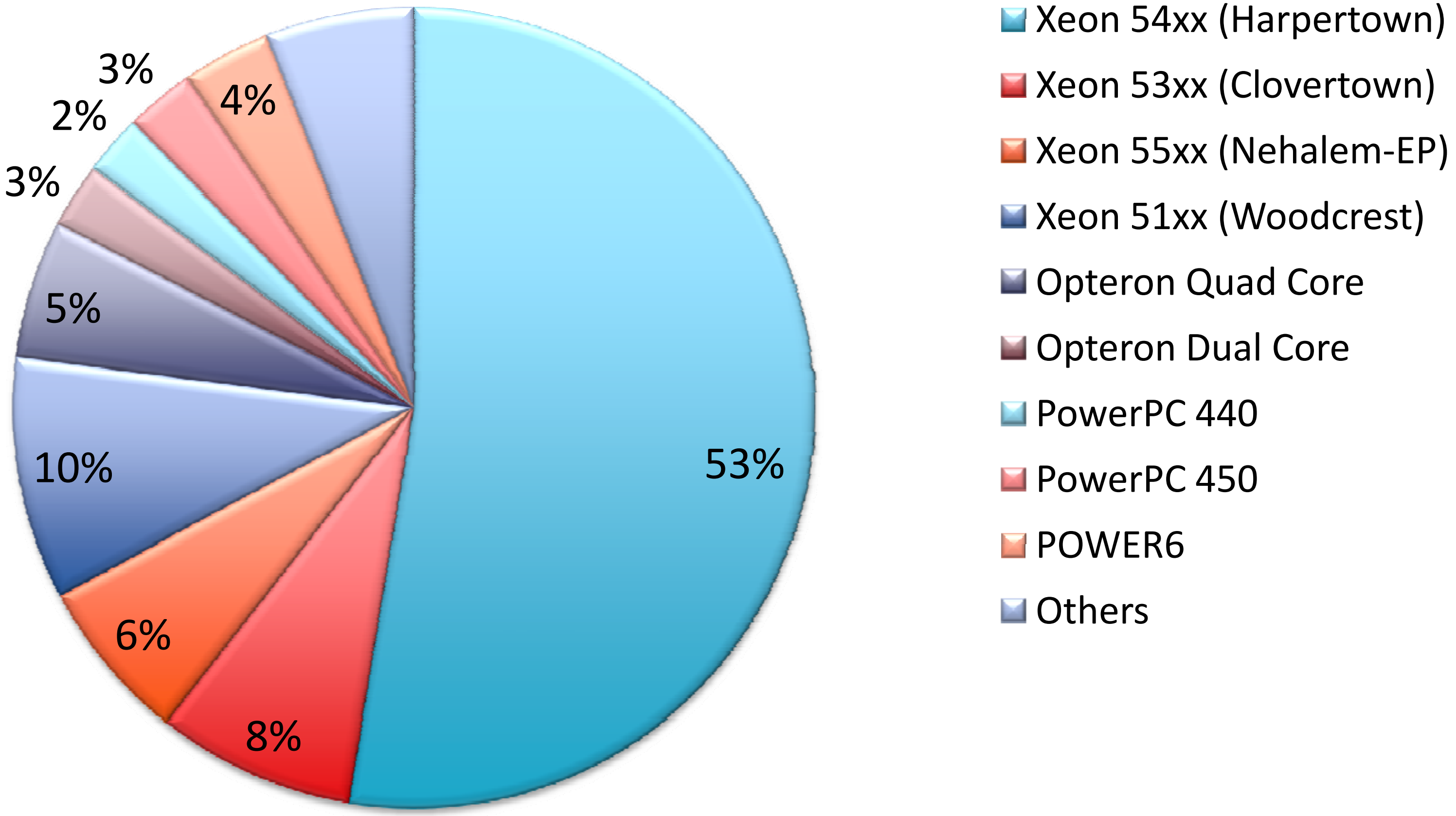
Architectures



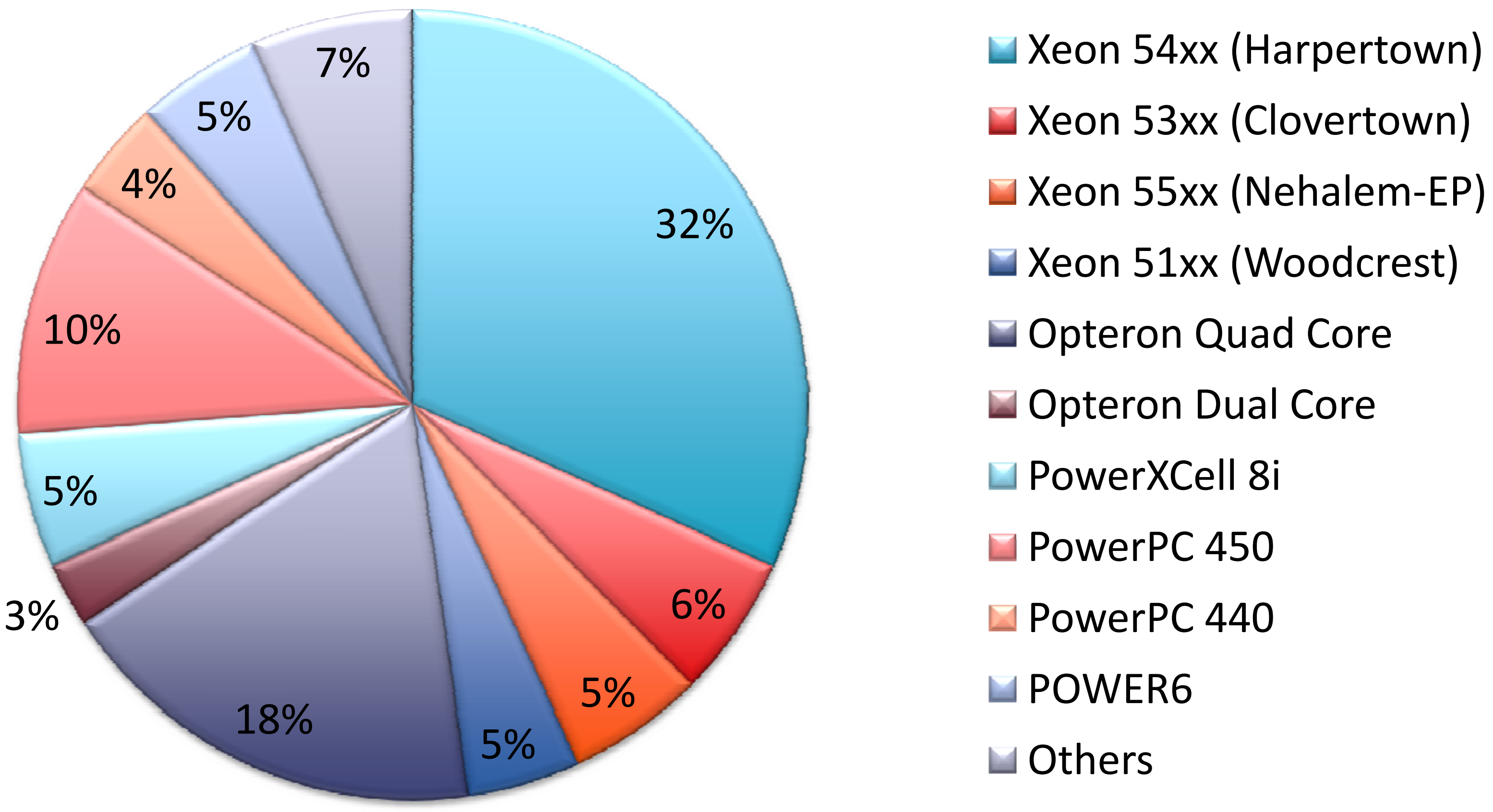
Architectures (TOP50)



