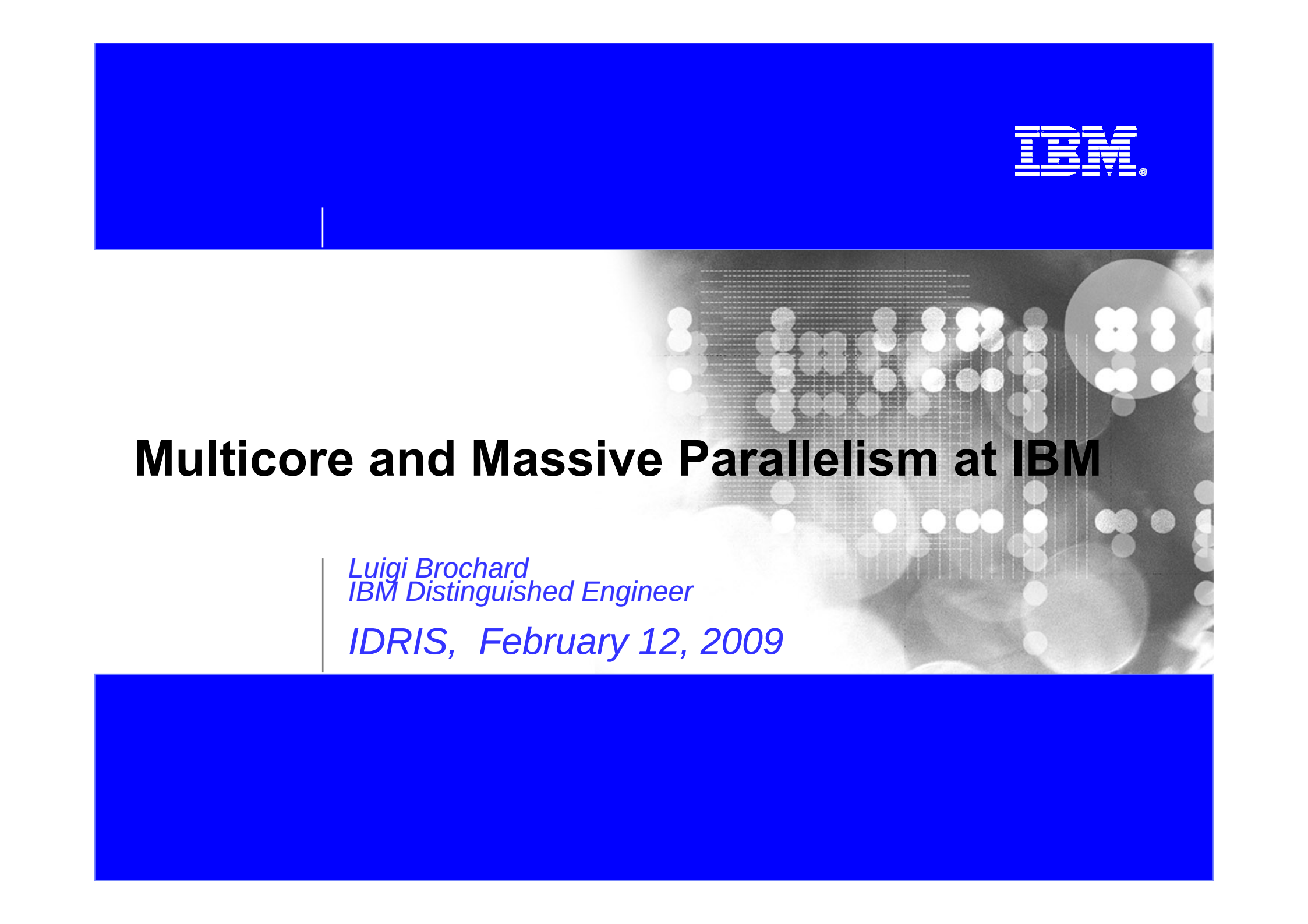


Multicore and Massive Parallelism at IBM

Luigi Brochard
IBM Distinguished Engineer

IDRIS, February 12, 2009

Agenda

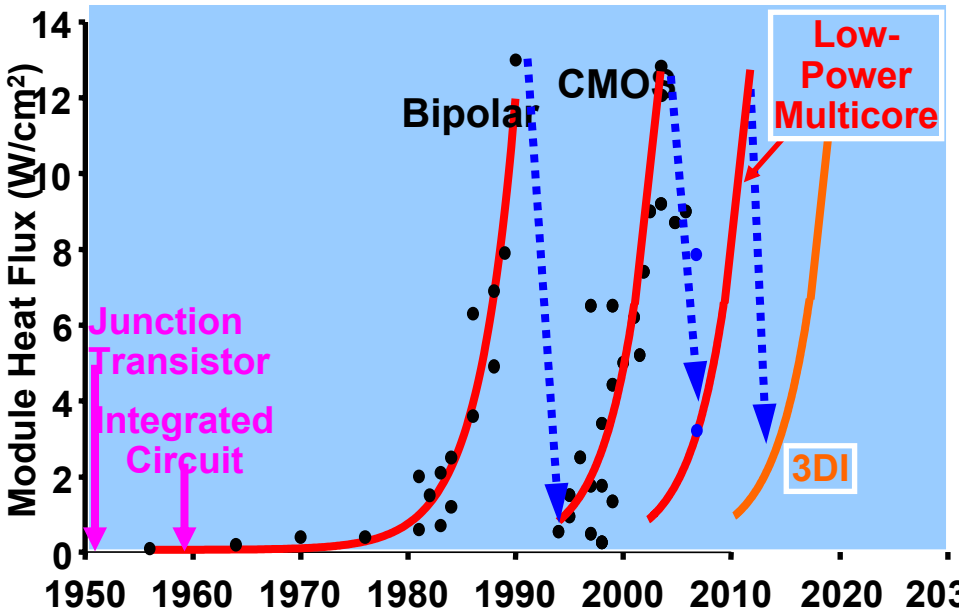
- Multi core and Massive parallelism
- IBM multi core and massive parallel systems
- Programming Models
- Toward a Convergence ?

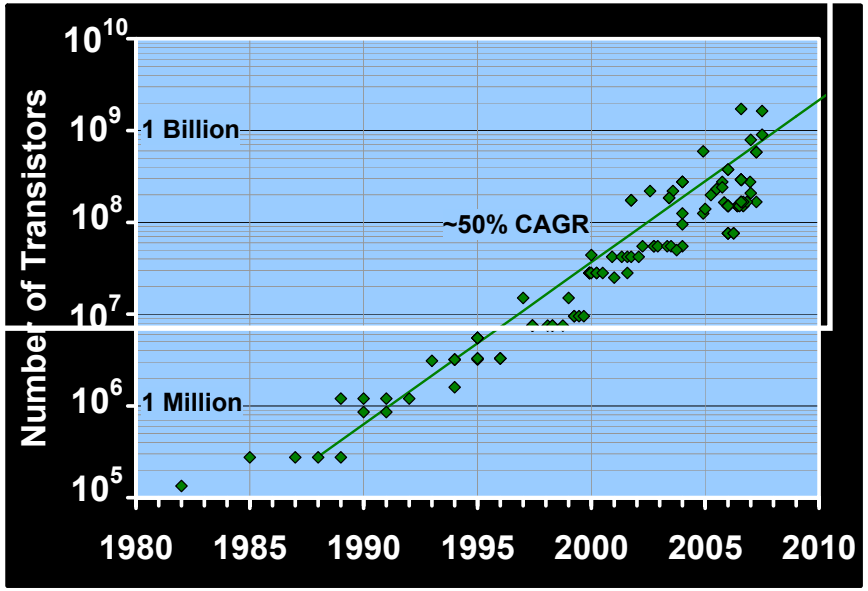
© 2007 IBM Corporation

Why Multicore?

- **Power ~ Voltage^2 * Frequency**
  ~ Frequency^3
- **We have a heat problem:**
  - Power per chip is constant due to cooling
  - => no more frequency improvment
- **But**
  - Smaller lithography => more transistors
  - Decrease frequency by 20% => 50% Power saving
    - => twice more transistors at same power consumption
- **And**
  - Simpler « cpu » => less transistors => more « cpu » per chip

Accelerators





© 2007 IBM Corporation

Why Parallelism?

