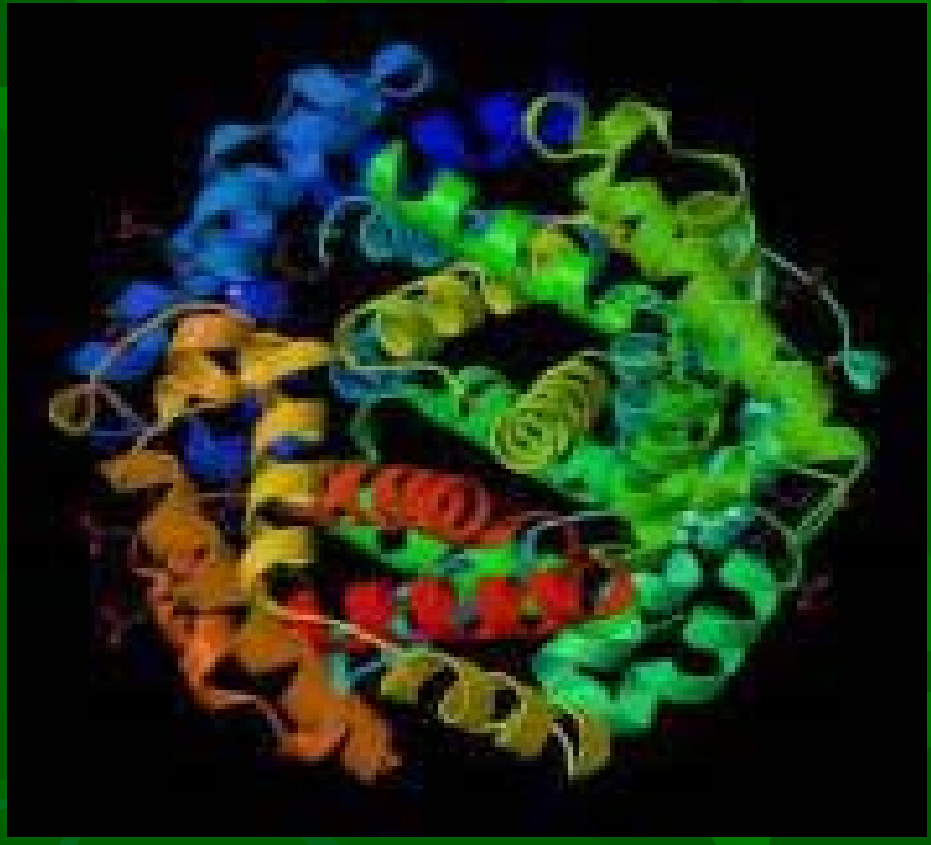
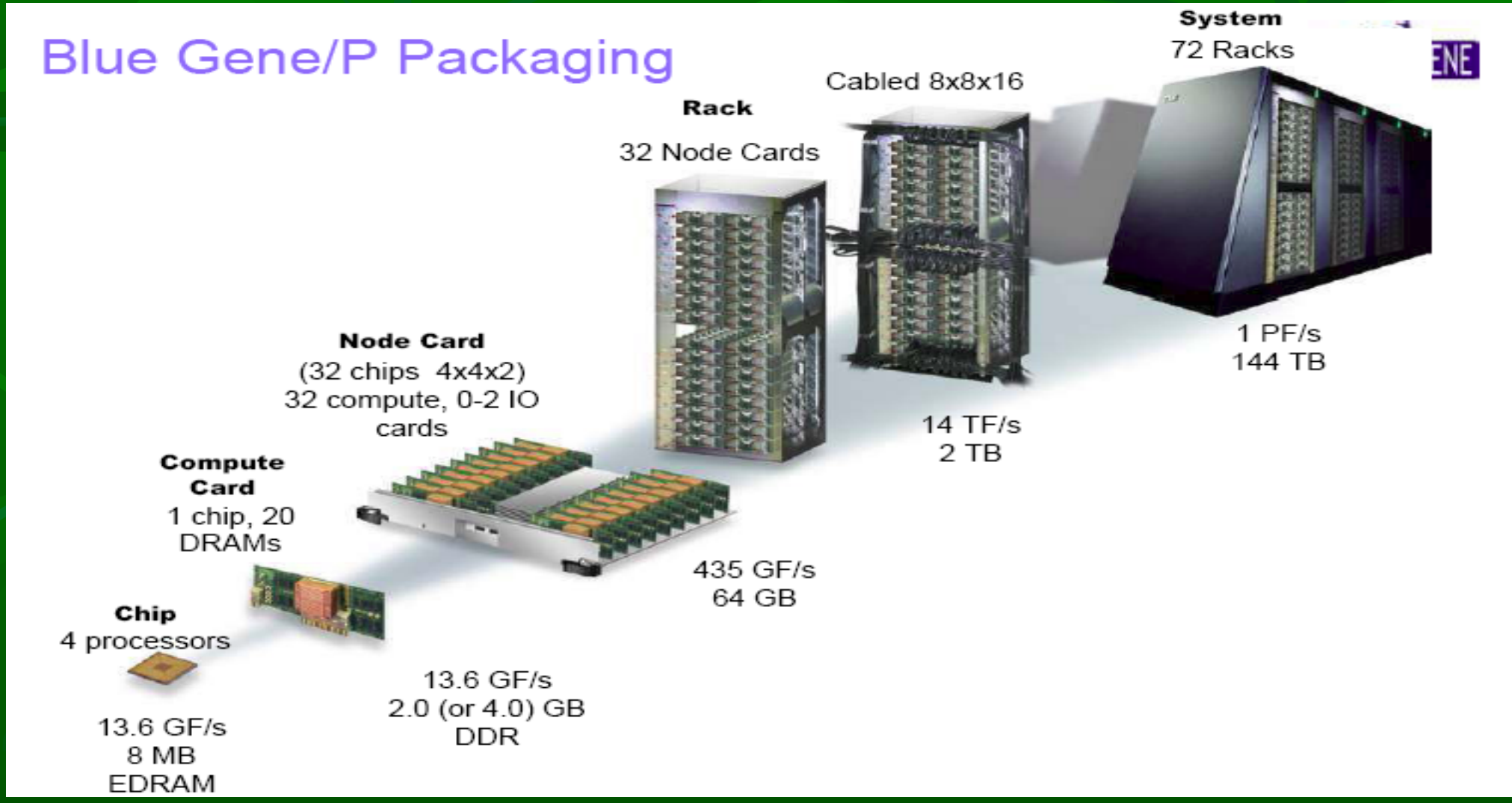
BlueGene/P
Българският Суперкомпютър