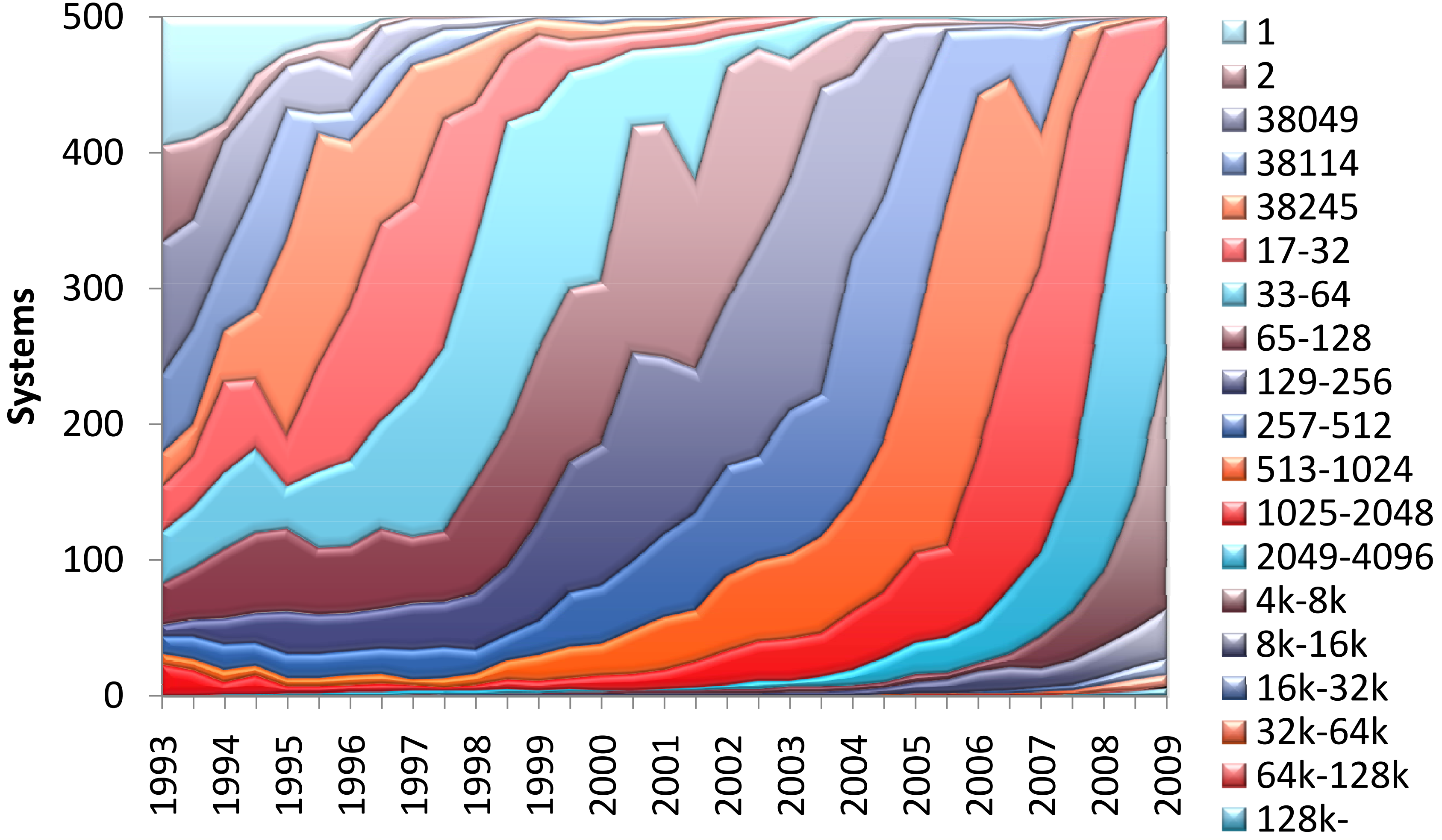
Processors / Systems



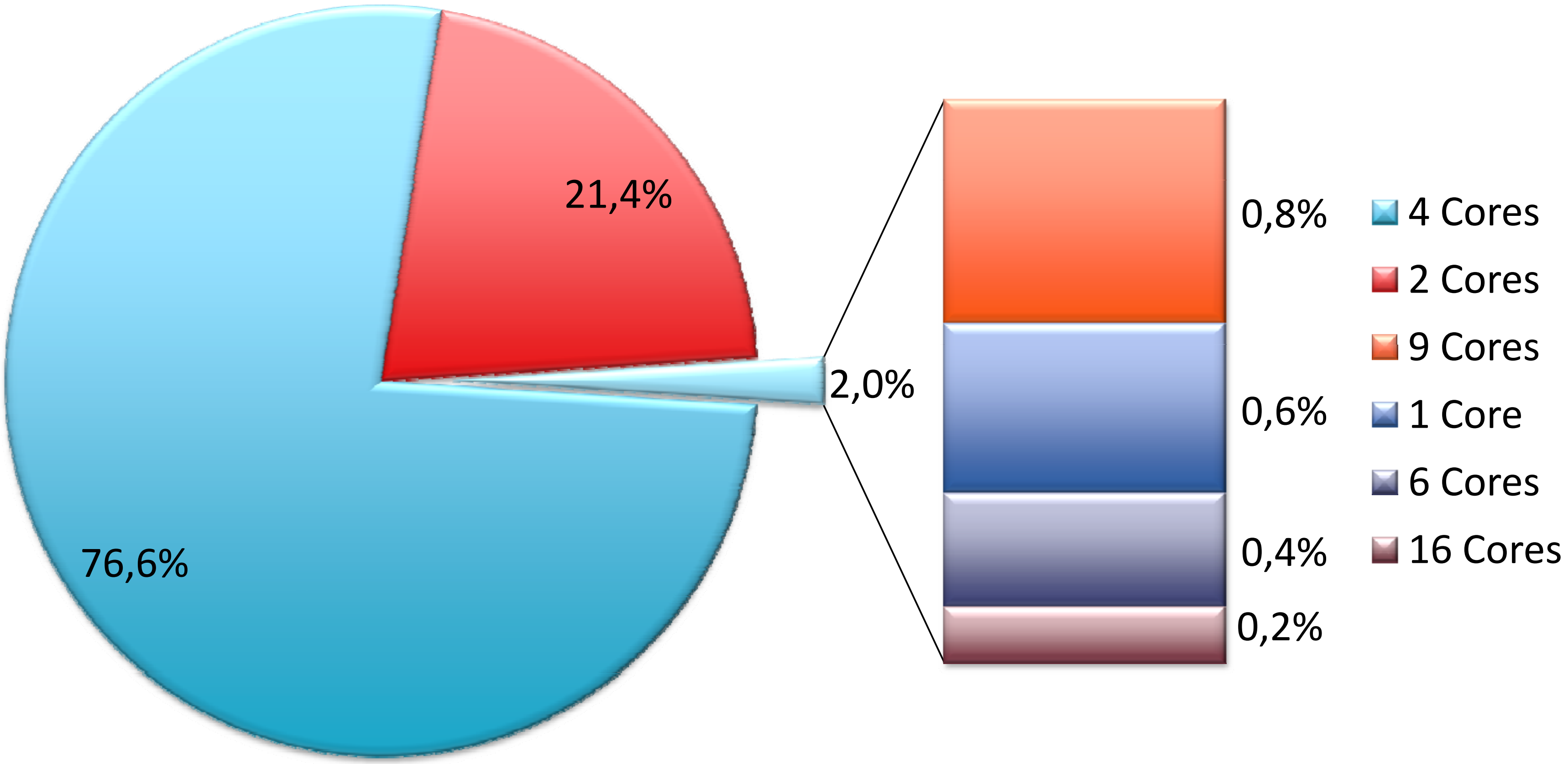
Processors / Performance



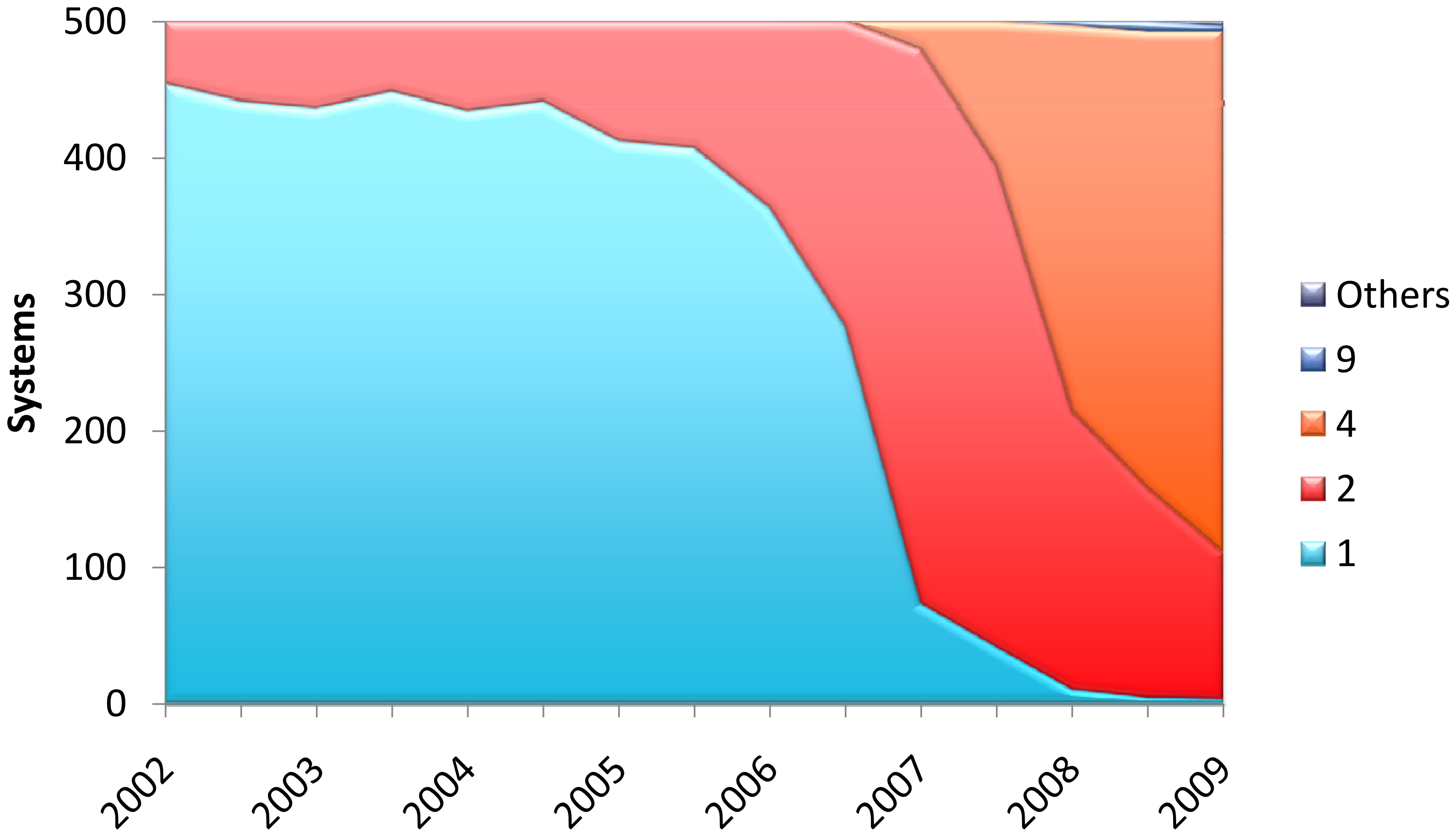
Core Count


Systems
500
400
300
200
100
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
1
2
38049
38114
38245
17-32
33-64
65-128
129-256
257-512
513-1024
1025-2048
2049-4096
4k-8k
8k-16k
16k-32k
32k-64k
64k-128k
128k-

