Chapter 1

Introduction

Why High Performance Computing?

Quote:

It is hard to understand an ocean because it is too big.

It is hard to understand a molecule because it is too small.

It is hard to understand nuclear physics because it is too fast.

It is hard to understand the greenhouse effect because it is too slow.

Supercomputers break these barriers to understanding.

They, in effect, shrink oceans, zoom in on molecules, slow down physics, and fastforward climates. Clearly a scientist who can see natural phenomena at the right size and the right speed learns more than one who is faced with a blur.

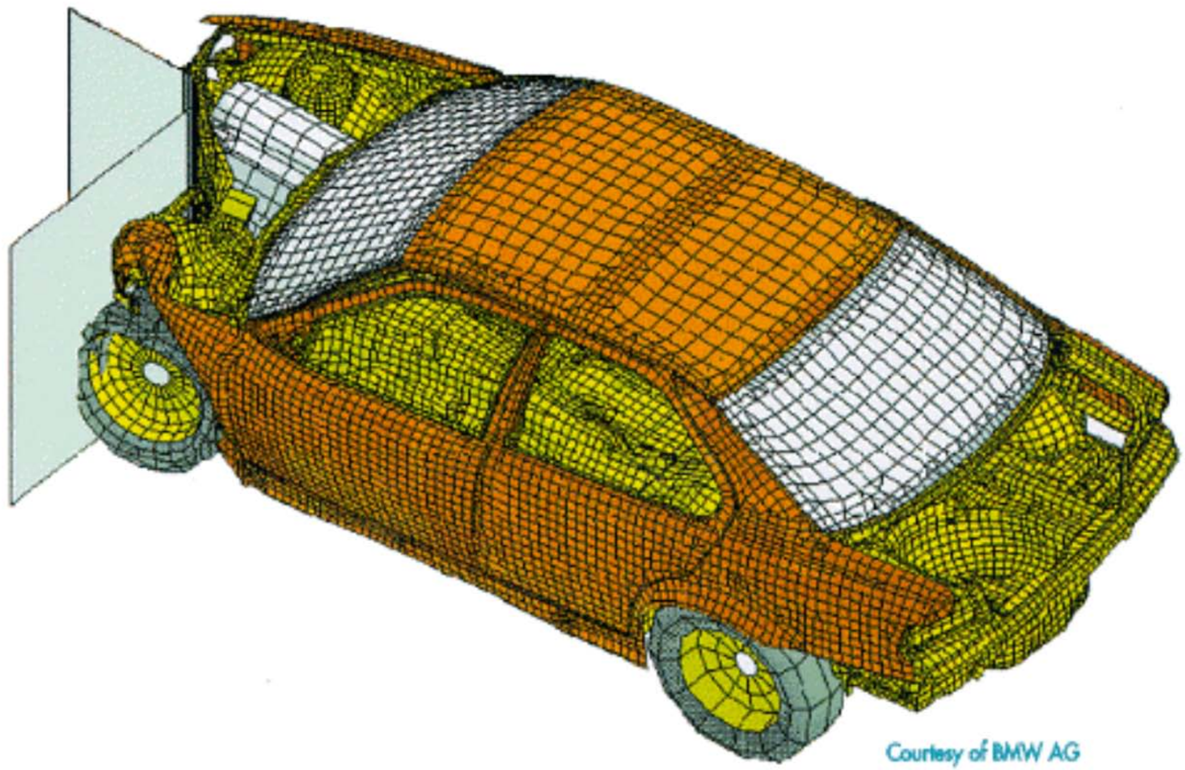
Al Gore, 1990

Why High Performance Computing?

- Grand Challenges (Basic Research)
  - Decoding of human genome
  - Kosmogenesis
  - Global Climate Changes
  - Biological Macro molecules
- Product development
  - Fluid mechanics
  - Crash tests
  - Material minimization
  - Drug design
  - Chip design
- IT Infrastructure
  - Search engines
  - Data Mining
- Virtual Reality
  - Rendering
  - Vision