Българският Суперкомпютърен Център

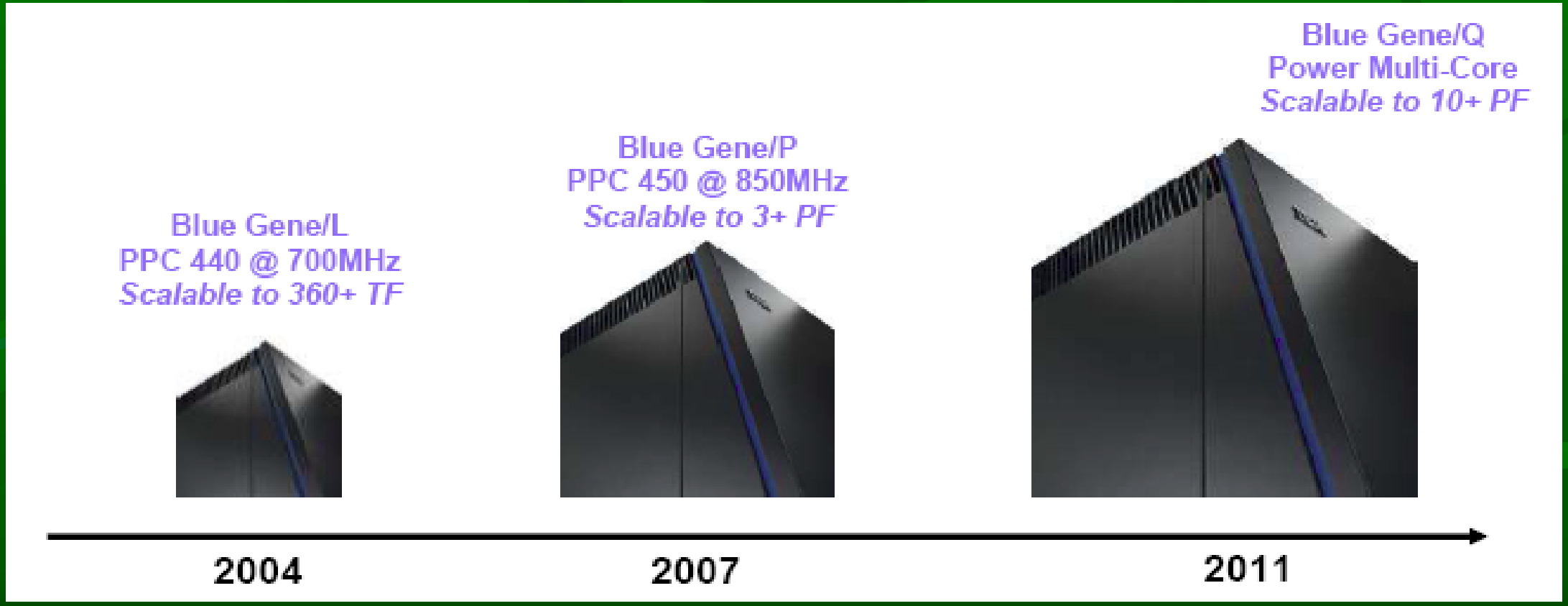
Българският суперкомпютърен център (БСЦ), разположен в сградата на Държавната агенция за информационни технологии и съобщения в София, експлоатира и осигурява достъп до суперкомпютър IBM Blue Gene/P, състоящ се от два шкафа:



Какво представлява BlueGene?