Cores per Socket

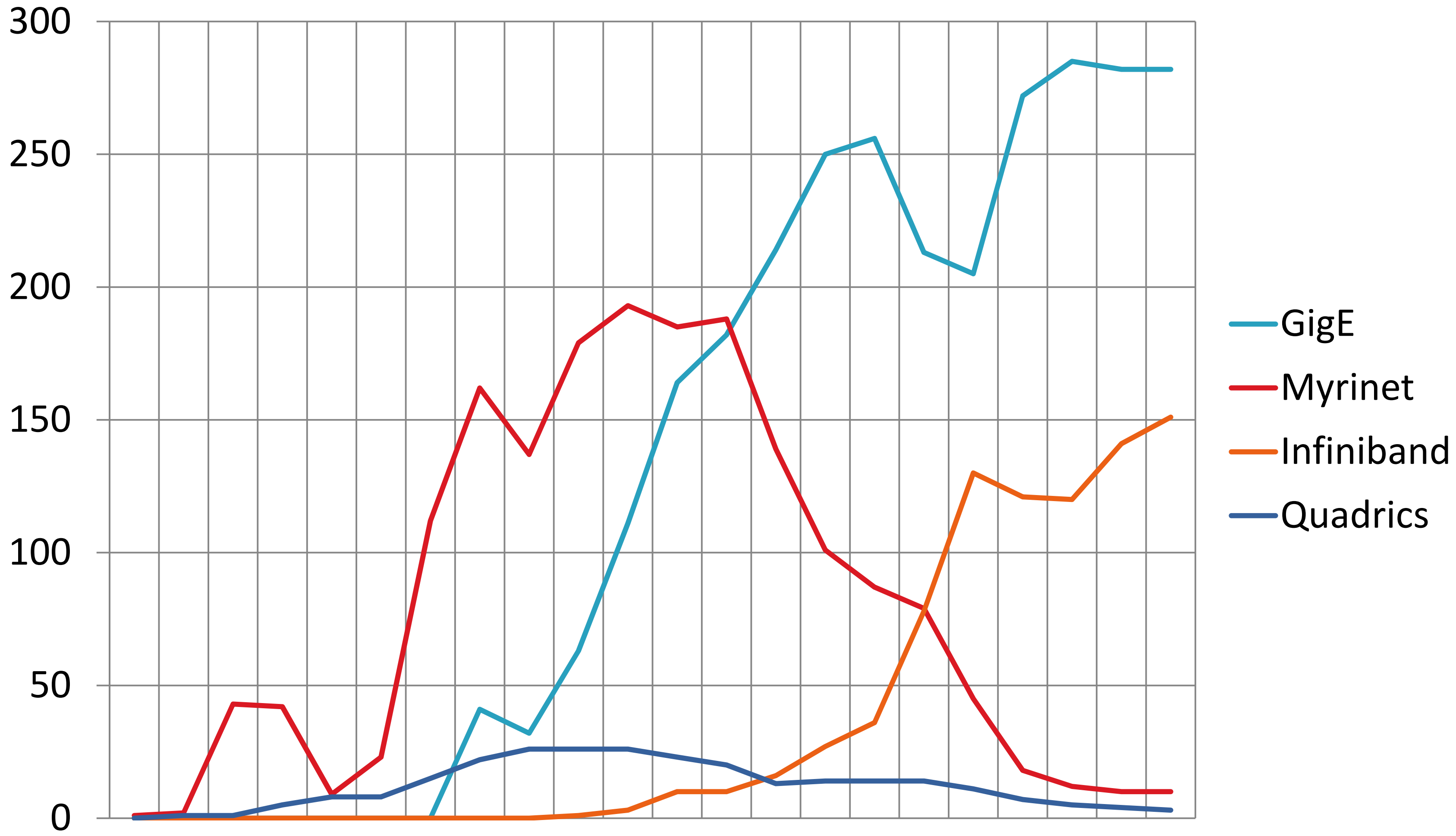


Cores per Socket

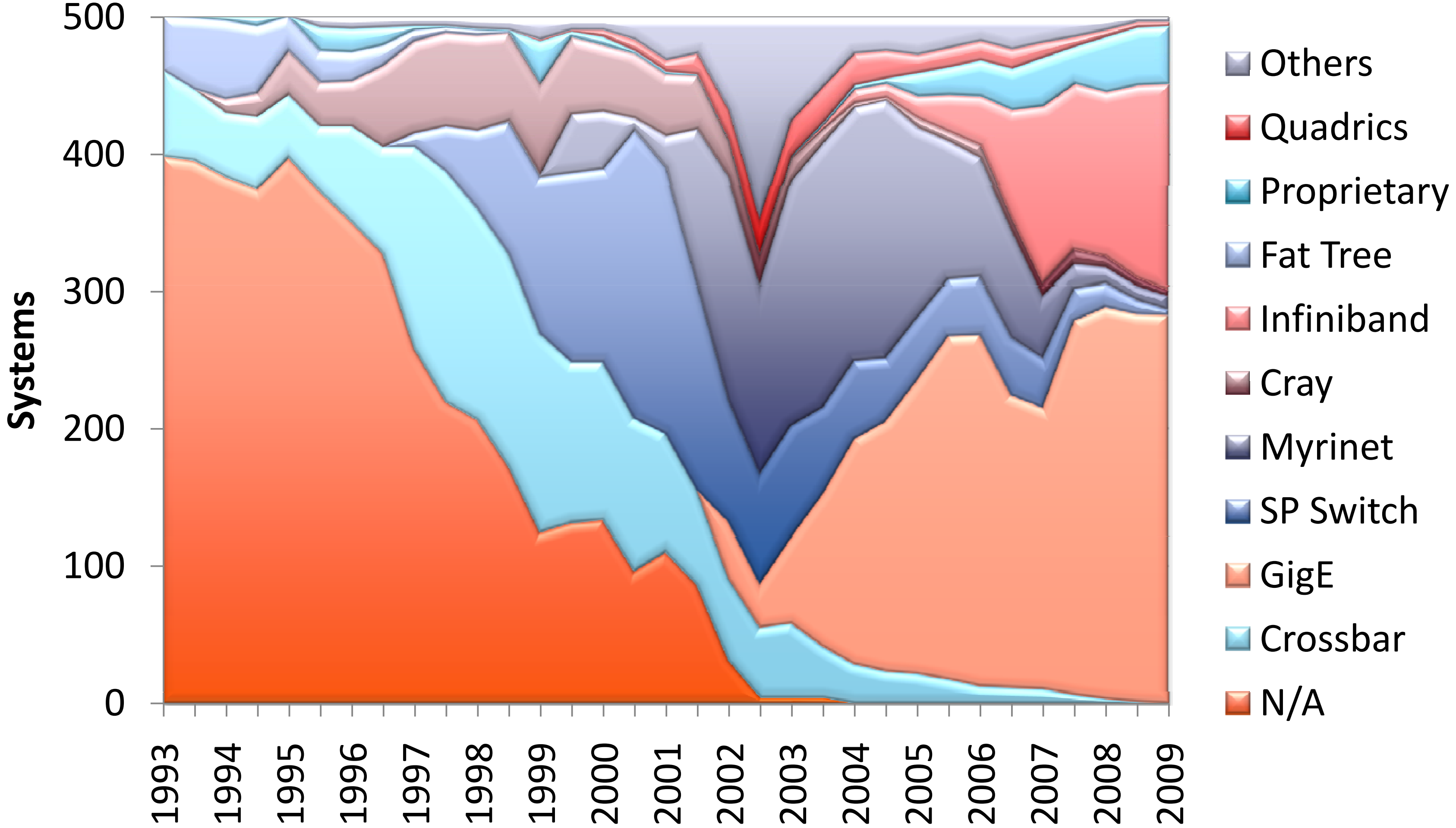


TOP500® SUPERCOMPUTER SITES

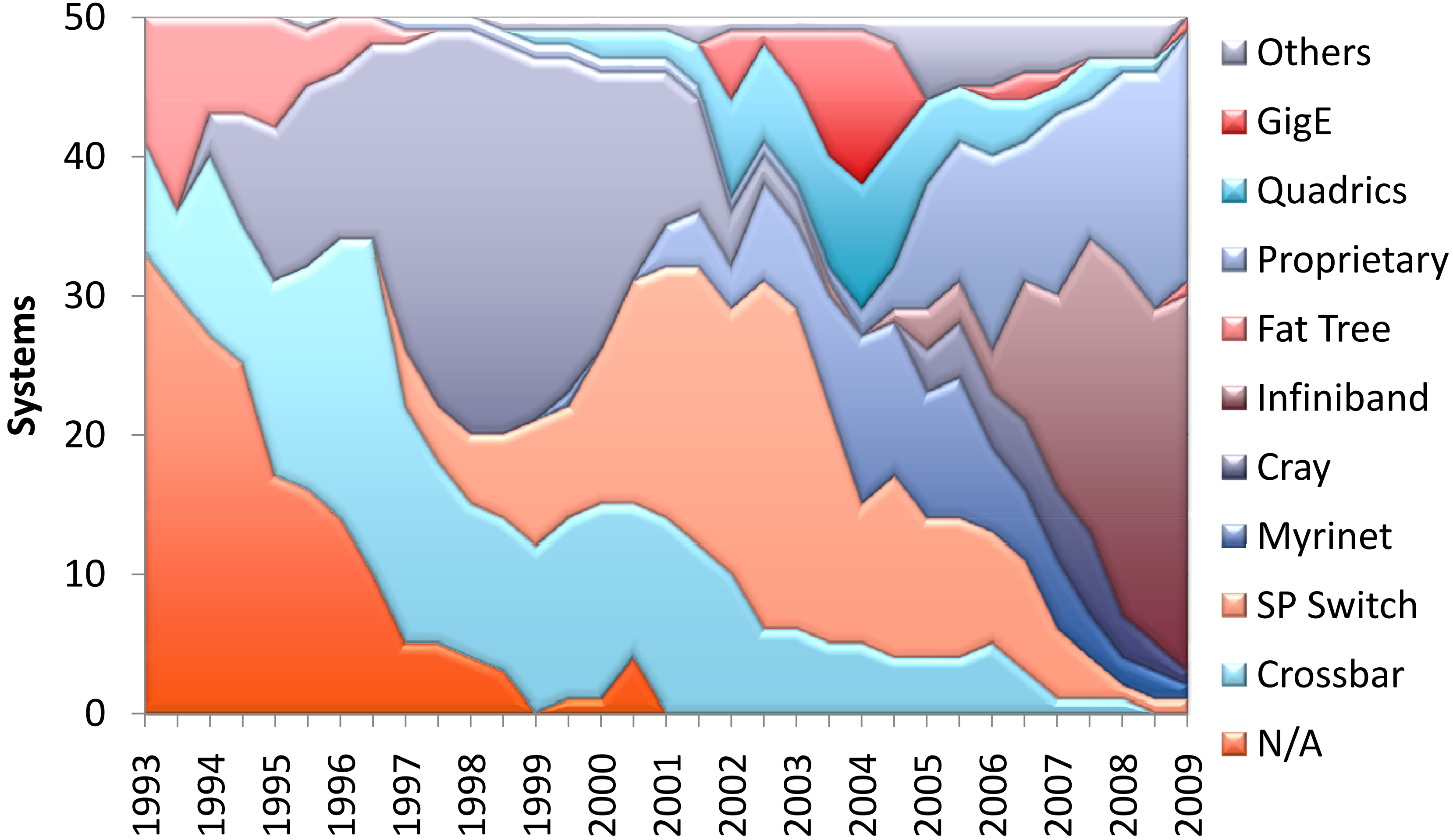
Cluster Interconnects



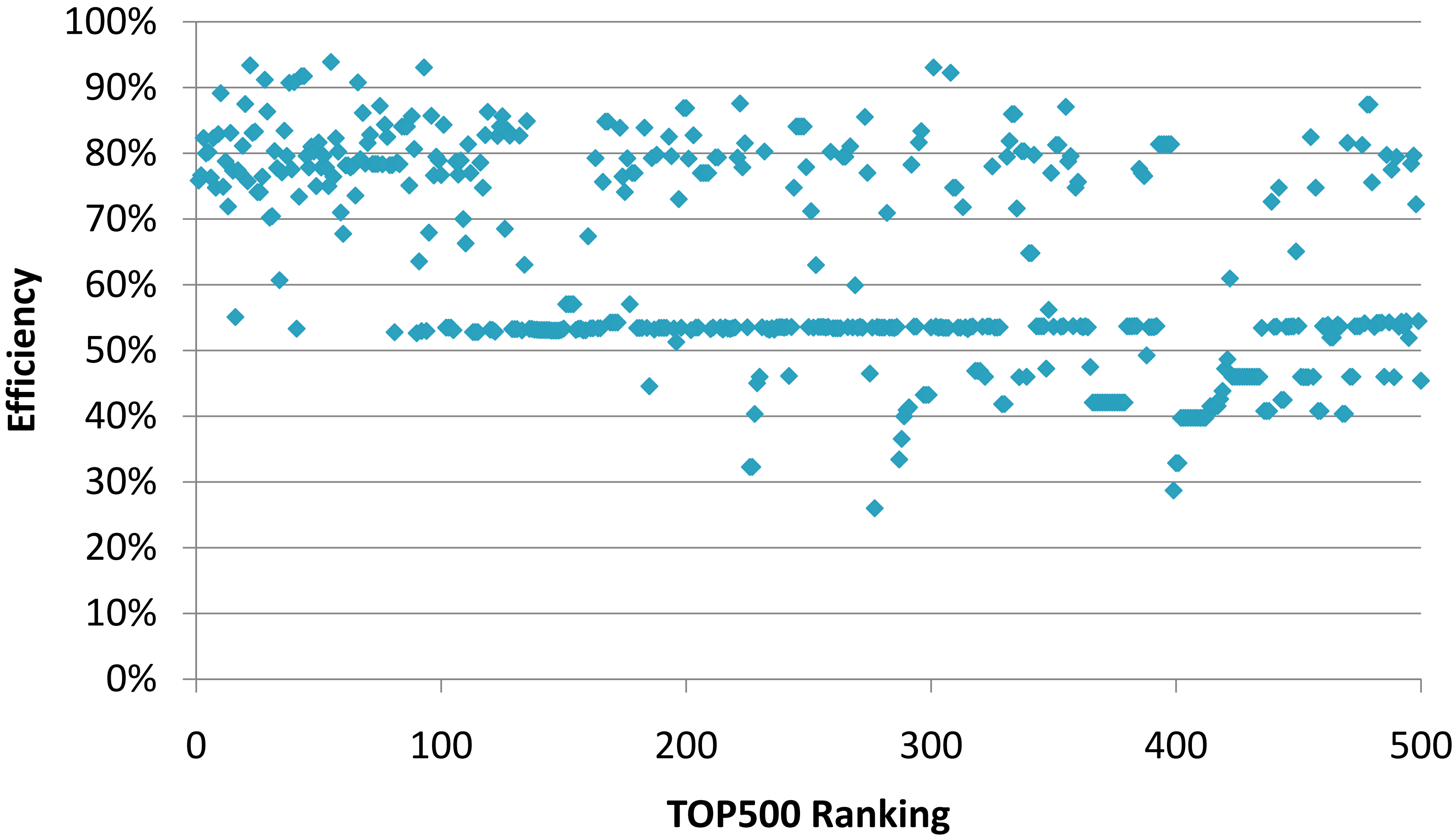
Interconnect Family



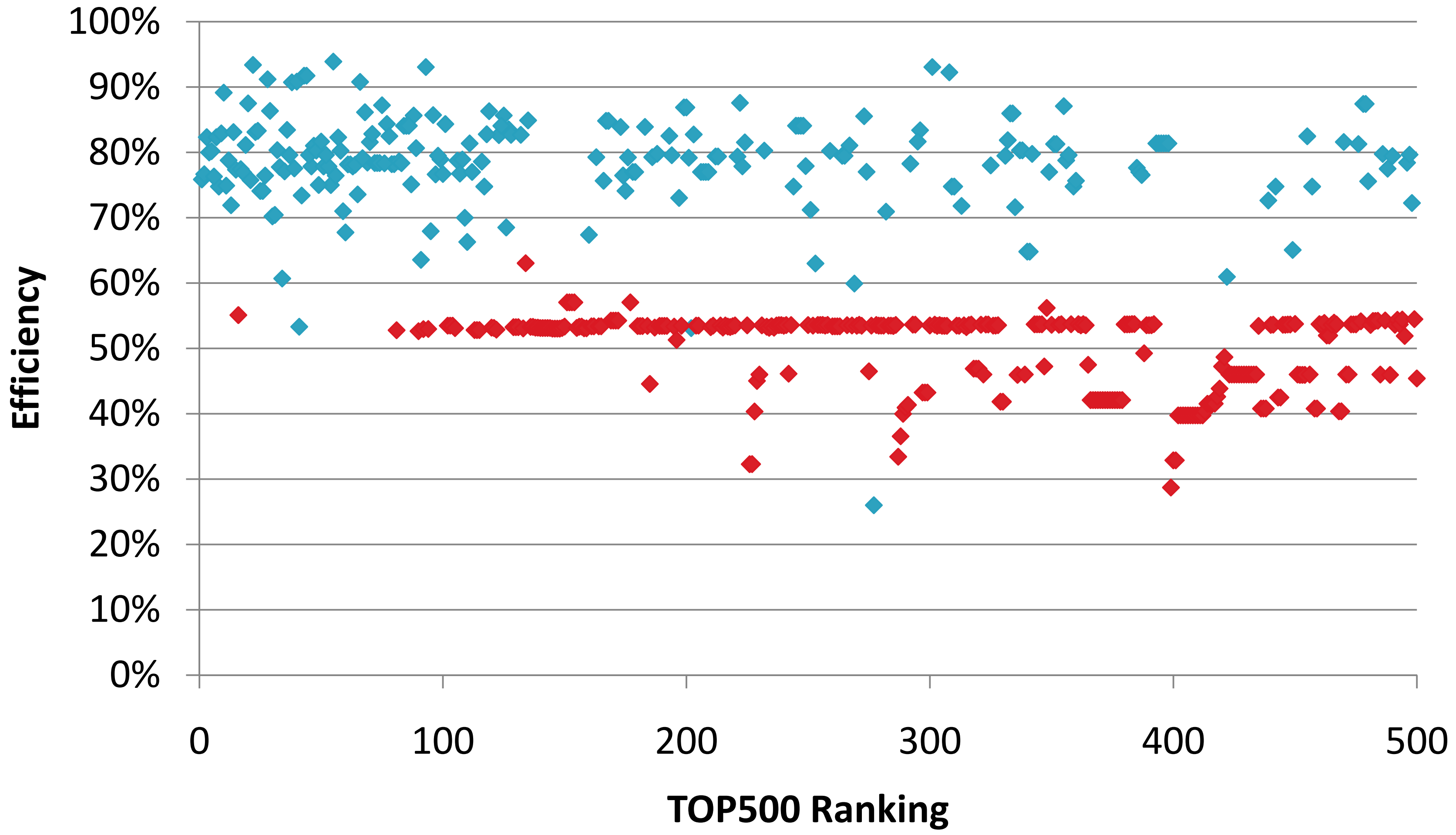
Interconnect Family (TOP50)



Linpack Efficiency



Linpack Efficiency



Bell’s Law (1972)

*“**Bell's Law of Computer Class formation** was discovered about 1972. It states that technology advances in semiconductors, storage, user interface and networking advance every decade enable a new, usually lower priced computing platform to form. Once formed, each class is maintained as a quite independent industry structure. This explains mainframes, minicomputers, workstations and Personal computers, the web, emerging web services, palm and mobile devices, and ubiquitous interconnected networks. We can expect home and body area networks to follow this path.”*