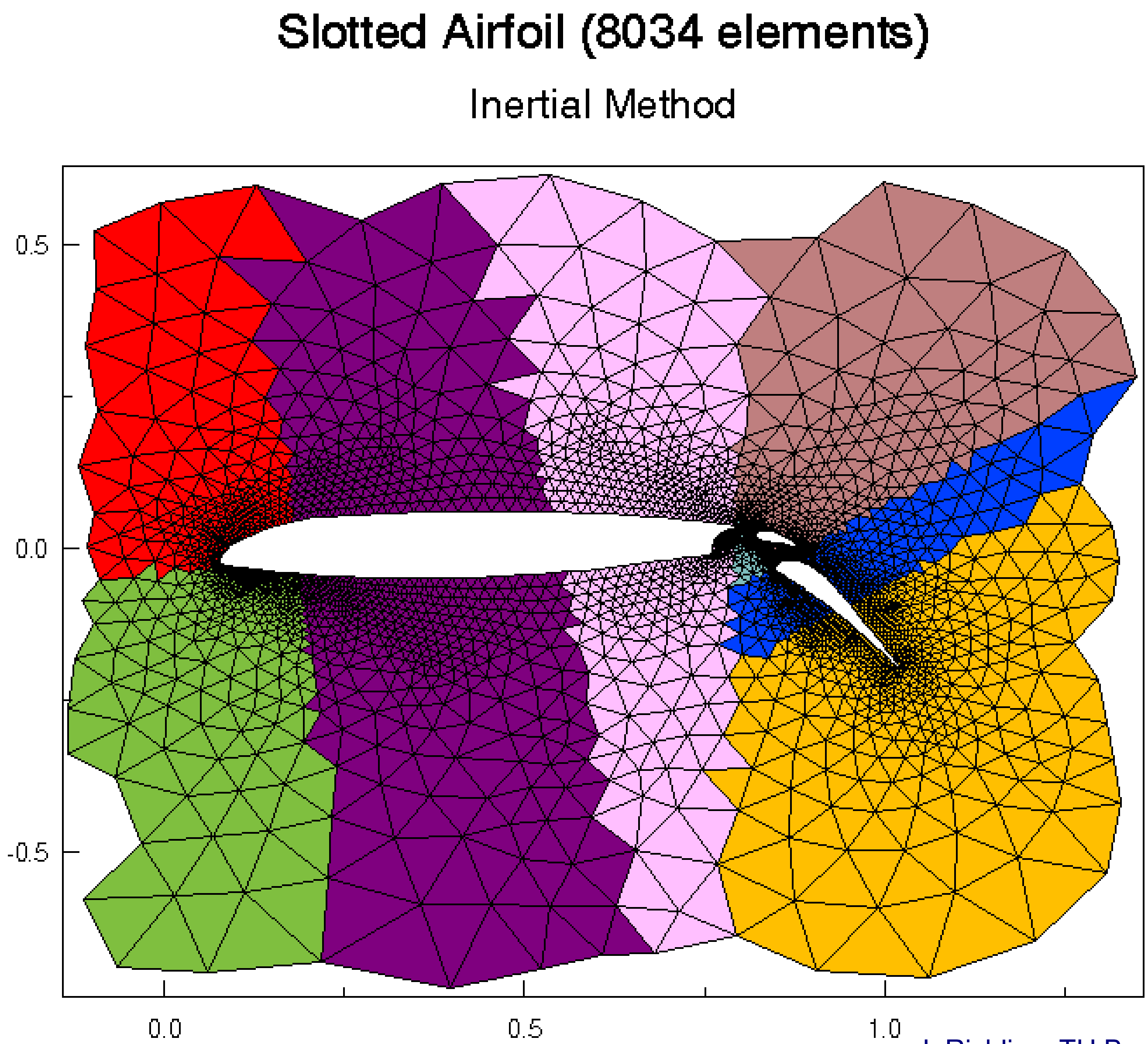
Examples for HPC-Applications

Finite Element Method: Crash-Analysis



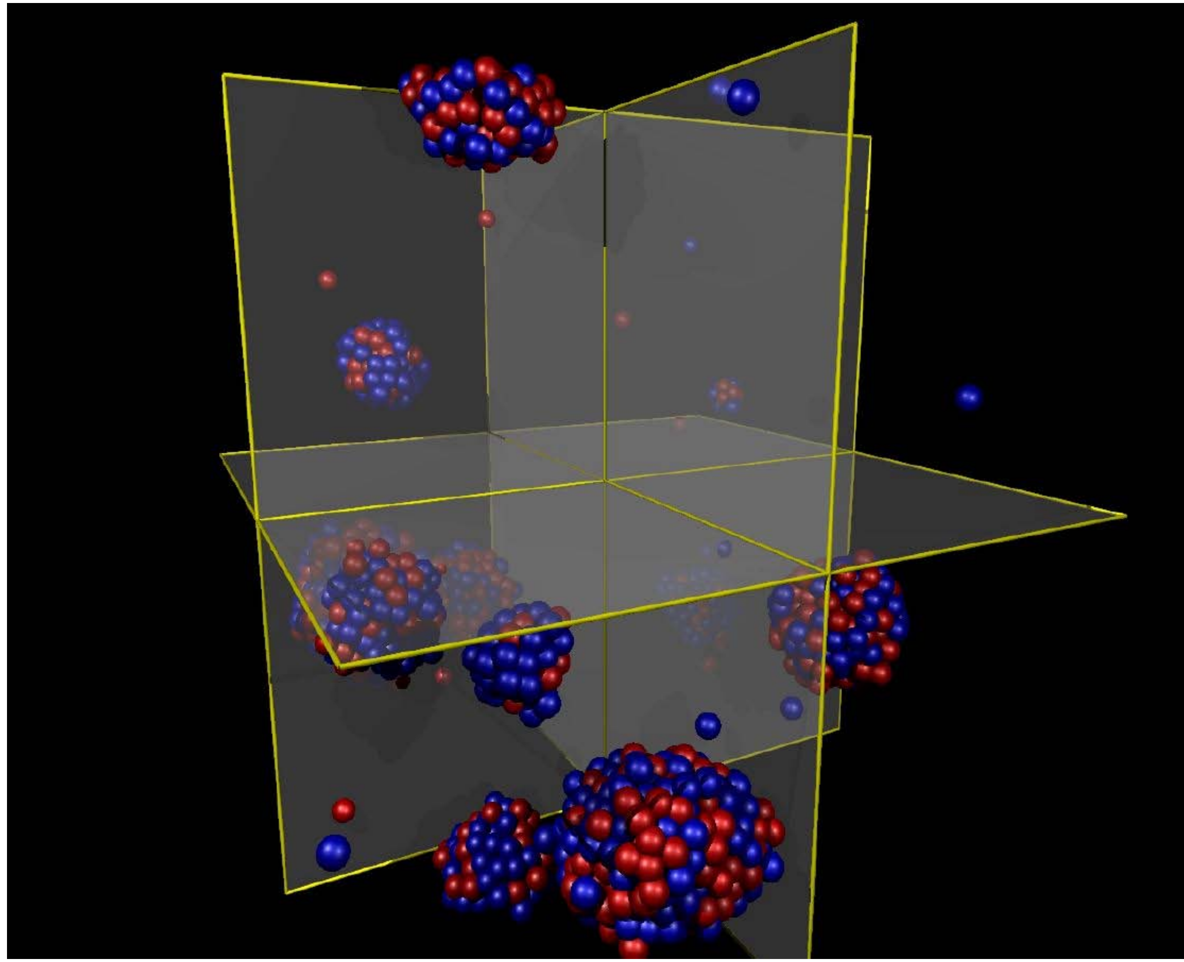
Asymmetric frontal impact of a car to a rigid obstacle

Examples for HPC-Applications: Aerodynamics



Examples for HPC-Applications: Molecular dynamics

Simulation of a noble gas (Argon) with a partitioning for 8 processors: 2048 molecules in a cube of16 nm edge length simulated in time steps of $2\text{x}10^{-14}$ sec.



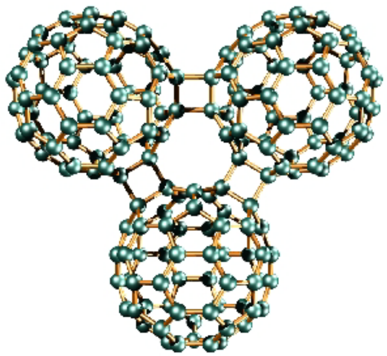
Examples for HPC-Applications:

Rendering (Interior architecture)

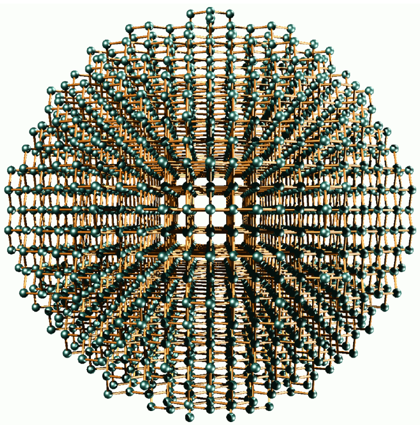
Rendering (Movie: Titanic)

Photo realistic presentation using sophisticated illumination modules

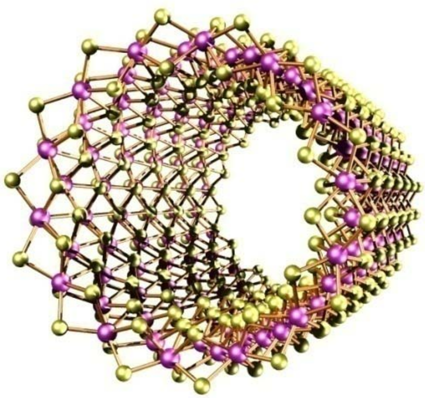
Examples for HPC-Applications:
Analysis of complex materials