Масивна Паралелна Архитектура

❖ Компютърна система с много независими модули и микро-процесори, който са сързани в една компютърна машина с голяма изчислителна мощност;



Сравнение BlueGene/L и BlueGene/P

Property		BG/L	BG/P
Node Properties	Node Processors	2 * 440 PowerPC	4 * 450 PowerPC
	Processor Frequency	0.7GHz	0.85GHz
	Coherency	Software managed	SMP
	L1 Cache (private)	32KB/processor	32KB/processor
	L2 Cache (private)	14 stream prefetching	14 stream prefetching
	L3 Cache size (shared)	4MB	8MB
	Main Store/Node	512MB and 1GB versions	2GB (studying 4GB) versions
	Main Store Bandwidth	5.6GB/s (16B wide)	13.6 GB/s (2*16B wide)
	Peak Performance	5.6GF/node	13.6 GF/node
Torus Network	Bandwidth	6*2*175MB/s = 2.1GB/s	6*2*425MB/s = 5.1GB/s
	Hardware Latency (Nearest Neighbor)	200ns (32B packet) 1.6us (256B packet)	64ns (32B packet) 512ns (256B packet)
	Hardware Latency (Worst Case)	6.4us (64 hops)	3.0us (64 hops)
Collective Network	Bandwidth	2*350MB/s = 700MB/s	2*0.85GB/s = 1.7GB/s
	Hardware Latency (Round Trip Worst Case)	5.0us	2.5us
System Properties	Peak Performance	360TF (64K nodes)	1PF (72K nodes)
	Total Power	1.5MW	> 2.0MW