From Gordon Bell (2007), http://research.microsoft.com/~GBell/Pubs.htm

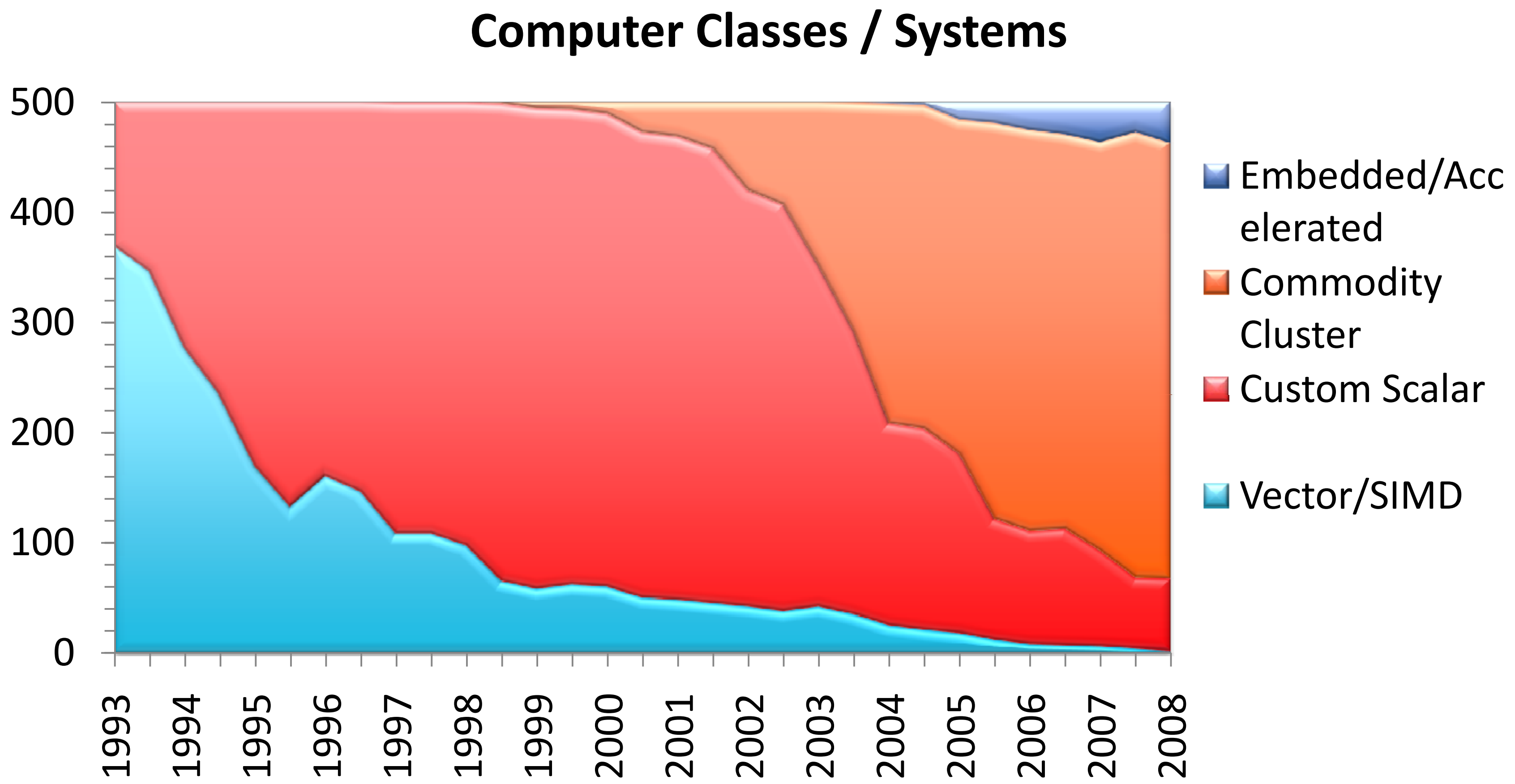
HPC Computer Classes and Bell’s Law

- Bell’s Law states, that:
- Important classes of computer architectures come in cycles of about 10 years.
- It takes about a decade for each phase
  - Early research
  - Early adoption and maturation
  - Prime usage
  - Phase out past its prime
- Can we use Bell’s Law to classify computer architectures in the TOP500?

HPC Computer Classes and Bell’s Law

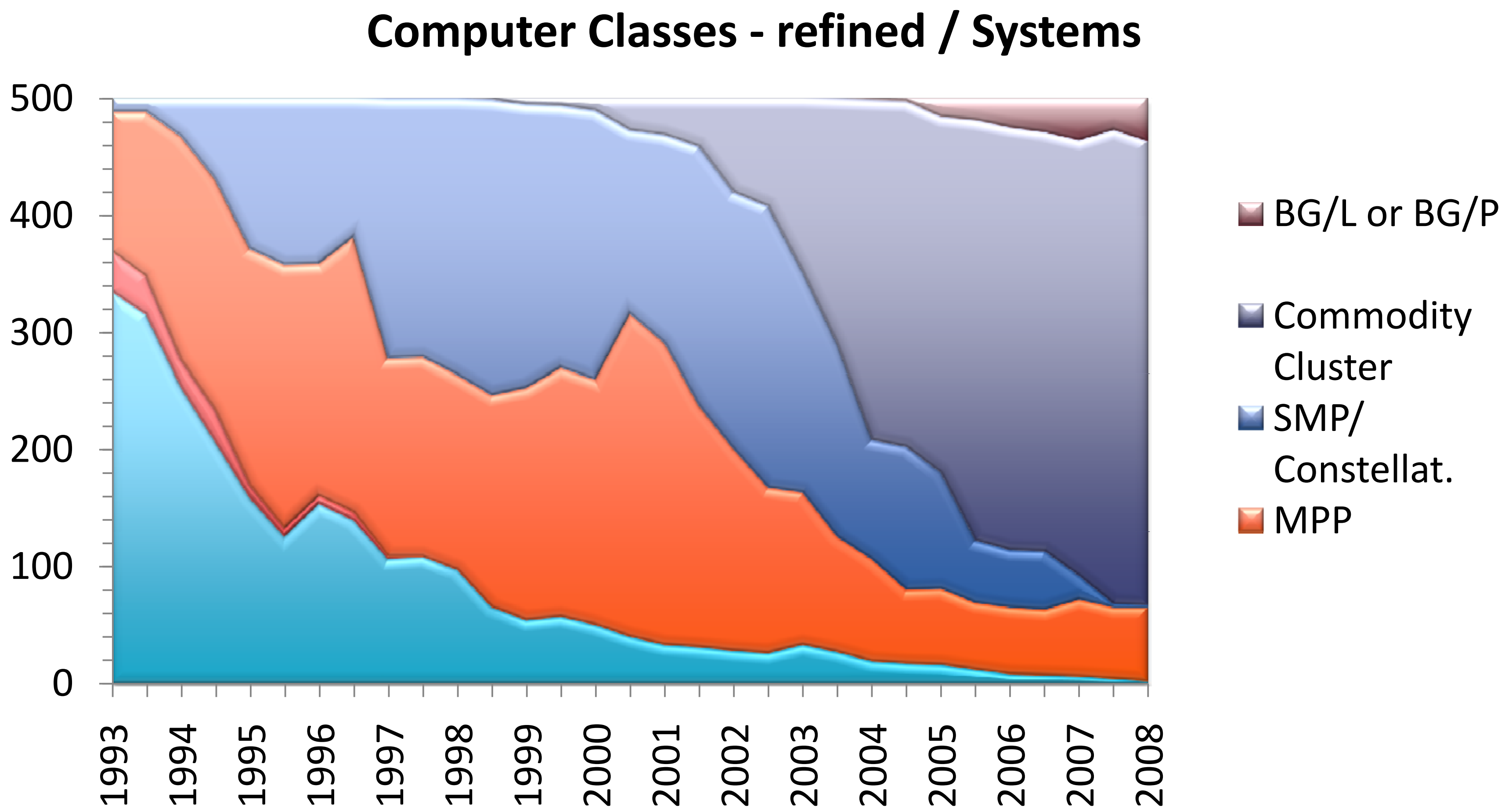
- Gordon Bell (1972): 10 year cycles for computer classes
- Computer classes in HPC based on the TOP500:
- Data Parallel Systems:
  - Vector (Cray Y-MP and X1, NEC SX, …)
  - SIMD (CM-2, …)
- Custom Scalar Systems:
  - MPP (Cray T3E and XT3, IBM SP, …)
  - Scalar SMPs and Constellations (Cluster of big SMPs)
- Commodity Cluster: NOW, PC cluster, Blades, …
- Power-Efficient Systems (BG/L as first example of low-power / embedded systems = potential new class ?)
  - Tsubame with Clearspeed, Roadrunner with Cell, ?
  - This is a fundamental rapture and new paradigm for computing at all levels!

Bell’s Law



Computer Classes / Systems
500
400
300
200
100
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
Embedded/Accelerated
Commodity Cluster
Custom Scalar
Vector/SIMD

Bell's Law


Computer Classes - refined / Systems
500
400
300
200
100
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
BG/L or BG/P
Commodity Cluster
SMP/ Constellat.
MPP

Bell’s Law

HPC Computer Classes

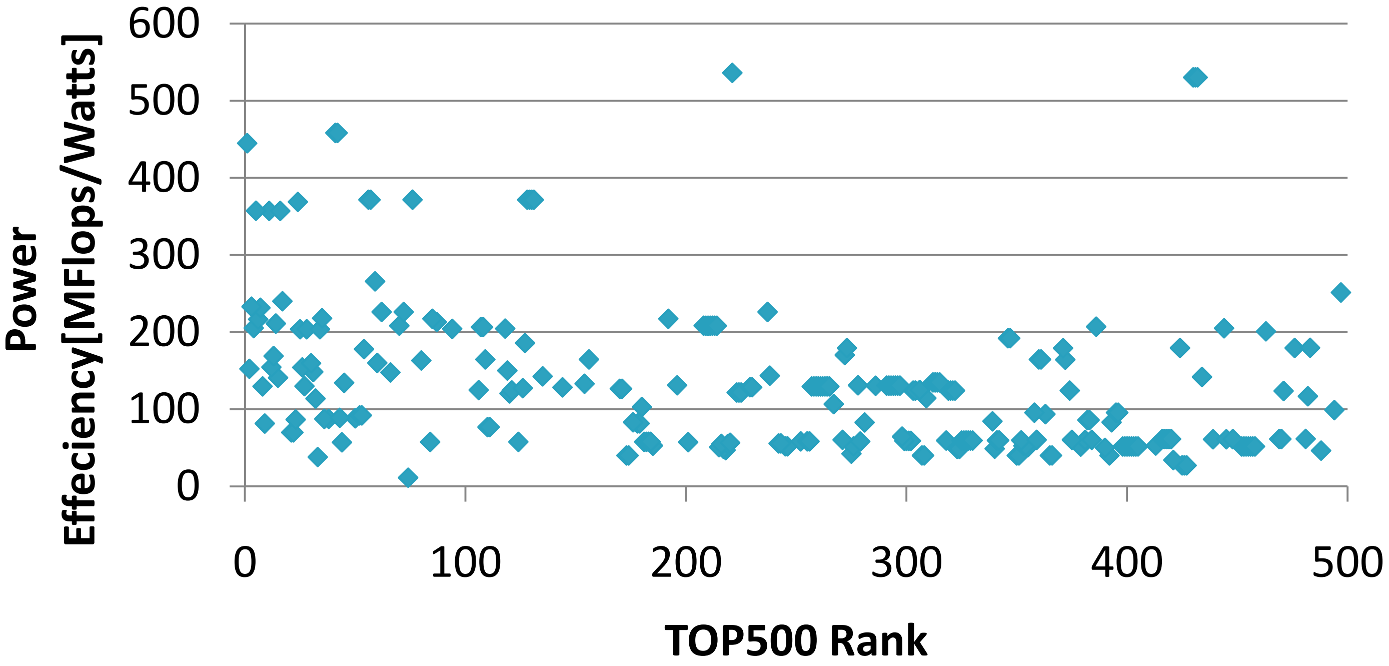
Class	Early Adoption starts:	Prime Use starts:	Past Prime Usage starts:
Data Parallel Systems	Mid 70’s	Mid 80’s	Mid 90’s
Custom Scalar Systems	Mid 80’s	Mid 90’s	Mid 2000’s
Commodity Cluster	Mid 90’s	Mid 2000’s	Mid 2010’s ???
BG/L or BG/P	Mid 2000’s	Mid 2010’s ???	Mid 2020’s ???