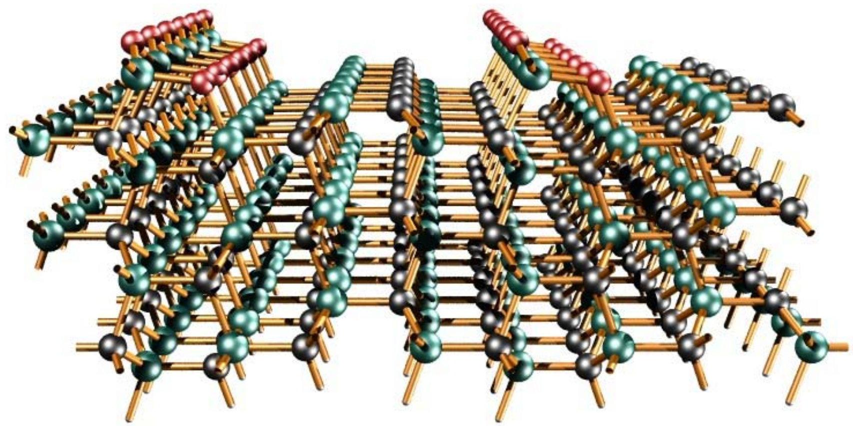
C_{60}-trimer



Si_{1600}



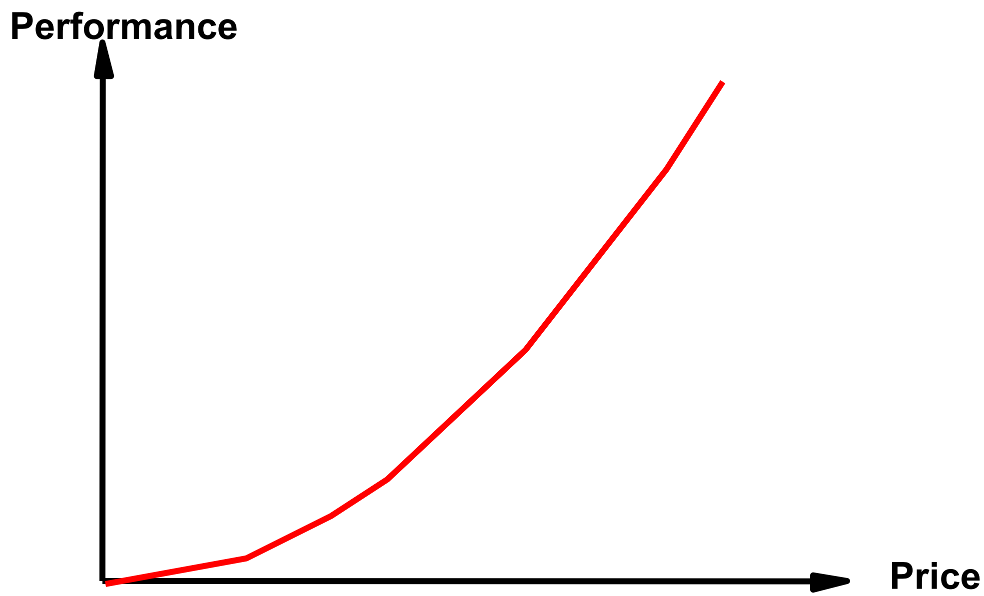
MoS_2



4H-SiC

Source: Th. Frauenheim, Paderborn

Grosch's Law



„The performance of a computer increases (roughly) quadratically with the prize"

Consequence:

It is better to buy a computer which is twice as fast than two buy two computers.

(The law was valid in the sixties and seventies over a wide range of universal computers.)

Eighties

Availability of powerful Microprocessors

- High Integration density (VLSI)
- Single-Chip-Processors
- Automated production process
- Automated development process
- High volume production

Consequence:
Grosch's Law no longer valid:
1000 cheap microprocessors render (theoretically)
more performance (MLOPS) than expensive Super computer (e.g. Cray)

Idea:
To achieve high performance at low cost use many microprocessors together
➔Parallel Processing

Eighties

- Wide spread use of workstations and PCs

  Terminals being replaced by PCs

  Workstations achieve performance of mainframes at a fraction of the price.

- Availability of local area networks (LAN) (Ethernet)

  Possibility to connect a larger number of autonomous computers using a low cost medium. Access to data of other computers. Usage of programs and other resources of remote computers.

- Network of Workstations as Parallel Computer

  Possibility to exploit unused compute capacity of other computers for compute-intensive calculations (idle time computing).

Nineties

- Parallel computers are built of a large number of microprocessors (Massively Parallel Processors, MPP), e. g. Transputer systems, Connection Machine CM-5, Intel Paragon, Cray T3E
- Alternative Architectures are built (Connection Machine CM-2, MasPar).
- Trend to use cost efficient standard components (“off-the-shelf”) leads to coupling of standard PCs to a “Cluster” (Beowulf, 1995)
- Exploitation of unused compute power for HPC (“idle time computing”)
- Program libraries like PVM and MPI allow for development of portable parallel programs
- Linux as operating system for HPC becomes prevailing

Top 10 of TOP500 List (06/2001)

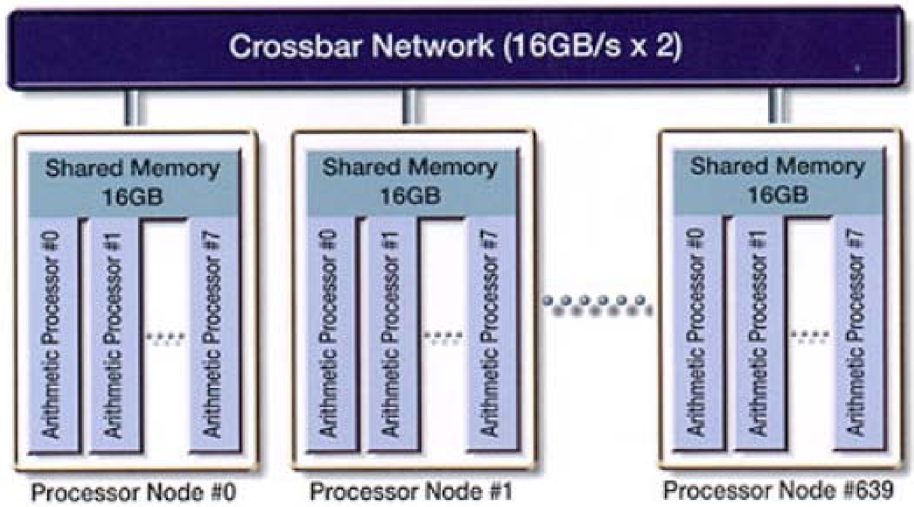
RANK	MANU-FACTURER	COMPUTER	R_{MAX} [TF/S]	INSTALLATION SITE	COUNTRY	YEAR	AREA OF INSTALLATION	# PROC
1	IBM	ASCI White SP Power3 375 MHz	7.23	Lawrence Livermore National Laboratory	USA	2000	Research Energy	8192
2	IBM	SP Power3 375 MHz 16 way	2.53	NERSC/LBNL, Berkeley	USA	2001	Research	2528
3	Intel	ASCI Red	2.38	Sandia National Laboratory, Albuquerque	USA	1999	Research	9632
4	IBM	ASCI Blue Pacific SST, IBM SP 604E	2.14	Lawrence Livermore National Laboratory	USA	1999	Research Energy	5808
5	Hitachi	SR8000/MPP	1.71	University of Tokyo	Japan	2001	Academic	1152
6	SGI	ASCI Blue Mountain	1.61	Los Alamos National Laboratory	USA	1998	Research	6144
7	IBM	SP Power3 375 MHz	1.42	Naval Oceanographic Office, Bay St. Louis	USA	2000	Research Aerospace	1336
8	NEC	SX-5/128 M3 3.2 ns	1.19	Osaka University	Japan	2001	Academic	128
9	IBM	SP Power3 375 MHz	1.18	National Centers for Environmental Predicition	USA	2000	Research Weather	1104
10	IBM	SP Power3 375 MHz	1.18	National Centers for Environmental Predicition	USA	2001	Research Weather	1104