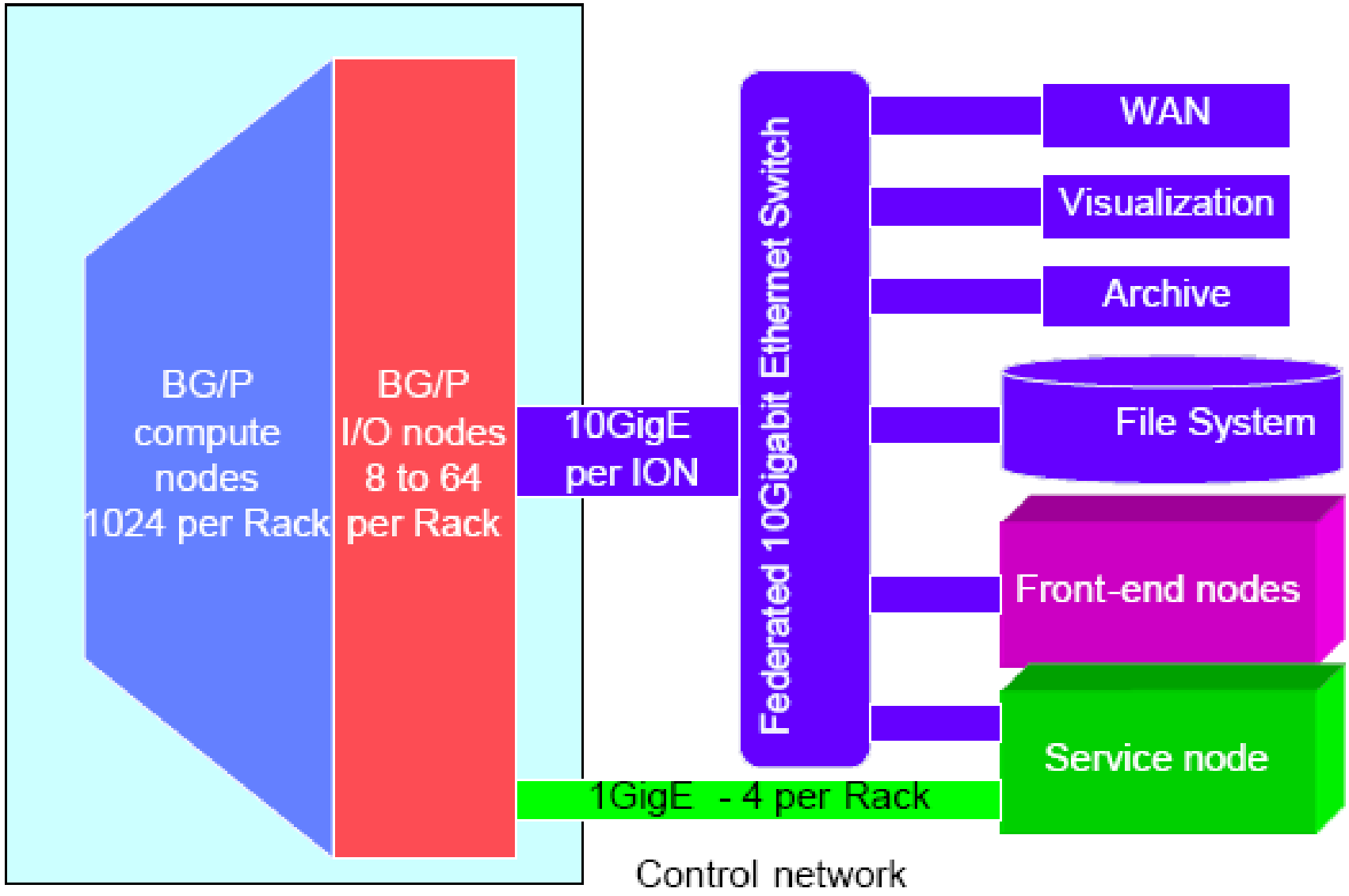
BlueGene Концепция

Паралелизацията води до висока производителност

- Висока Производителност постигната чрез използването на паралелни приложения

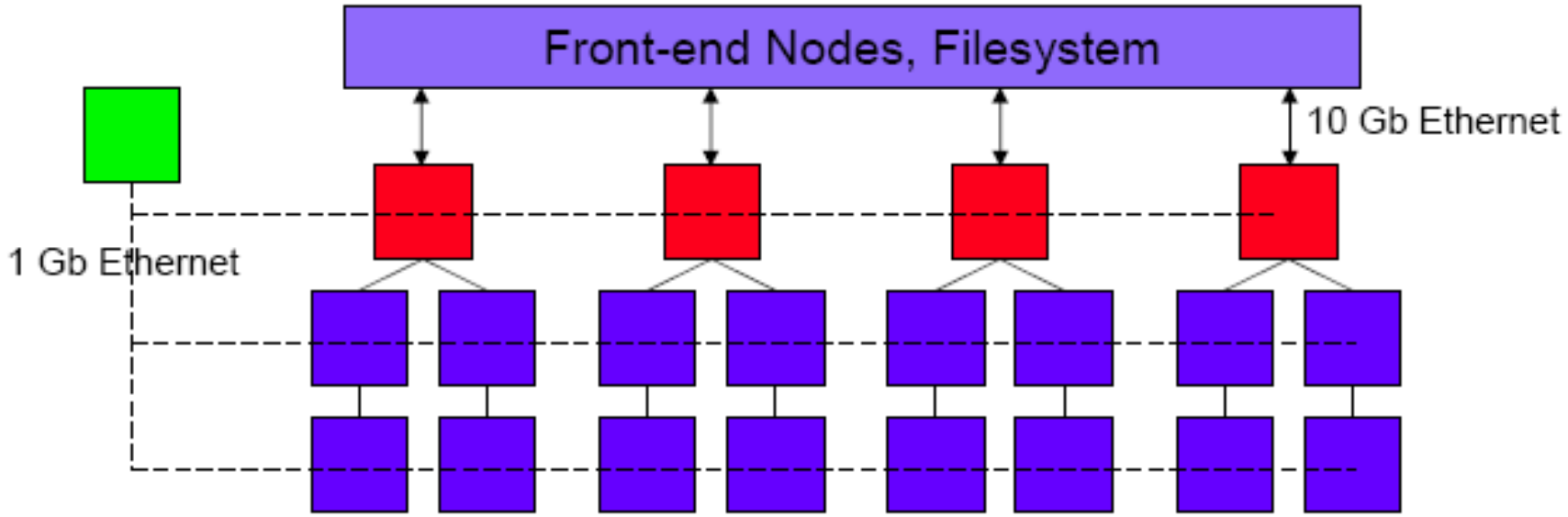
- Паралелизъм на ниво данни (Data level parallelism) с SIMD
  - Работа с множество елементи;
  - SIMD е ефективен
- Паралелизъм на ниво нишки (Thread level parallelism) с многоядрен “multi-core” дизайн
  - 2×ядра = 2×производителност = 2×сила
  - IBM е пионер в сферата на “multicore” с POWER4 от преди повече от 10 години

Пълна конфигурация на BlueGene/P

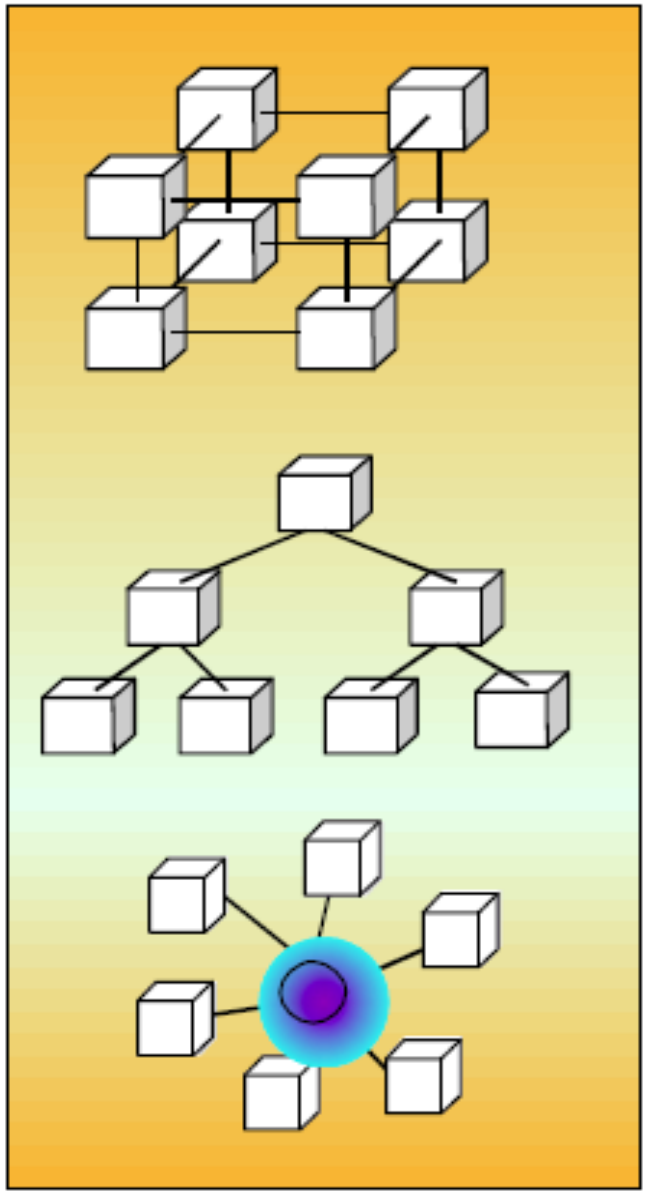


Йерархична организация на BlueGene блоковете

- Compute Nodes dedicated to running user application, and almost nothing else - simple compute node kernel (CNK)
- I/O Nodes run Linux and provide a more complete range of OS services – files, sockets, process launch, signaling, debugging, and termination
- Service Node performs system management services (e.g., partitioning, heart beating, monitoring errors) - transparent to application software