Power Consumption Data

- Have been planning this for years.
- Started in June 2008
- Independent from the Green500, but we try to learn from each other.
- Collect power consumption for:
  - Linpack as workload
  - Including all essential parts of a system (processor, memory, …)
  - Excluding features related to machine room (Most disk, UPS, …)
  - Full system or part large enough to include all shared components (fans, power supplies, …)
- Analyze these data carefully!

Power Ranking and How Not to do it!

- To rank ob
  - Weight (
  - Rmax (T
    - A 'lar
- The ratio c
  - (weight/
  - Perform
- One can-n
  - Density
  - A piece of lead is not heavier or larger than one piece of wood.



- Linpack (sub-linear) / Power (linear)
  will always sort smaller systems before larger ones!

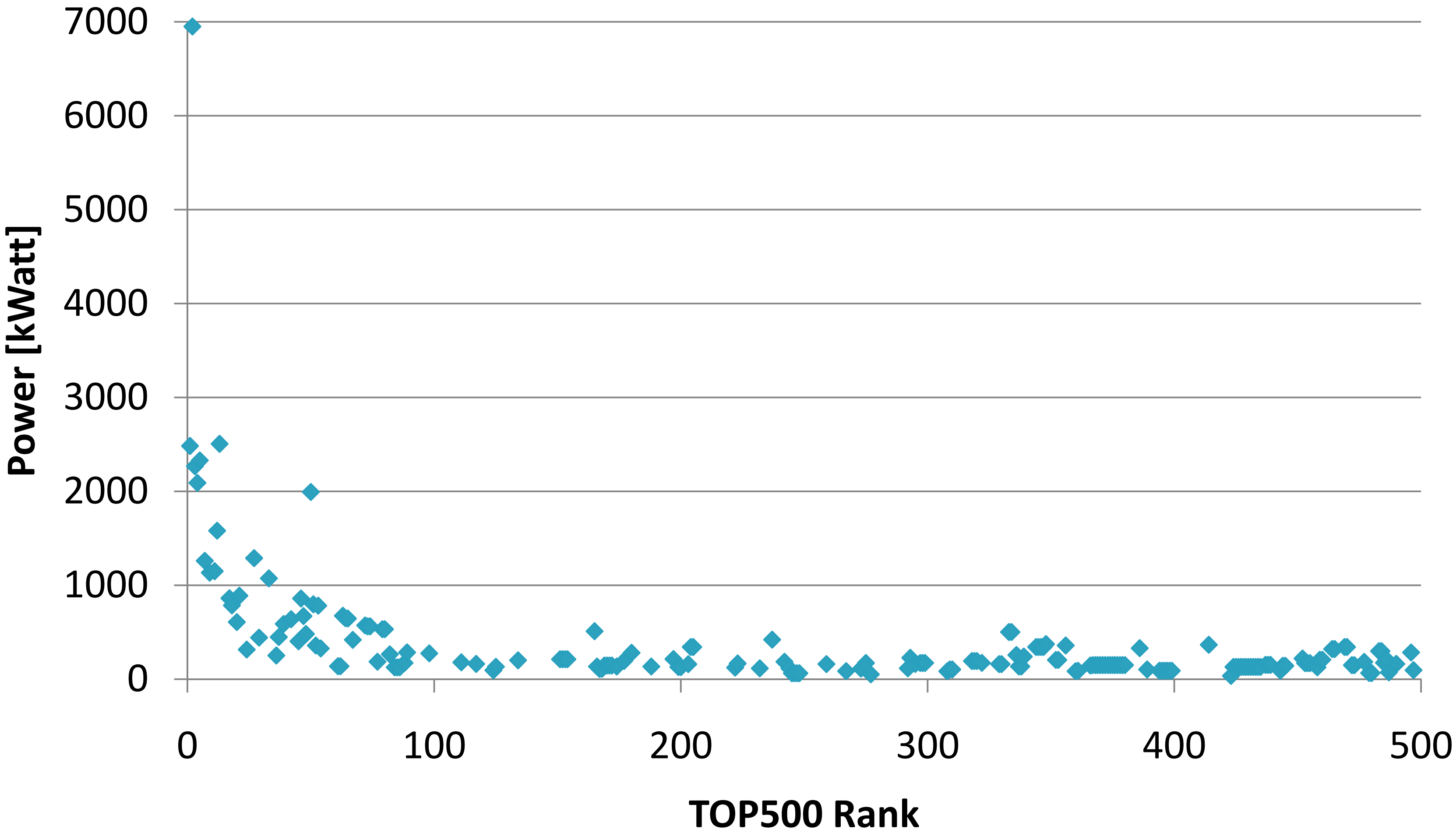
Power Ranking and How Not to do it!

- To rank objects by “size” one needs extensive properties:
  - Weight or Volume
  - Rmax (TOP500)
    - A ‘larger’ system should have a larger Rmax.
- The ratio of 2 extensive properties is an intensive one:
  - (weight/volume = density)
  - Performance / Power Consumption = Power_efficiency
- One can-not ‘rank’ objects with densities BY SIZE:
  - Density does not tell anything about size of an object
  - A piece of lead is not heavier or larger than one piece of wood.
- Linpack (sub-linear) / Power (linear)
  will always sort smaller systems before larger ones!

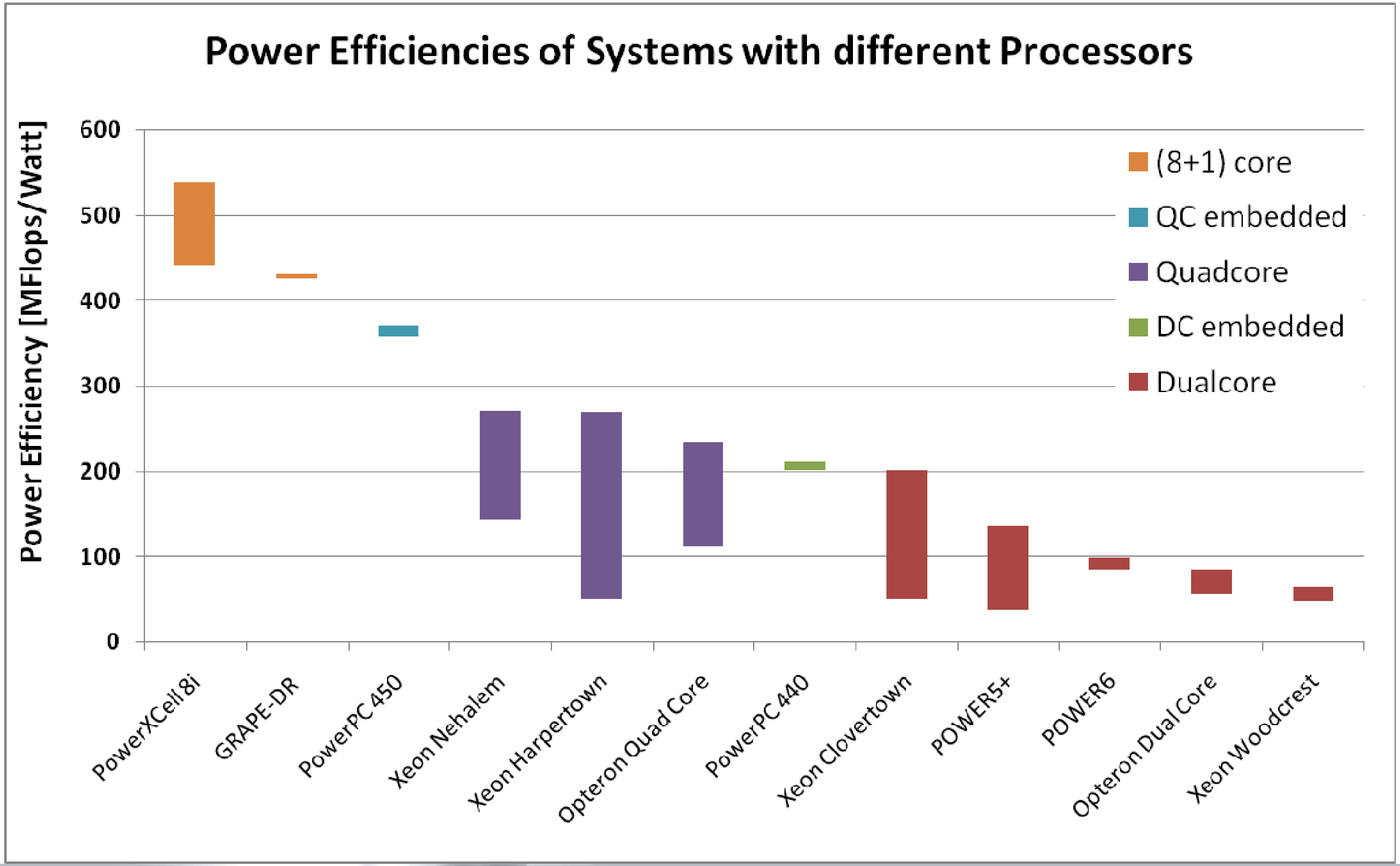
Multi-Core and Many-Core

- Power consumption of chips and systems has increased tremendously, because of ‘cheap’ exploitation of Moore’s Law.
  - Free lunch has ended
  - Stall of frequencies forces increasing concurrency levels, Multi-Cores
  - Optimal core sizes/power are smaller than current ‘rich cores’, which leads to Many-Cores
- Many-Cores, more (10-100x) but smaller cores:
  - Intel Polaris – 80 cores,
  - Clearspeed CSX600 – 96 cores,
  - nVidia G80 – 128 cores, or
  - CISCO Metro – 188 cores