- Always best power efficient solution
  - 2 processors at frequency F/2 vs 1 processor at frequency F
    - Produce the same work if NO OVERHEAD
    - Consume ¼ of the energy
- Two types of parallelism
  - Internal parallelism is limited by:
    - Number of cores on a chip
    - Performance per wattin internal parallelism
  - External parallelism is limited by:
    - Floor Space
    - Total power consumption !
- Future is Multicore Massive Parallel Systems
  - How to manage/program million cores system/application

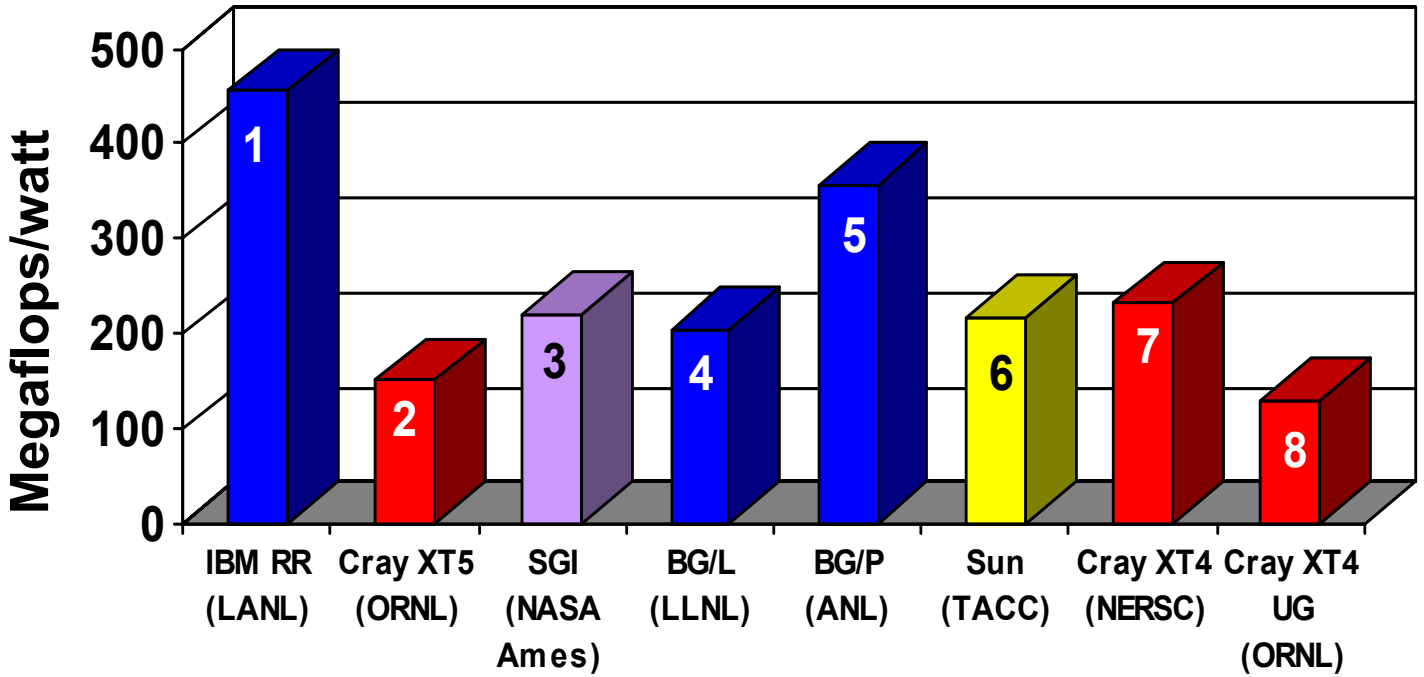
© 2007 IBM Corporation

So far

- **Multicore and Massive Parallelism have developed independently**
  - Multicore :
    - Power4 : first dual core homogeneous microprocessor in 2001
    - Cell BE : first nine core heterogeneous microprocessor in 2004
  - Massively Parallelism
    - Long time ago : CM1, CM2, Cosmic Cube , ...GF11
    - Recently : BG/L in 2004