BlueGene Мрежи



- **3-Dimensional Torus**
  - Interconnects all compute nodes
  - Virtual cut-through hardware routing
  - 3.4 Gb/s on all 12 node links (5.1 GB/s per node)
  - 0.5 μs latency between nearest neighbors, 5 μs to the farthest
  - MPI: 3 μs latency for one hop, 10 μs to the farthest
  - Communications backbone for computations
  - 1.7/3.9 TB/s bisection bandwidth, 188TB/s total bandwidth
- **Collective Network**
  - One-to-all broadcast functionality
  - Reduction operations functionality
  - 6.8 Gb/s of bandwidth per link
  - Latency of one way tree traversal 1.3 μs, MPI 5 μs
  - ~62TB/s total binary tree bandwidth (72k machine)
  - Interconnects all compute and I/O nodes (1152)
- **Low Latency Global Barrier and Interrupt**
  - Latency of one way to reach all 72K nodes 0.65 μs (MPI 1.6 μs)
- **Other networks**
  - 10Gb Functional Ethernet
  - I/O nodes only
  - 1Gb Private Control Ethernet
  - Provides JTAG access to hardware. Accessible only from Service Node system

Конфигурация I

IBM Blue Gene/P:

- 2048 изчислителни възела с PowerPC 450 процесори, 8192 ядра и общо 4 TB оперативна памет.
- Всяко ядро може да обработва по два потока от данни (с двойна точност) с плаваща запетая.

Конфигурация II

IBM Blue Gene/P:

- Шестнайсет входно-изходни възела са свързани посредством оптични влакна към 10 Gb/s Ethernet комутатор, а други допълнителни 16 входно-изходни възела ще бъдат добавени към системата в близко бъдеще.

Конфигурация III

IBM Blue Gene/P консумация:

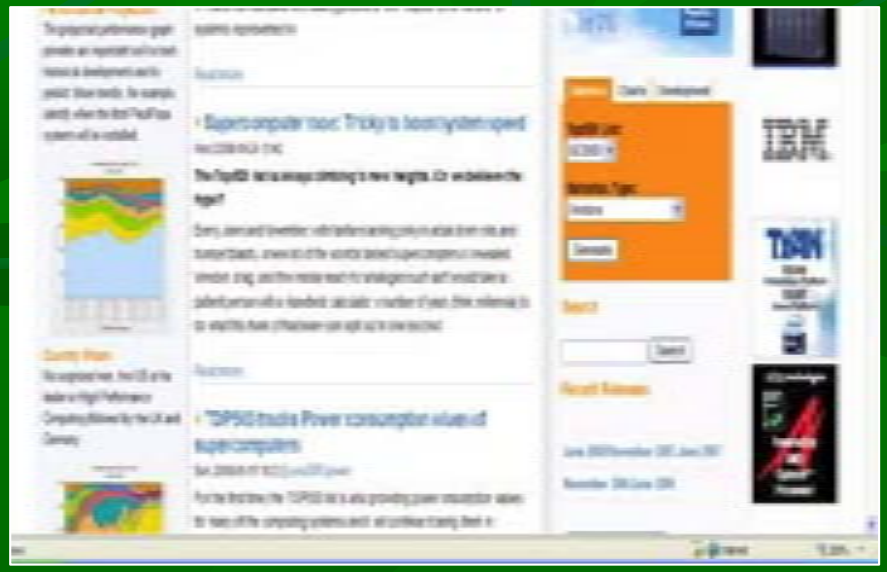
- Начална: 34kW;
- Средна: 50 kW;
- Висока: 68 kW;

Конфигурация IV

<u>IBM Blue Gene/P – Охлаждаща Система:</u>

- Uniflair TDAR 1422A – two units;
- Airflow –15600 m3/h per unit;
- Total cooling capacity –51.6 kW per unit;
- Refrigerant circuits/compressors –two per unit;

Бенчмарк I



Постигната максимална LINPACK производителност: Rmax= 23.42 Tflops;

- ❑ Теоретична производителност: Rpeak= 27.85 Tflops;
- ❑ Blue Gene/P системата на БСЦ е на 126-то място по производителност в света според 32-то издание на TOP500 класацията за суперкомпютри, публикувана на 17 ноември 2008

Бенчмарк II

Енергийна Производителност: 371.67 MFlops/W

- **Споделя 7 място през юли в листа на GREEN500 редом със суперкомпютри инсталирани в**: RZG/Max-Planck-Gesellschaft MPI/IPP, Stony Brook/BNL –New York Center for Computational Sciences, ASTRON/University Groningen, IBM –Rochester and DOE/Oak Ridge National Laboratory;

Софтуер

Компилатори

- IBM XL C/C++ Advanced Edition за Blue Gene/P V9.0
- IBM XL Fortran Advanced Edition за Blue Gene/P V11.1
- GNU Toolchain (gcc, glibc, binutlils, gdb, python)

Системни Библиотеки

- Engineering and Scientific Subroutine Library (ESSL) за Linux на Power V4.3.1
- MPI (MPICH2)

3. Приложен Софтуер и Библиотеки

Следните приложения и библиотеки са успешно компилирани, както и е потвърдено че се изпълняват добре и нямат проблеми със скалирумостта на Blue Gene/P машината в БСЦ:

- **CPMD:** паралелна имплементация на теорията на функционала на плътността, използваща плоски вълни (псевдопотенциали) и специално разработена за ab initio молекулна динамика
- **GAMMES-US:** общ пакет за ab initio квантова химия
- **GotoBLAS:** най-бързата имплементация на Basic Linear Algebra Subroutines (BLAS) в момента
- **LAMMPS:** код за симулиране на класическа молекулна динамика по ефективен начин на паралелни компютри
- **NAMD:** паралелен код за молекулна динамика на големи биомолекулни системи

В процес на разработка (портване, компилиране, тестване и настройка) са следните приложения и/или библиотеки:

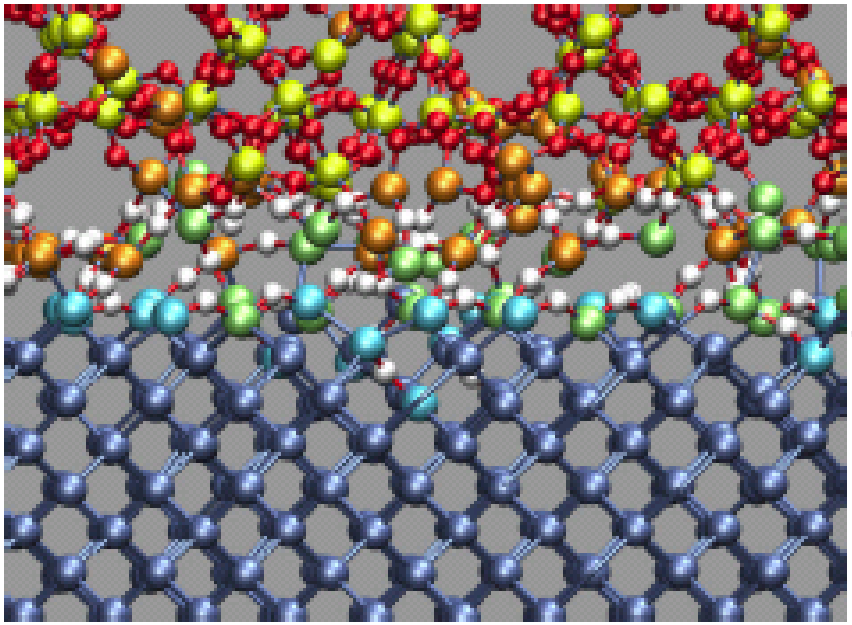
- **GROMACS:** класическа молекулна динамика, симулиране на Нютоновите уравнения за движение за системи с хиляди или милиони частици;
- **mpiBLAST-PIO:** паралелна имплементация на NCBI BLAST (Basic Local Alignment Search Tool, алгоритъм използван в биоинформатиката);
- **OpenAtom:** скалируемо и лесно за портване паралелно приложение за симулиране на молекулна динамика на квантово ниво, което имплементира метода за ab initio молекулна динамика на Car-Parrinelo (CPAIMD);

BlueGene – Приложения I

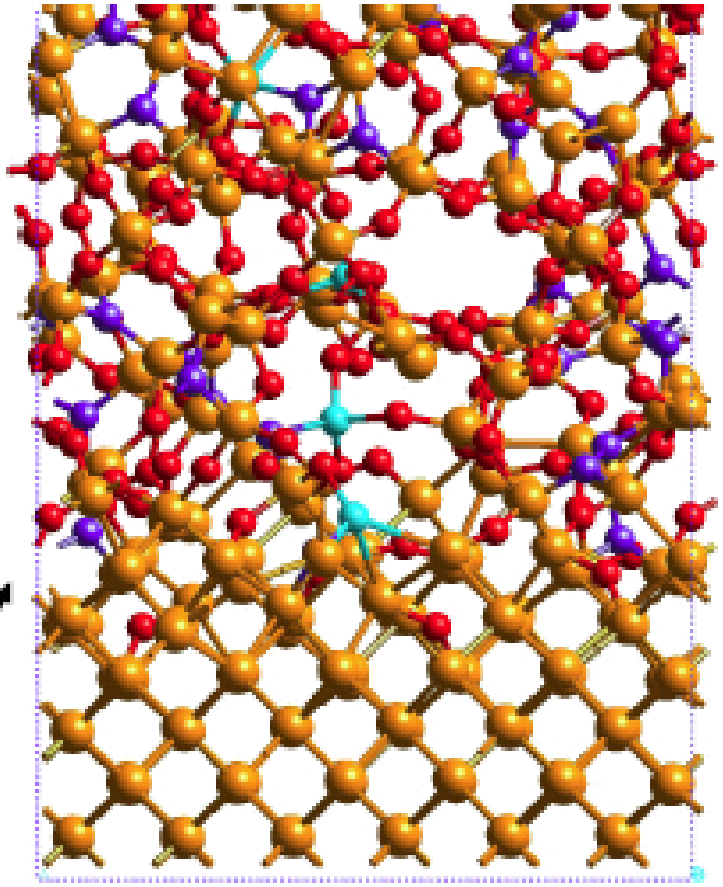
Car-Parrinello Molecular Dynamics (CPMD): Studying the effect on dopants on SiO_2/Si boundaries

Simulations from first principles to understand the physics and chemistry of current technology and guide the design of next-generation materials

Characterization of materials currently under experimental test



Formation of a Non-abrupt SiO_2/Si interface correctly predicted "from scratch"