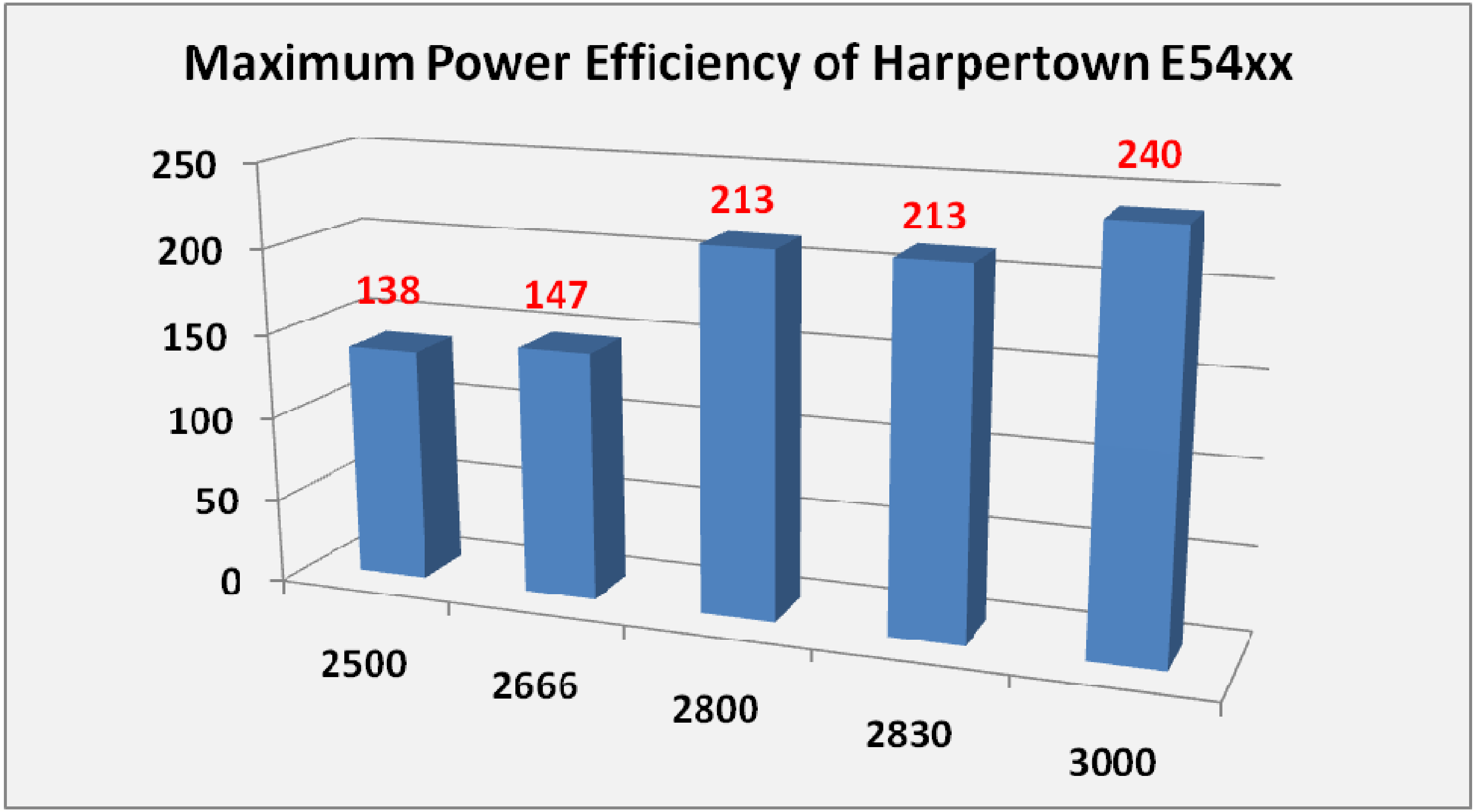
Absolute Power Levels



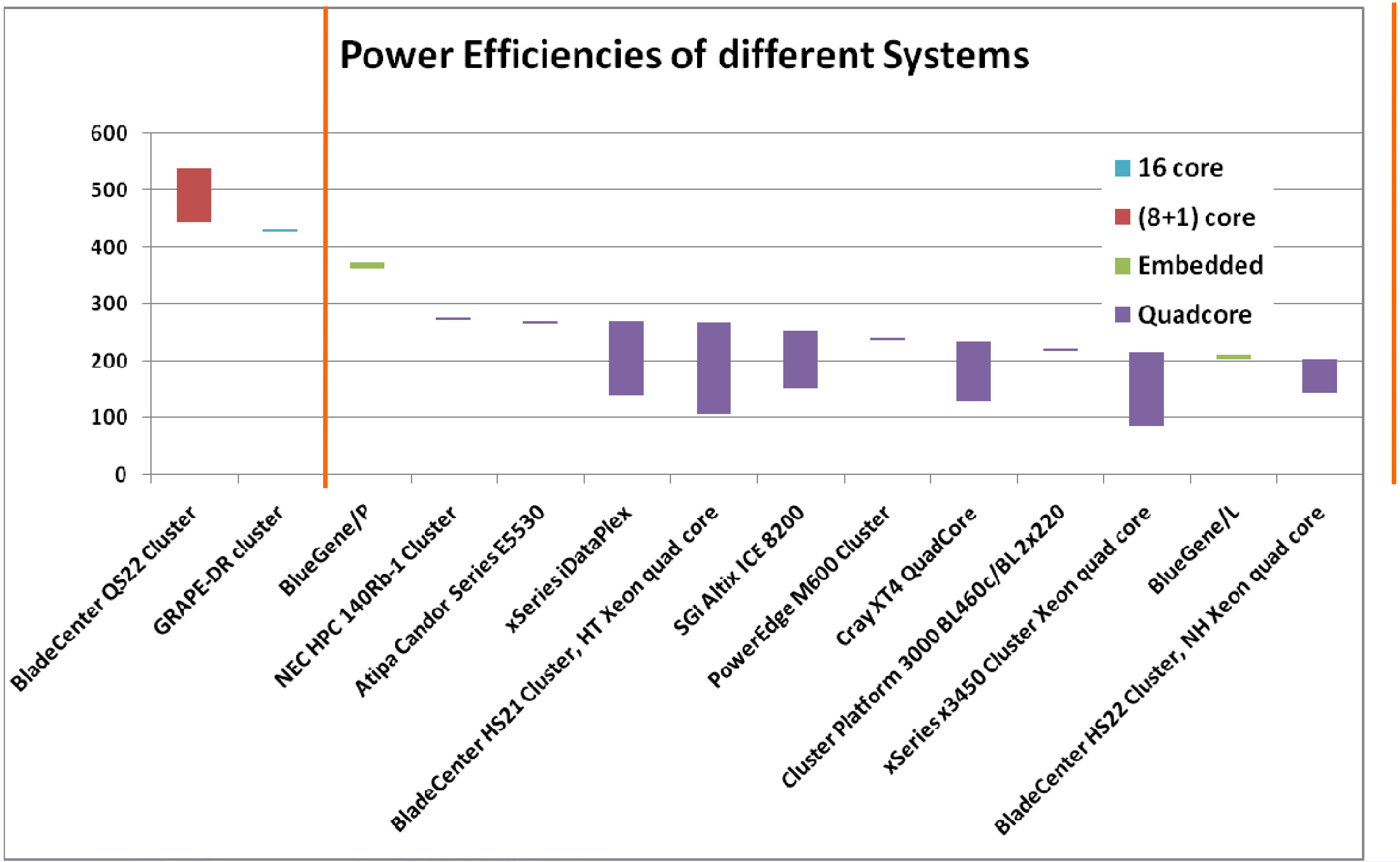
Power Efficiency related to Processors



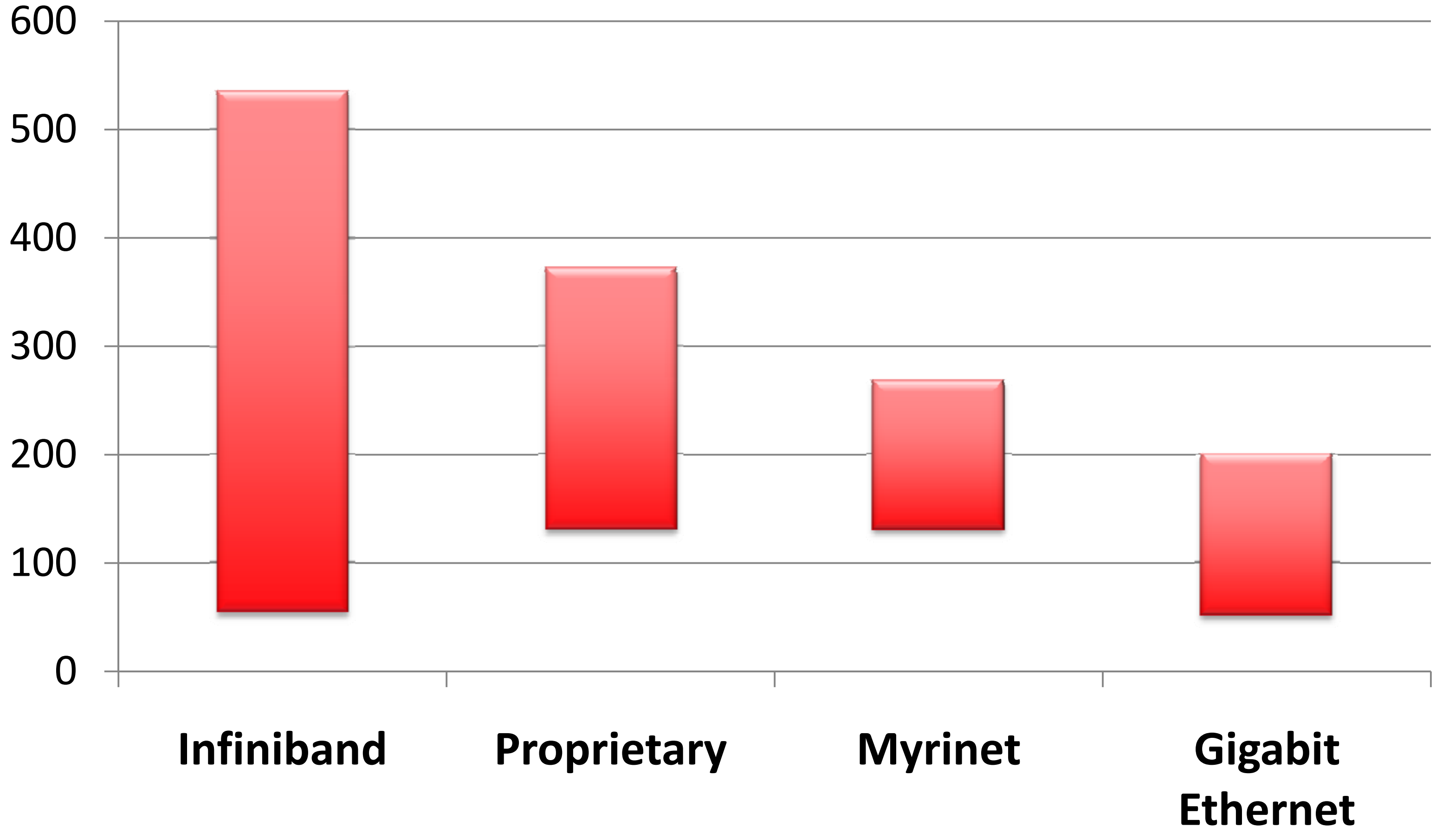
Frequencies and Power Efficiency



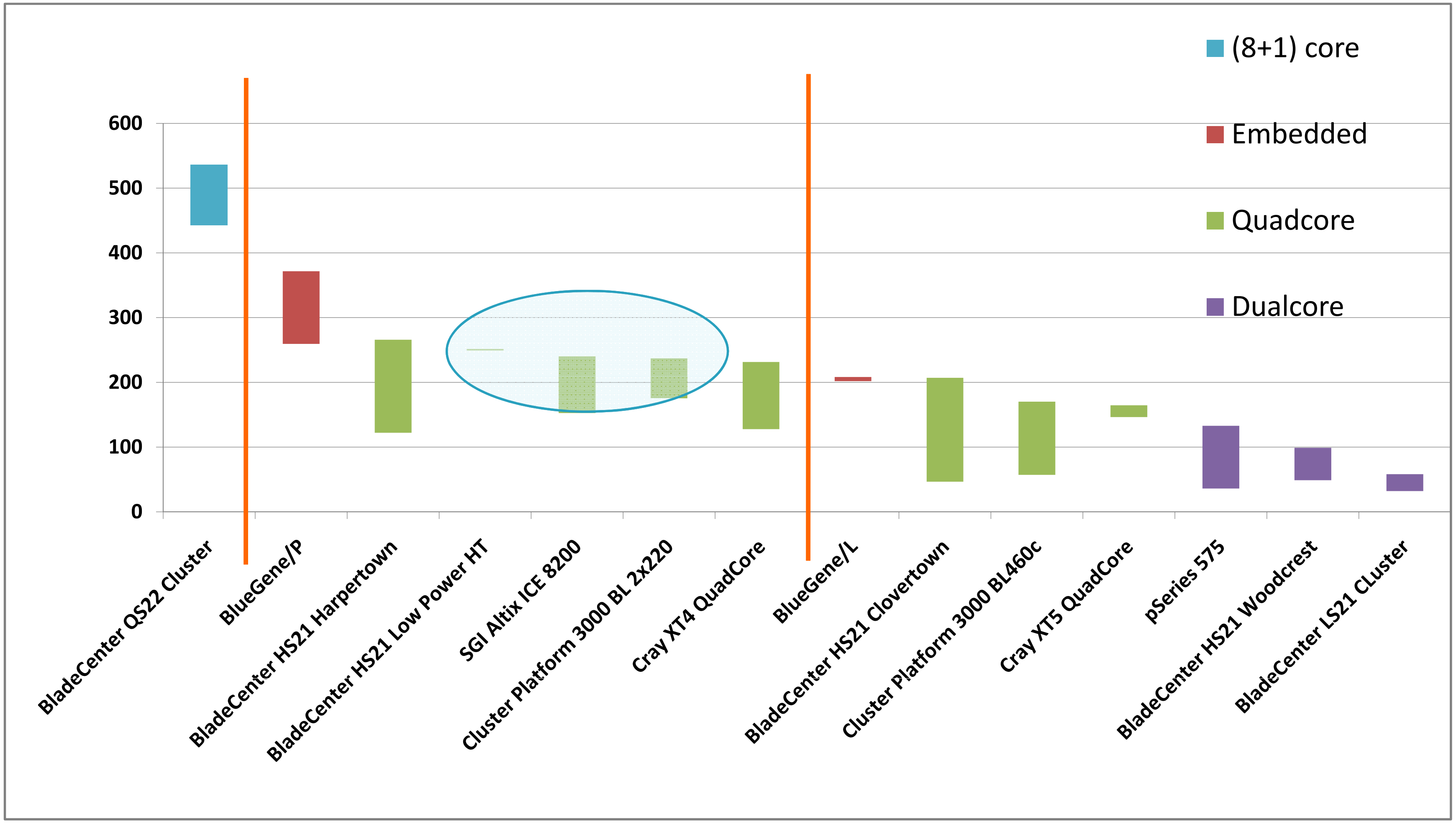
Power Efficiencies of different Systems