© 2007 IBM Corporation

Power Efficency (MF/w) of Top 8 TOP500 Systems

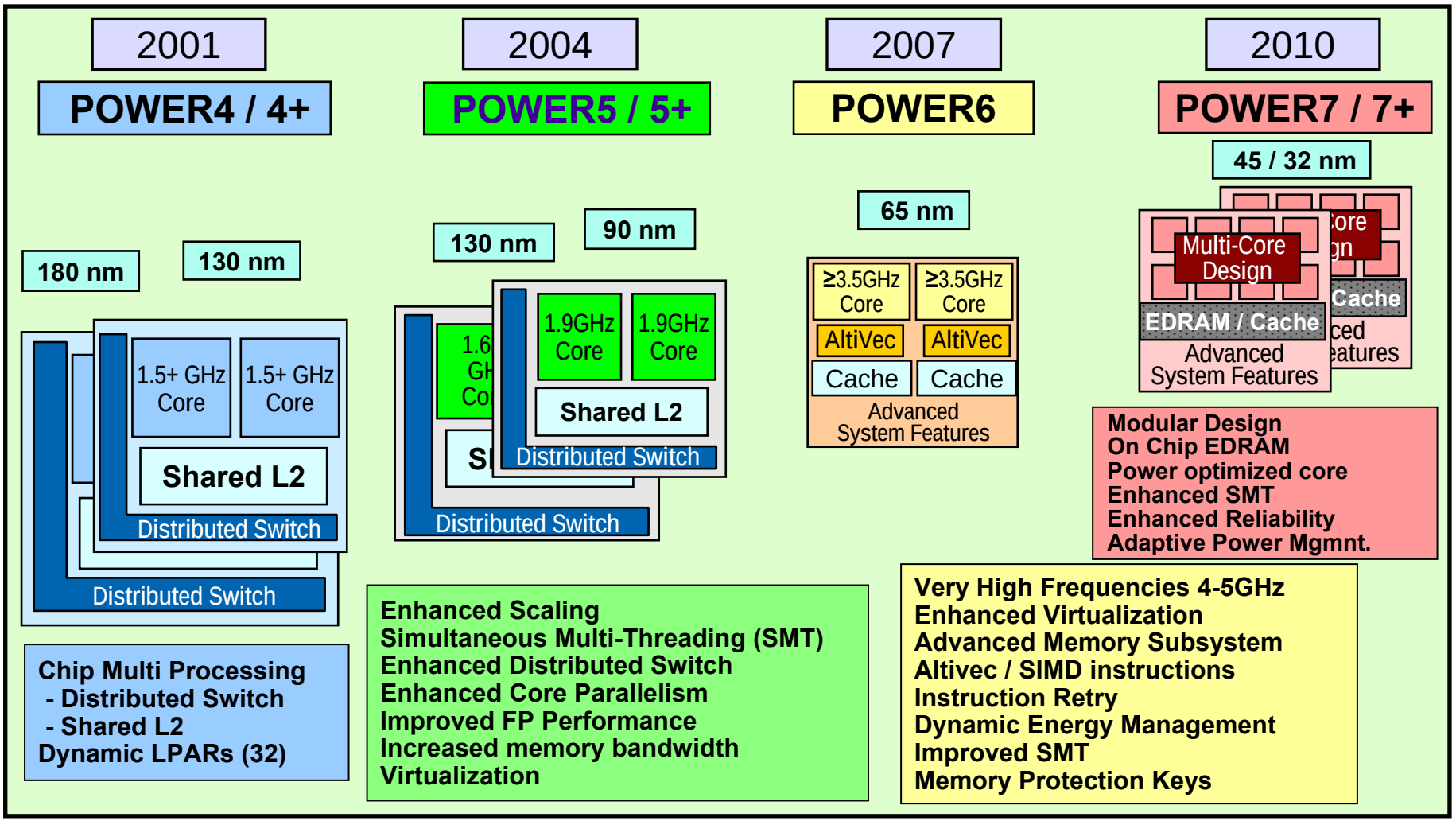


* Number shown in column is Nov 2008 TOP500 rank

Rank	Site	Mfgr	System	MF/w	Relative
1	LANL	IBM	Roadrunner QS22/LS21	445	1.00
2	ORNL	Cray	Jaguar XT5 2.3 GHz QC Opteron	152	2.92
3	NASA Ames	SGI	QC 3.0 Xeon	233	1.91
4	LLNL	IBM	Blue Gene/L	205	2.17
5	ANL	IBM	Blue Gene/P	357	1.25
6	TACC	Sun	2.3 GHz QC Opteron	217	2.05
7	NERSC/LBNL	Cray	Jaguar XT4 2.3 GHz QC Opteron	232	1.92
8	ORNL	Cray	XT4 2.1 GHz QC Opteron	130	3.43