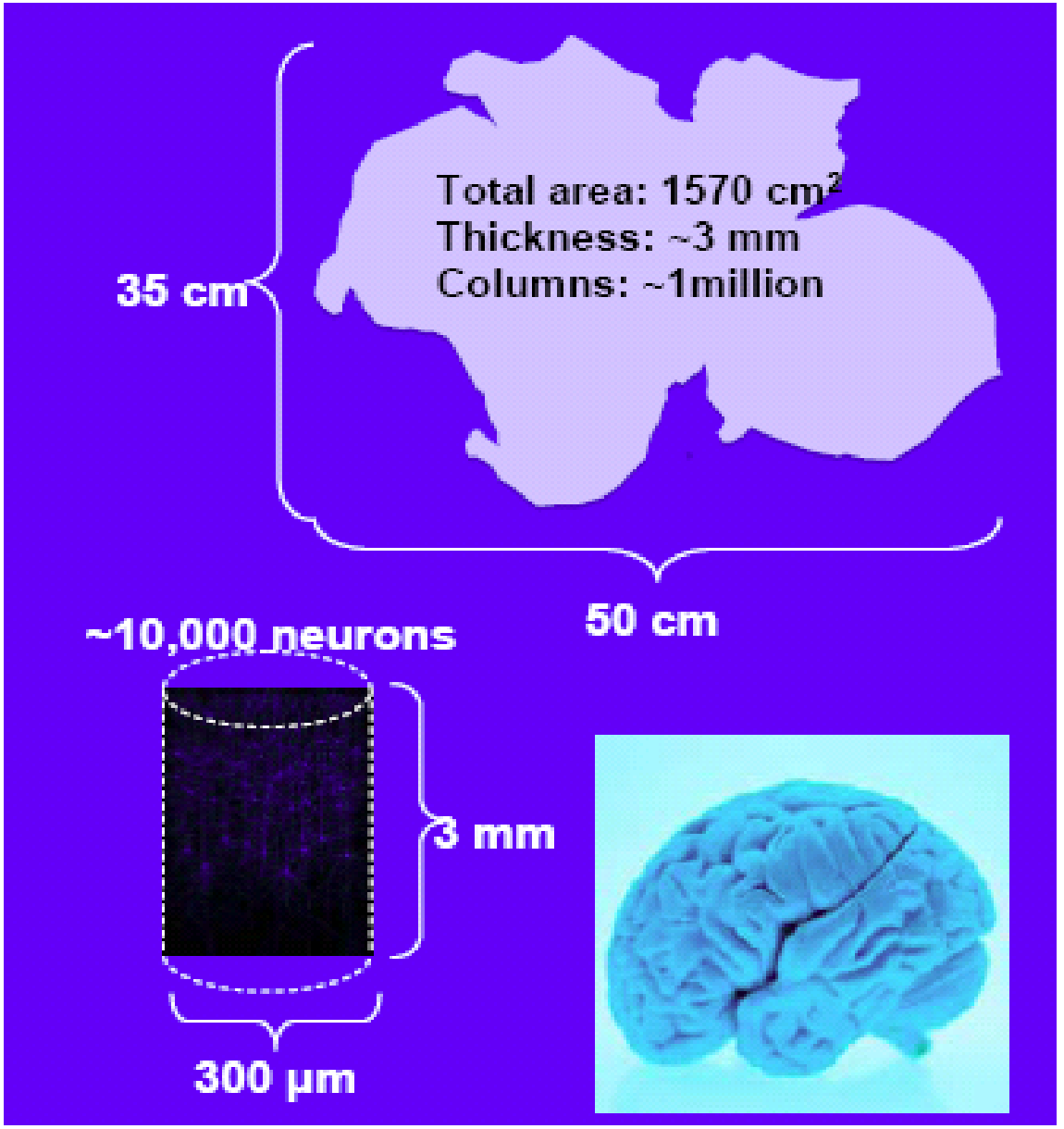


When nitrogen and hafnium are introduced during the simulation process, detrimental defects are unraveled

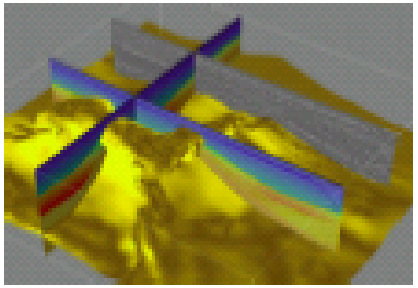
BlueGene – Приложения II

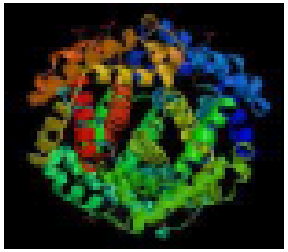
Blue Brain

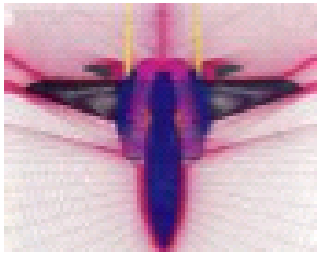
- EPFL to simulate the neocortical column
- Our understanding of the brain is limited by insufficient information and complexity
  - Overcome limitations of neuroscientific experimentation
- Inform experimental design and theory
- Enable scientific discovery for understanding brain function and diseases
- Finally feasible!!! (although by no means finshed)
- 8096 processors (BG/L)
  - 100,000 morphologically complex neurons in real-time