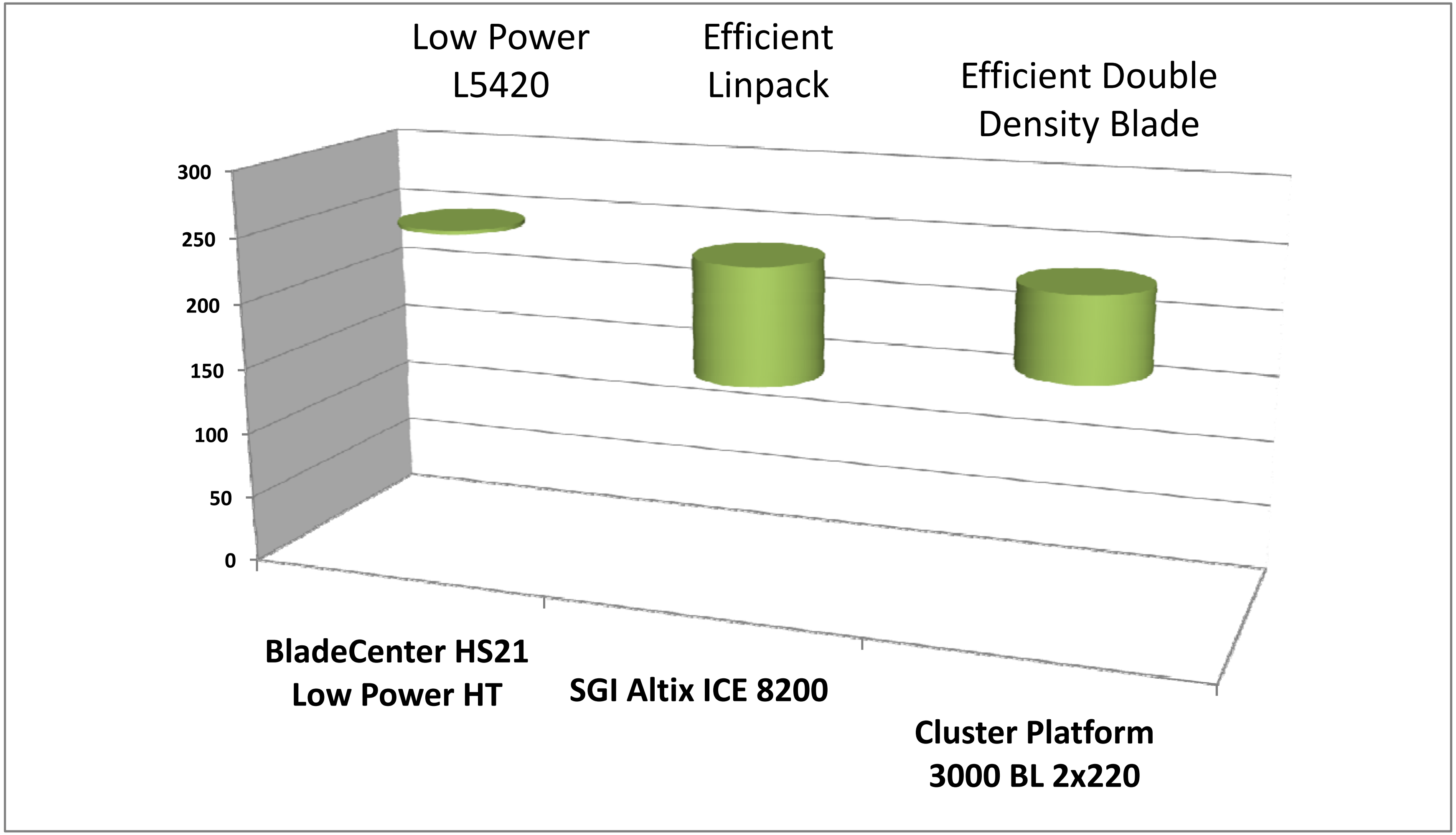
Power Efficiency related to Interconnects



Power Efficiencies of different Systems



Power Efficiencies of different Systems


Low Power
L5420
Efficient
Linpack
Efficient Double
Density Blade
BladeCenter HS21
Low Power HT
SGI Altix ICE 8200
Cluster Platform
3000 BL 2x220