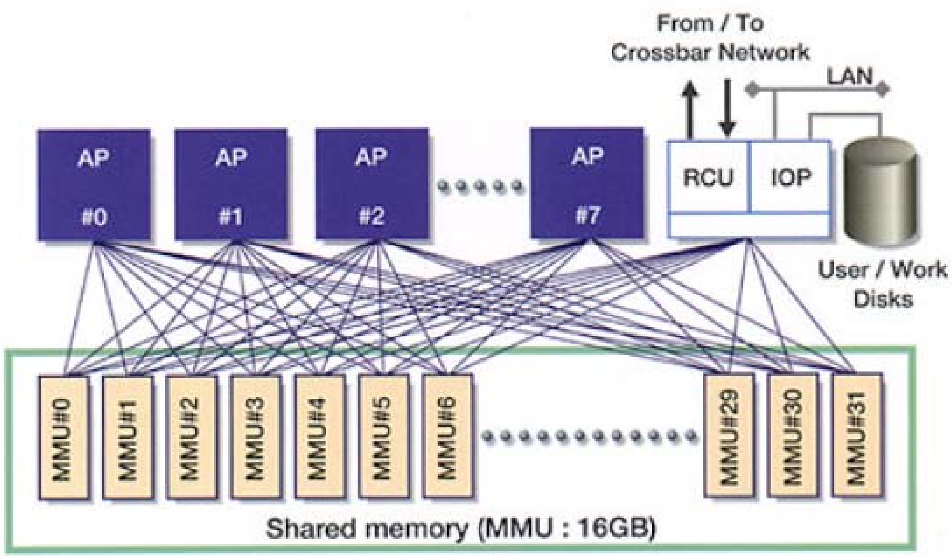
Top 10 of TOP500 List (06/2002)

Rank	Manuf.	Computer	R_{max}(GF)	Installation Site	Country	Year	# Proc	Peak GF
1	NEC	Earth-Simulator	35860.00	Earth Simulator Center	Japan	2002	5120	40960.00
2	IBM	ASCI White, SP Power3 375 MHz	7226.00	Lawrence Livermore National Laboratory	USA	2000	8192	12288.00
3	Hewlett-Packard	AlphaServer SC ES45/1 GHz	4463.00	Pittsburgh Supercomputing Center	USA	2001	3016	6032.00
4	Hewlett-Packard	AlphaServer SC ES45/1 GHz	3980.00	Commissariat a l'Energie Atomique (CEA)	France	2001	2560	5120.00
5	IBM	SP Power3 375 MHz 16 way	3052.00	NERSC/LBNL	USA	2001	3328	4992.00
6	Hewlett-Packard	AlphaServer SC ES45/1 GHz	2916.00	Los Alamos National Laboratory	USA	2002	2048	4096.00
7	Intel	ASCI Red	2379.00	Sandia National Laboratories	USA	1999	9632	3207.00
8	IBM	pSeries 690 Turbo 1.3GHz	2310.00	Oak Ridge National Laboratory	USA	2002	864	4493.00
9	IBM	ASCI Blue-Pacific SST, IBM SP 604e	2144.00	Lawrence Livermore National Laboratory	USA	1999	5808	3868.00
10	IBM	pSeries 690 Turbo 1.3GHz	2002.00	IBM/US Army Research Laboratory (ARL)	USA	2002	768	3994.00

Earth Simulator (Rank 1, 2002-2004)

Earth Simulator





Source: H.Simon, NERSC

- 640 Nodes
- 8 vector processors (each 8 GFLOPS) and 16 GB per node
- 5120 CPUs
- 10 TB main memory
- 40 TFLOPS peak performance
- 65m x 50m physical dimension