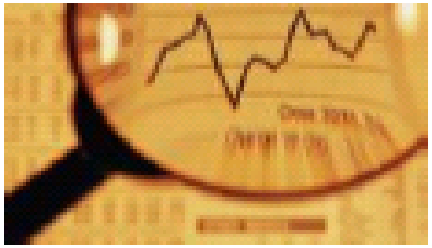


BlueGene – Приложения III





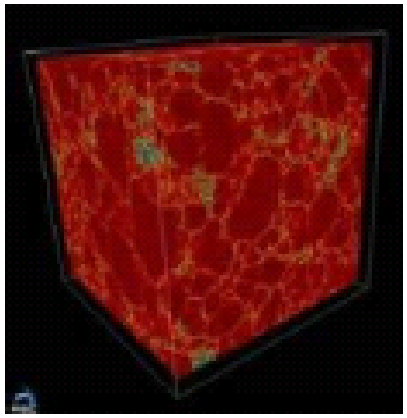




- 2 racks of Blue Gene /P with 8000 processors will allow you to process approximately 1500 square kilometers of seismic data per day, with 41 Million traces per line, which is 3-5 times faster than what's possible on typical clusters in use today. (With the near linear scalability of seismic processing, if you double the BG/P racks, you double the amount of work done.)
- The aggregate memory available in 2 racks of Blue Gene/P is enough to load and run the largest petroleum reservoir models in the world.
- Using 10 racks of BG/P, computation time for a detailed simulation of a human heart beating can be reduced from months to minutes compared to a traditional large cluster.
- Using 2 racks of BG/P one can conduct an in-silico clinical trial on 27 million simulated patients in 5 hours. A typical clinical trial may have several thousand patients and take years.
- A transonic flow simulation of a full-body aircraft structure with 300 million cells can be done in less than 1 hour on a single rack of BG/P.
- One BG/P rack can run 4000 independent pricing calculations in parallel while only consuming 35kW and occupying 1 square meter of data center power and space.
- One BG/P rack, with 2TB of memory and 4000 compute engines, can efficiently perform real-time analysis of multi-Petabyte data streams, which can be used for tracking news streams, updating trade data and monitoring events that impact financial markets.

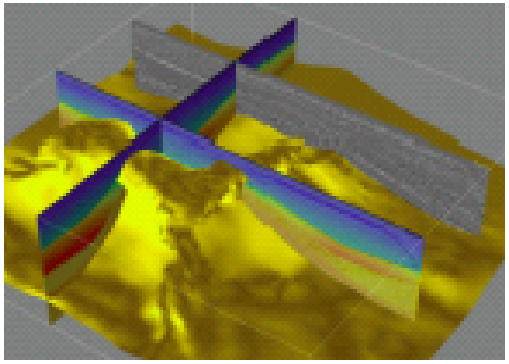
BlueGene – Приложения IV

- **Improve understanding** – significantly larger scale, more complex and higher resolution models; new science applications
- **Multiscale and multiphysics** – From atoms to mega-structures; coupled applications
- **Shorter time to solution** – Answers from months to minutes



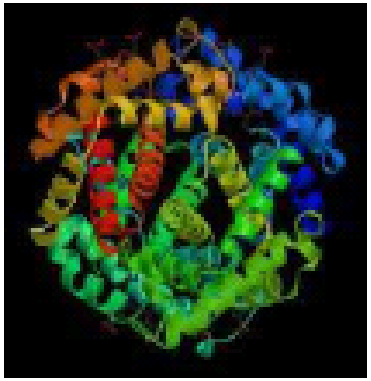
Physics – Materials Science
Molecular Dynamics

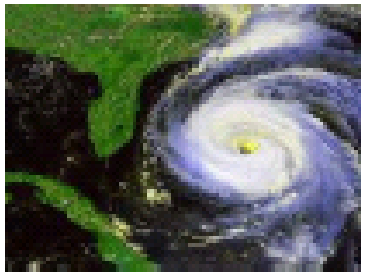
Geophysical Data Processing
Upstream Petroleum



Biological
Modeling – Brain Science

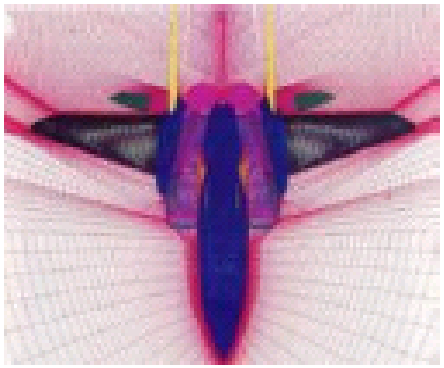
Life Sciences: In-Silico
Trials, Drug Discovery

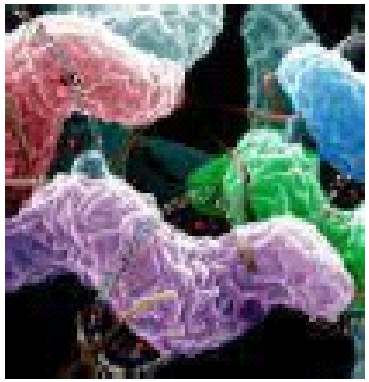




Environment and Climate Modeling

Computational Fluid Dynamics





Life Sciences: Sequencing

Financial Modeling
Streaming Data Analysis

More Information

- IBM Redbooks for Blue Gene
  - Application Development Guide
  - System Administration Guide
  - Performance Tools

www.research.ibm.com/bluegene