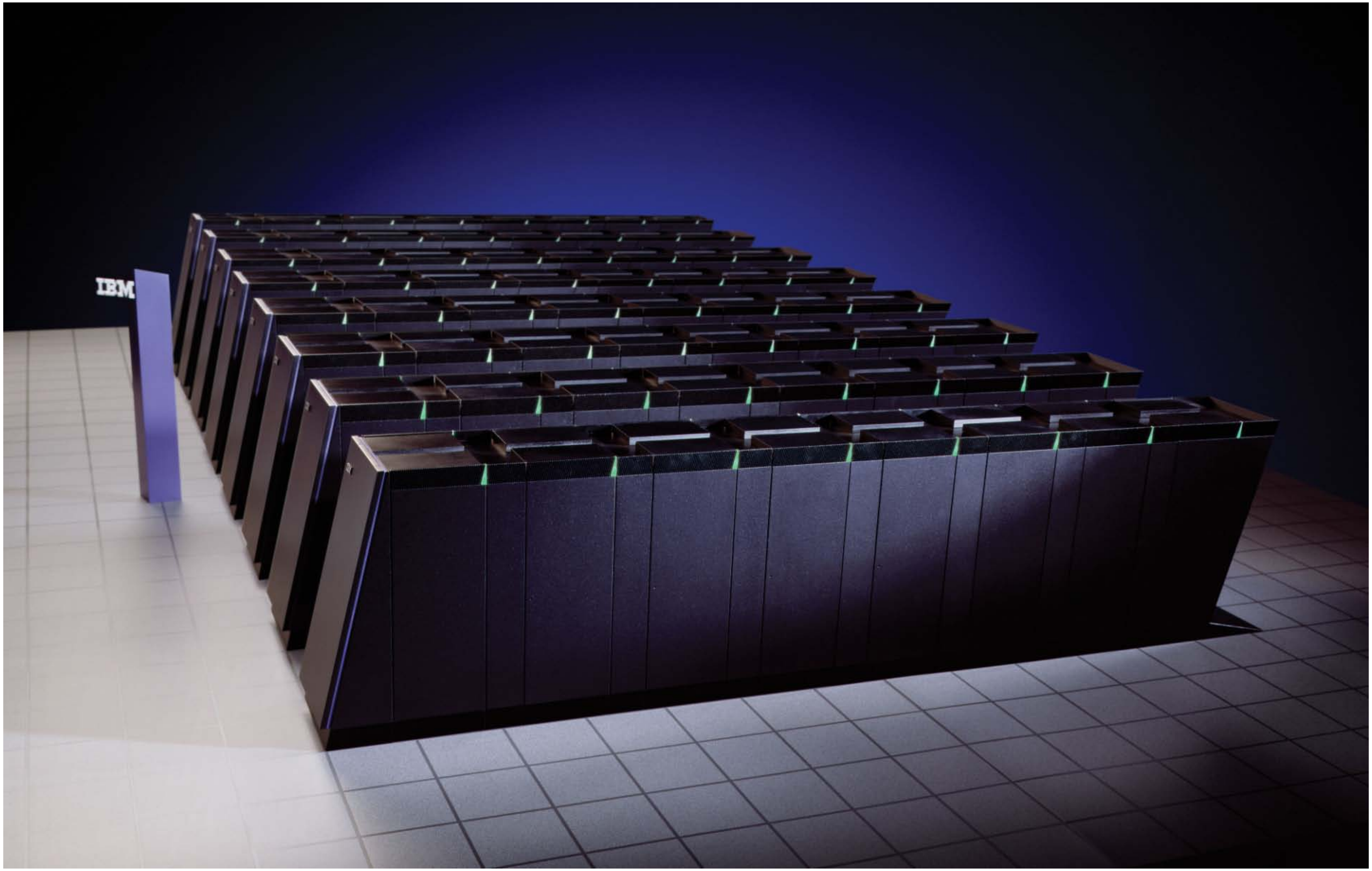
TOP 10 / Nov. 2004

Rank	Site Country/Year	Computer / Processors Manufacturer	R_{max} R_{peak}
1	IBM/DOE United States/2004	*BlueGene/L beta-System* **BlueGene/L DD2 beta-System (0.7 GHz PowerPC 440)** / 32768 IBM	70720 91750
2	NASA/Ames Research Center/NAS United States/2004	*Columbia* **SGI Altix 1.5 GHz, Voltaire Infiniband** / 10160 SGI	51870 60960
3	The Earth Simulator Center Japan/2002	**Earth-Simulator** / 5120 NEC	35860 40960
4	Barcelona Supercomputer Center Spain/2004	*MareNostrum* **eServer BladeCenter JS20 (PowerPC970 2.2 GHz), Myrinet** / 3564 IBM	20530 31363
5	Lawrence Livermore National Laboratory United States/2004	*Thunder* **Intel Itanium2 Tiger4 1.4GHz - Quadrics** / 4096 California Digital Corporation	19940 22938
6	Los Alamos National Laboratory United States/2002	*ASCI Q* **ASCI Q - AlphaServer SC45, 1.25 GHz** / 8192 HP	13880 20480
7	Virginia Tech United States/2004	*System X* **1100 Dual 2.3 GHz Apple XServe/Mellanox Infiniband 4X/Cisco GigE** / 2200 Self-made	12250 20240
8	IBM - Rochester United States/2004	**BlueGene/L DD1 Prototype (0.5GHz PowerPC 440 w/Custom)** / 8192 IBM/ LLNL	11680 16384
9	Naval Oceanographic Office United States/2004	**eServer pSeries 655 (1.7 GHz Power4+)** / 2944 IBM	10310 20019.2
10	NCSA United States/2003	*Tungsten* **PowerEdge 1750, P4 Xeon 3.06 GHz, Myrinet** / 2500 Dell	9819 15300

TOP 10 / Nov. 2005

Rank	Site	Computer	Processors	Year	R_{max}
1	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution IBM	131072	2005	280600
2	IBM Thomas J. Watson Research Center United States	BGW - eServer Blue Gene Solution IBM	40960	2005	91290
3	DOE/NNSA/LLNL United States	ASC Purple - eServer pSeries p5 575 1.9 GHz IBM	10240	2005	63390
4	NASA/Ames Research Center/NAS United States	Columbia - SGI Altix 1.5 GHz, Voltaire Infiniband SGI	10160	2004	51870
5	Sandia National Laboratories United States	Thunderbird - PowerEdge 1850, 3.6 GHz, Infiniband Dell	8000	2005	38270
6	Sandia National Laboratories United States	Red Storm Cray XT3, 2.0 GHz Cray Inc.	10880	2005	36190
7	The Earth Simulator Center Japan	Earth-Simulator NEC	5120	2002	35860
8	Barcelona Supercomputer Center Spain	MareNostrum - JS20 Cluster, PPC 970, 2.2 GHz, Myrinet IBM	4800	2005	27910
9	ASTRON/University Groningen Netherlands	Stella - eServer Blue Gene Solution IBM	12288	2005	27450
10	Oak Ridge National Laboratory United States	Jaguar - Cray XT3, 2.4 GHz Cray Inc.	5200	2005	20527

IBM Blue Gene



IBM Blue Gene

TOP 10 / Nov. 2006

Rank	Site	Manufacturer	Computer	Country	Processors	RMax
1	DOE/NNSA/LLNL	IBM	eServer Blue Gene Solution	United States	131072	280600
2	NNSA/Sandia National Laboratories	Cray Inc.	Sandia/ Cray Red Storm, Opteron 2.4 GHz dual core	United States	26544	101400
3	IBM Thomas J. Watson Research Center	IBM	eServer Blue Gene Solution	United States	40960	91290
4	DOE/NNSA/LLNL	IBM	eServer pSeries p5 575 1.9 GHz	United States	12208	75760
5	Barcelona Supercomputing Center	IBM	BladeCenter JS21 Cluster, PPC 970, 2.3 GHz, Myrinet	Spain	10240	62630
6	NNSA/Sandia National Laboratories	Dell	PowerEdge 1850, 3.6 GHz, Infiniband	United States	9024	53000
7	Commissariat a l'Energie Atomique (CEA)	Bull SA	NovaScale 5160, Itanium2 1.6 GHz, Quadrics	France	9968	52840
8	NASA/Ames Research Center/NAS	SGI	SGI Altix 1.5 GHz, Voltaire Infiniband	United States	10160	51870
9	GSIC Center, Tokyo Institute of Technology	NEC/Sun	Sun Fire x4600 Cluster, Opteron 2.4/2.6 GHz, Infiniband	Japan	11088	47380
10	Oak Ridge National Laboratory	Cray Inc.	Cray XT3, 2.6 GHz dual Core	United States	10424	43480

Top 500 Nov 2008

Rank	Computer	Country	Year	Cores	RMax	Processor Family	System Family	Inter-connect
1	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz , Voltaire Infiniband	United States	2008	129600	1105000	Power	IBM Cluster	Infiniband
2	Cray XT5 QC 2.3 GHz	United States	2008	150152	1059000	AMD x86_64	Cray XT	Proprietary
3	SGI Altix ICE 8200EX, Xeon QC 3.0/2.66 GHz	United States	2008	51200	487005	Intel EM64T	SGI Altix	Infiniband
4	eServer Blue Gene Solution	United States	2007	212992	478200	Power	IBM BlueGene	Proprietary
5	Blue Gene/P Solution	United States	2007	163840	450300	Power	IBM BlueGene	Proprietary
6	SunBlade x6420, Opteron QC 2.3 Ghz, Infiniband	United States	2008	62976	433200	AMD x86_64	Sun Blade System	Infiniband
7	Cray XT4 QuadCore 2.3 GHz	United States	2008	38642	266300	AMD x86_64	Cray XT	Proprietary
8	Cray XT4 QuadCore 2.1 GHz	United States	2008	30976	205000	AMD x86_64	Cray XT	Proprietary
9	Sandia/ Cray Red Storm, XT3/4, 2.4/2.2 GHz dual/quad core	United States	2008	38208	204200	AMD x86_64	Cray XT	Cray Inter-connect
10	Dawning 5000A, QC Opteron 1.9 Ghz, Infiniband, Windows HPC 2008	China	2008	30720	180600	AMD x86_64	Dawning Cluster	Infiniband

Top 500 Nov 2008

Rank	Computer	Country	Year	Cores	RMax	Processor Family	System Family	Inter-connect
11	Blue Gene/P Solution	Germany	2007	65536	180000	Power	IBM BlueGene	Proprietary
12	SGI Altix ICE 8200, Xeon quad core 3.0 GHz	United States	2007	14336	133200	Intel EM64T	SGI Altix	Infiniband
13	Cluster Platform 3000 BL460c, Xeon 53xx 3GHz, Infiniband	India	2008	14384	132800	Intel EM64T	HP Cluster Platform 3000BL	Infiniband
14	SGI Altix ICE 8200EX, Xeon quad core 3.0 GHz	France	2008	12288	128400	Intel EM64T	SGI Altix	Infiniband
15	Cray XT4 QuadCore 2.3 GHz	United States	2008	17956	125128	AMD x86_64	Cray XT	Proprietary
16	Blue Gene/P Solution	France	2008	40960	112500	Power	IBM BlueGene	Proprietary
17	SGI Altix ICE 8200EX, Xeon quad core 3.0 GHz	France	2008	10240	106100	Intel EM64T	SGI Altix	Infiniband
18	Cluster Platform 3000 BL460c, Xeon 53xx 2.66GHz, Infiniband	Sweden	2007	13728	102800	Intel EM64T	HP Cluster Platform 3000BL	Infiniband
19	DeepComp 7000, HS21/x3950 Cluster, Xeon QC HT 3 GHz/2.93 GHz, Infiniband	China	2008	12216	102800	Intel EM64T	Lenovo Cluster	Infiniband

Top 500 Nov 2009

Rank	Site	Computer/Year Vendor	Cores	R_{max}	R_{peak}	Power
1	Oak Ridge National Laboratory United States	Jaguar - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	224162	1759.00	2331.00	6950.60
2	DOE/NNSA/LANL United States	Roadrunner - BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband / 2009 IBM	122400	1042.00	1375.78	2345.50
3	National Institute for Computational Sciences/University of Tennessee United States	Kraken XT5 - Cray XT5-HE Opteron Six Core 2.6 GHz / 2009 Cray Inc.	98928	831.70	1028.85	
4	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution / 2009 IBM	294912	825.50	1002.70	2268.00
5	National SuperComputer Center in Tianjin/NUDT China	Tianhe-1 - NUDT TH-1 Cluster, Xeon E5540/E5450, ATI Radeon HD 4870 2, Infiniband / 2009 NUDT	71680	563.10	1206.19	
6	NASA/Ames Research Center/NAS United States	Pleiades - SGI Altix ICE 8200EX, Xeon QC 3.0 GHz/Nehalem EP 2.93 Ghz / 2009 SGI	56320	544.30	673.26	2348.00
7	DOE/NNSA/LLNL United States	BlueGene/L - eServer Blue Gene Solution / 2007 IBM	212992	478.20	596.38	2329.60
8	Argonne National Laboratory United States	Blue Gene/P Solution / 2007 IBM	163840	458.61	557.06	1260.00
9	Texas Advanced Computing Center/Univ. of Texas United States	Ranger - SunBlade x6420, Opteron QC 2.3 Ghz, Infiniband / 2008 Sun Microsystems	62976	433.20	579.38	2000.00
10	Sandia National Laboratories / National Renewable Energy Laboratory United States	Red Sky - Sun Blade x6275, Xeon X55xx 2.93 Ghz, Infiniband / 2009 Sun Microsystems	41616	423.90	487.74	

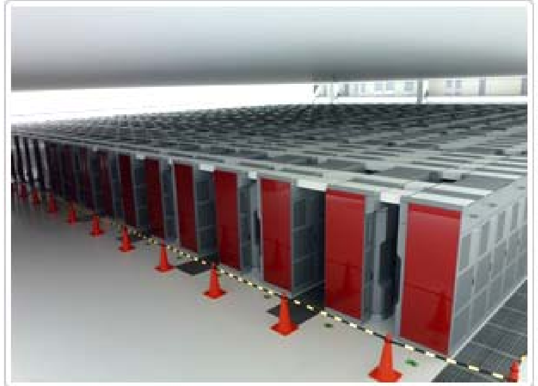
Top 500 Nov 2011

Rank	Computer	Site	Country	Total Cores	Rmax	Rpeak
1	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	RIKEN Advanced Institute for Computational Science (AICS)	Japan	705024	10510000	11280384
2	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050	National Supercomputing Center in Tianjin	China	186368	2566000	4701000
3	Cray XT5-HE Opteron 6-core 2.6 GHz	DOE/SC/Oak Ridge National Laboratory	United States	224162	1759000	2331000
4	Dawning TC3600 Blade System, Xeon X5650 6C 2.66GHz, Infiniband QDR, NVIDIA 2050	National Supercomputing Centre in Shenzhen (NSCS)	China	120640	1271000	2984300
5	HP ProLiant SL390s G7 Xeon 6C X5670, Nvidia GPU, Linux/Windows	GSIC Center, Tokyo Institute of Technology	Japan	73278	1192000	2287630
6	Cray XE6, Opteron 6136 8C 2.40GHz, Custom	DOE/NNSA/LANL/SNL	United States	142272	1110000	1365811
7	SGI Altix ICE 8200EX/8400EX, Xeon HT QC 3.0/Xeon 5570/5670 2.93 Ghz, Infiniband	NASA/Ames Research Center/NAS	United States	111104	1088000	1315328
8	Cray XE6, Opteron 6172 12C 2.10GHz, Custom	DOE/SC/LBNL/NERSC	United States	153408	1054000	1288627
9	Bull bullx super-node S6010/S6030	Commissariat a l'Energie Atomique (CEA)	France	138368	1050000	1254550
10	BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2 Ghz / Opteron DC 1.8 GHz, Voltaire Infiniband	DOE/NNSA/LANL	United States	122400	1042000	1375776

K-computer (AICS), Rank 1

- 800 racks with 88,128 SPARC64 CPUs / 705,024 cores
- 200,000 cables with total length over 1,000km
  - third floor (50m x 60m) free of structural pillars.
- New problem: floor load capacity of 1t/m^2 (avg.)
  - Each rack has up to 1.5t
  - High-tech construction required
- first floor with global file system has just three pillars
- Own power station (300m^2)

The K-Computer



Jugene (Forschungszentrum Jülich), Rank 13

Top 500 Nov 2012

Rank	Name	Computer	Site	Country	Total Cores	Rmax	Rpeak
1	Titan	Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x	DOE/SC/Oak Ridge National Laboratory	United States	560640	17590000	27112550
2	Sequoia	BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	DOE/NNSA/LLNL	United States	1572864	16324751	20132659,2
3	K computer	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect	RIKEN Advanced Institute for Computational Science (AICS)	Japan	705024	10510000	11280384
4	Mira	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	DOE/SC/Argonne National Laboratory	United States	786432	8162376	10066330
5	JUQUEEN	BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect	Forschungszentrum Juelich (FZJ)	Germany	393216	4141180	5033165
6	SuperMUC	iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR	Leibniz Rechenzentrum	Germany	147456	2897000	3185050
7	Stampede	PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi	Texas Advanced Computing Center/Univ. of Texas	United States	204900	2660290	3958965
8	Tianhe-1A	NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050	National Supercomputing Center in Tianjin	China	186368	2566000	4701000
9	Fermi	BlueGene/Q, Power BQC 16C 1.60GHz, Custom	CINECA	Italy	163840	1725492	2097152
10	DARPA Trial Subset	Power 775, POWER7 8C 3.836GHz, Custom Interconnect	IBM Development Engineering	United States	63360	1515000	1944391,68

Titan (Oak Ridge National Laboratory), Rank 1

Cray XK7

- 18,688 nodes with 299,008 cores, 710 TB (598 TB CPU and 112 TB GPU)
    - AMD Opteron 6274 16-core CPU
    - Nvidia Tesla K20X GPU
    - Memory: 32GB (CPU) + 6 GB (GPU)
- Cray Linux Environment
- Power 8.2 MW

JUQUEEN (Forschungszentrum Jülich), Rank 5

- 28 racks (7 rows à 4 racks)
- 28,672 nodes (458,752 cores)
- Rack: 2 midplanes à 16 nodeboards (16,384 cores)
- Nodeboard: 32 compute nodes
- Node: 16 cores
  IBM PowerPC A2, 1.6 GHz

- Main memory: 448 TB (16 GB per node)

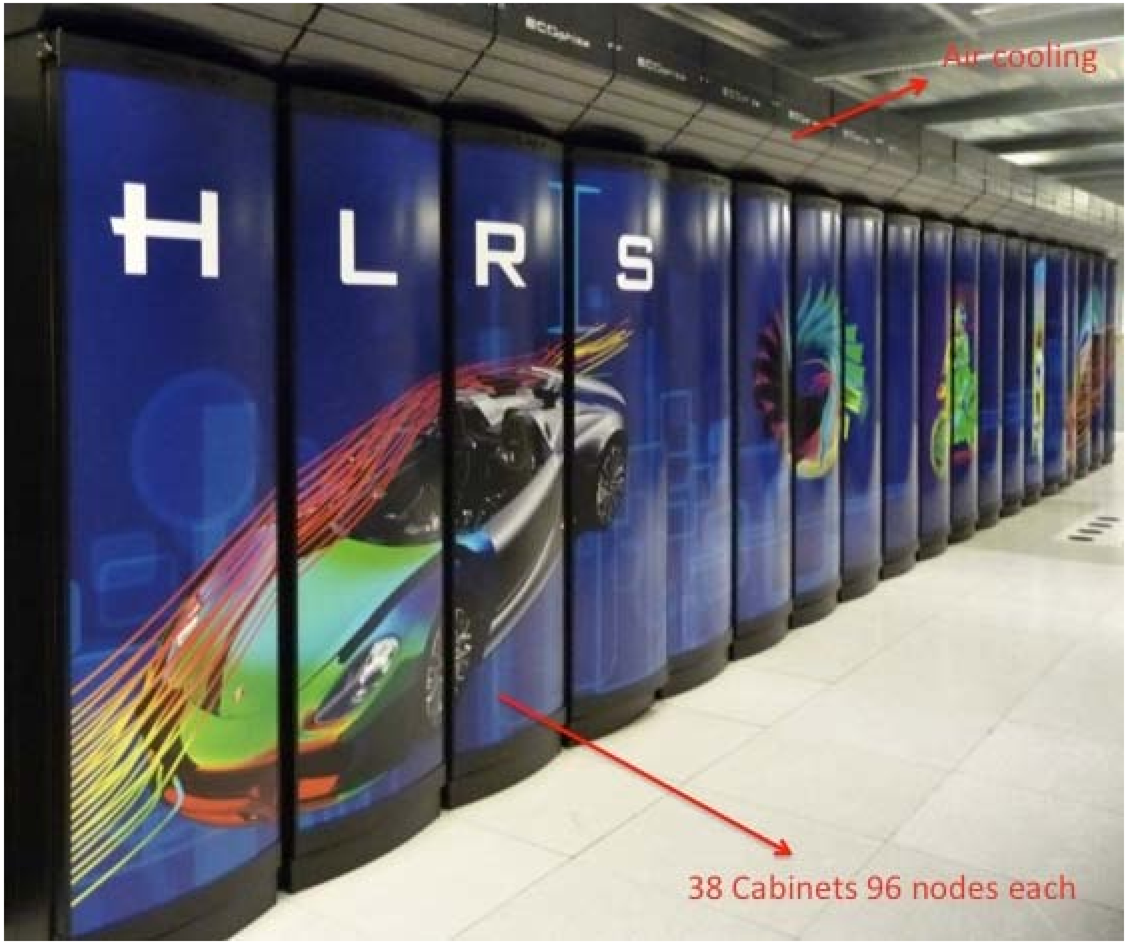
- Network: 5D Torus — 40 GBps; 2.5 μsec latency (worst case)

- Overall peak performance: 5.9 Petaflops
- Linpack: > 4.141 Petaflops

HERMIT (HLRS Stuttgart), Rank 27

- Cray XE6

- 3.552 Nodes
- 113.664 Cores
- 126TB RAM



HLRN II (Hannover + Berlin), Rank 274 + 275

Volunteer Computing

2^P-1 MAY BE PRIME!

Great Internet Mersenne Prime Search

GIMPS

Finding World Record Primes Since 1996

SETI@HOME

POWERED BY BOINC

distributed.net

Folding@home

Grid'5000

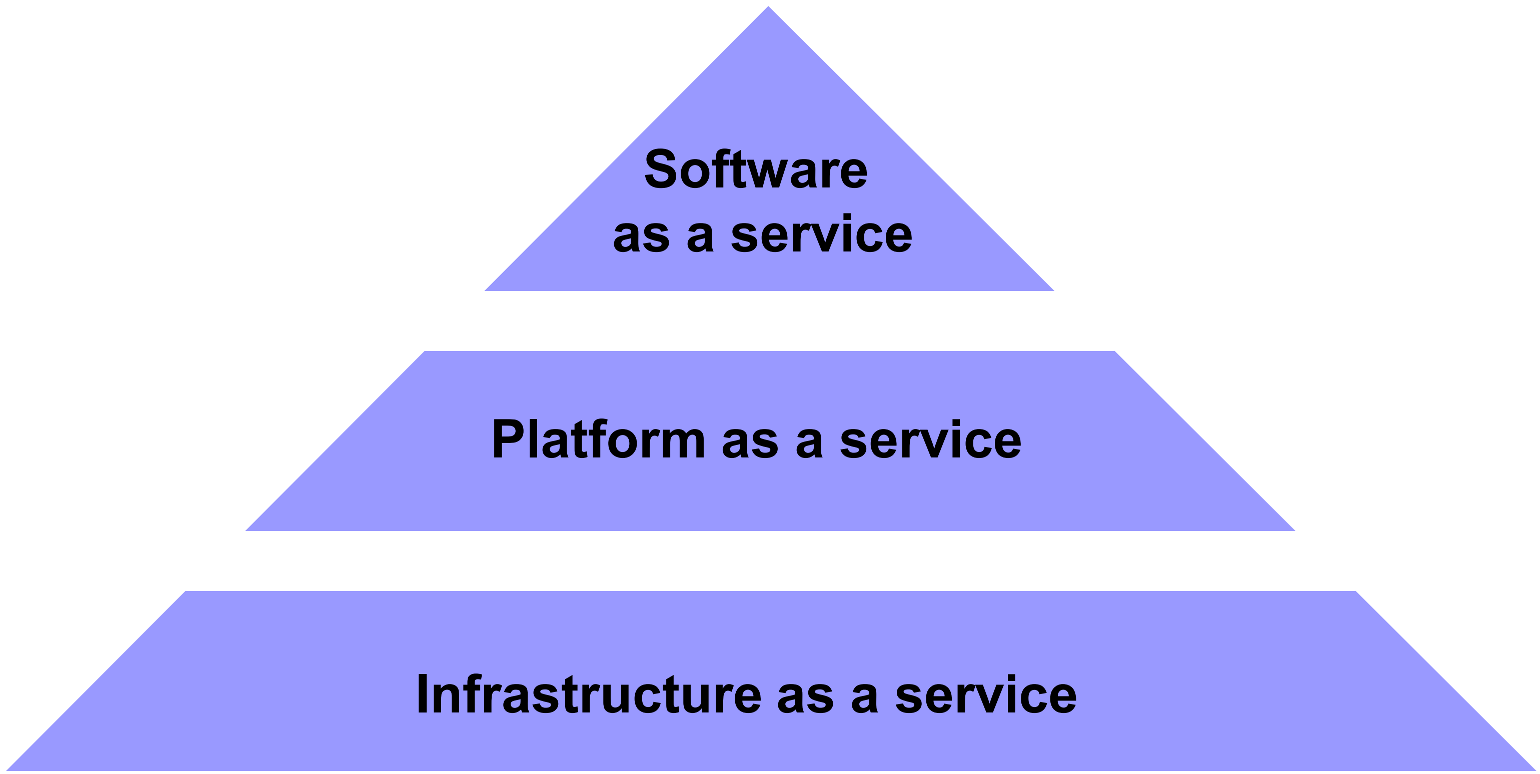
D-GRID

Grid'5000 architecture

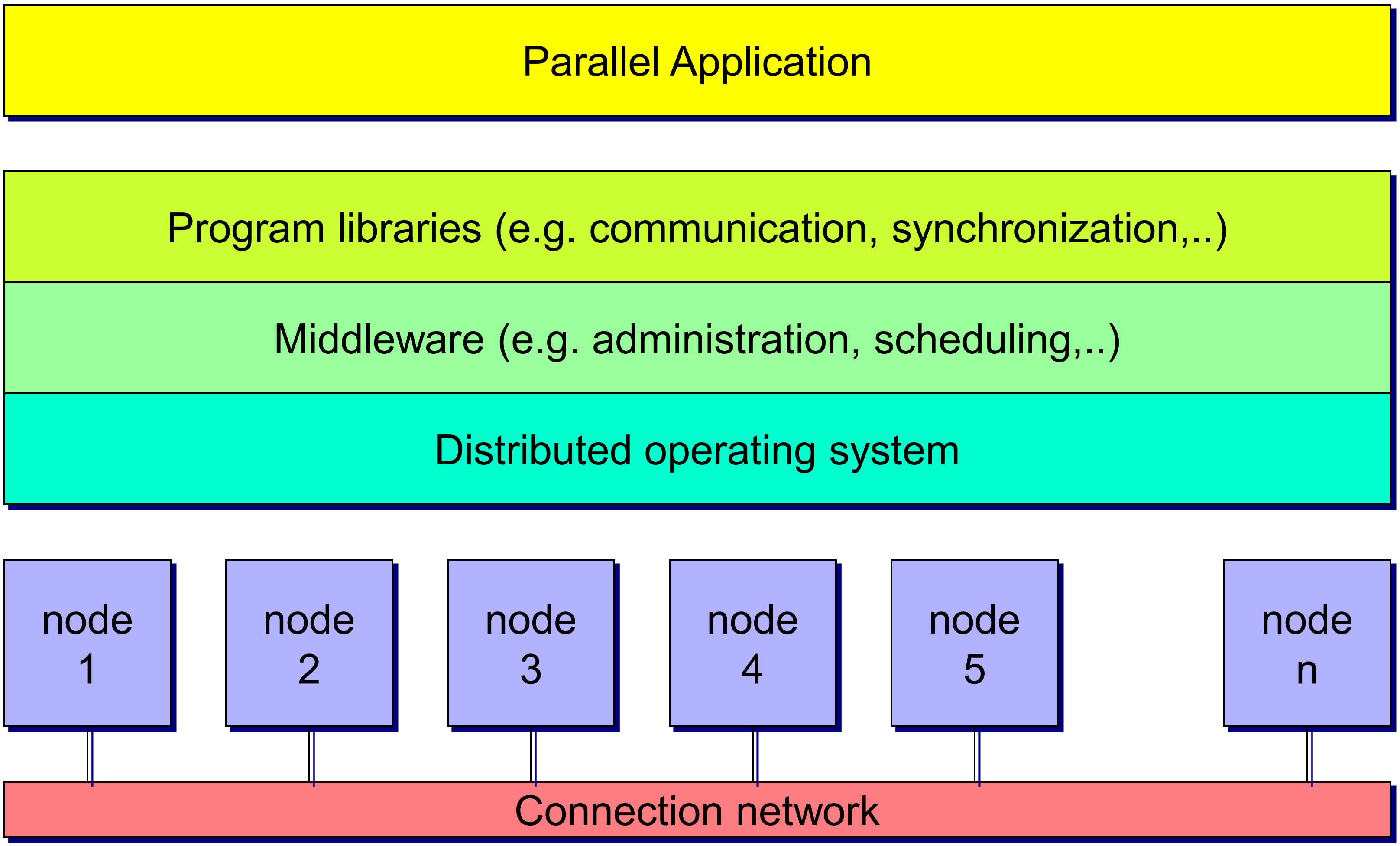
10 Gbps lambda activated for Grid'5000

Cloud Computing (XaaS)

- Usage of remote physical and logical resources



Support for parallel programs



Tasks

User‘s point of view (programming comfort, short response times)

- Efficient Interaction (Information exchange)
  - Message passing
  - Shared memory
- Synchronization
- Automatic distribution and allocation of code and data
- Load balancing
- Debugging support
- Security
- Machine independence

Operator‘s point of view (High utilization, high throughput)

- Multiprogramming / Partitioning
  - Time sharing
  - Space sharing
- Load distribution / Load balancing
- Stability
- Security