© 2007 IBM Corporation

POWER Processor Roadmap



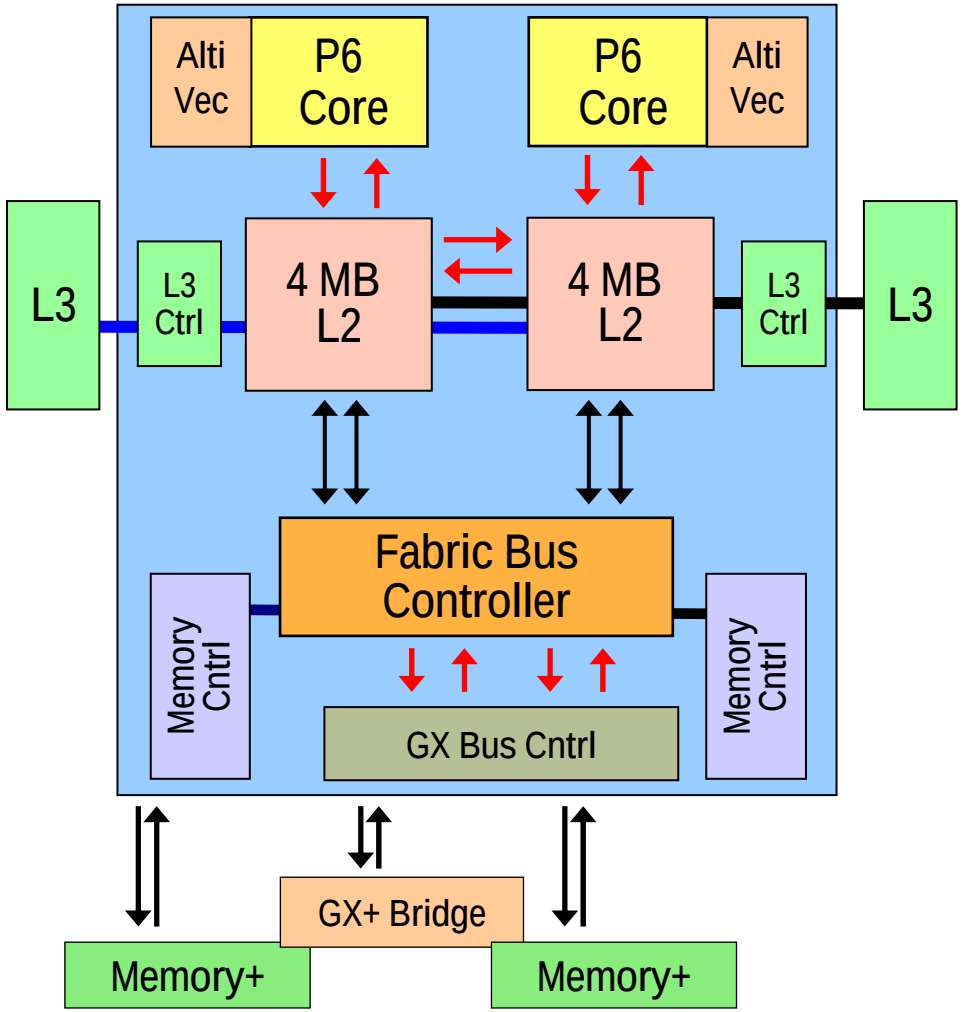
BINARY COMPATIBILITY

© 2007 IBM Corporation

POWER6 Architecture

Features:

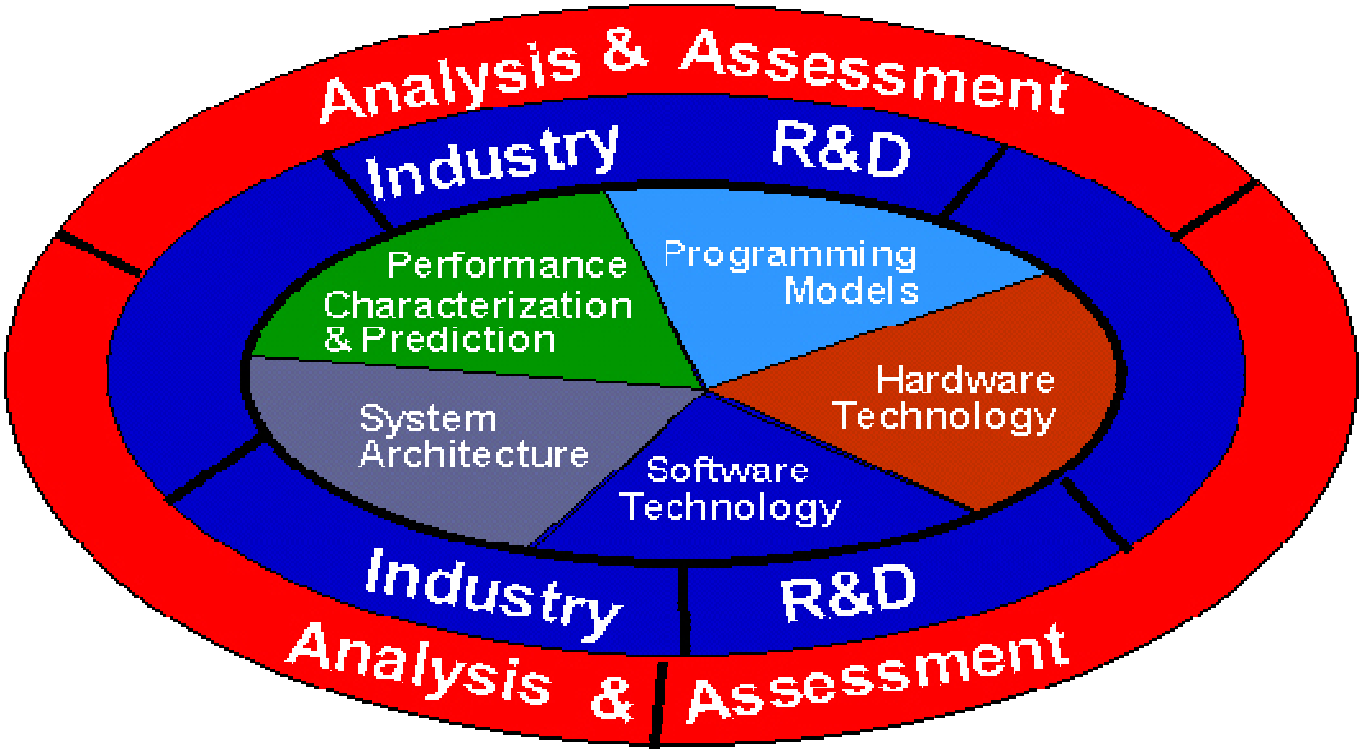
- Ultra High Frequency dual core chip
- 7way superscalar, 2-way SMT core
  - 5 instruction per thread,
- 8 execution units
  - 2LS, 2 FP, 2 FX, 1 BR, 1 VMX
- 790 M transistors, 341 mm2 die
- Upto 128 core SMP systems
- 8MB on chip L2 – point of coherency
- One chip L3 and memory controller
- Two memory controller on chip



© 2007 IBM Corporation

PERCS

PERCS – Productive, Easy-to-use, Reliable Computing System



© 2007 IBM Corporation

High Productivity Computing Systems Overview

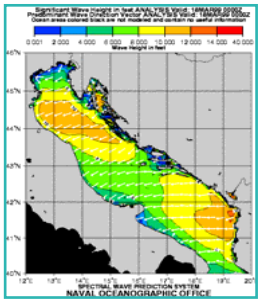
Goal: Provide a new generation of economically viable high productivity computing systems

Impact:

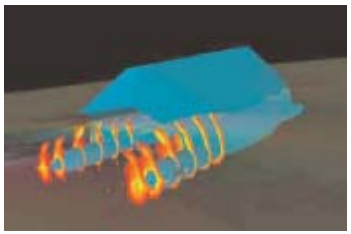
- **Performance** (time-to-solution): speedup by **10X to 40X**
- **Programmability** (idea-to-first-solution): dramatically reduce cost & development time
- **Portability** (transparency): insulate software from system
- **Robustness** (reliability): continue operating in the presence of localized hardware failure, contain the impact of software defects, & minimize likelihood of operator error

Applications:

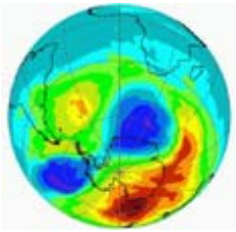
Weather Prediction



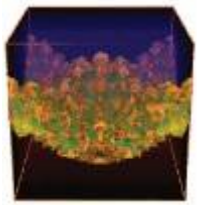
Ocean/wave Forecasting



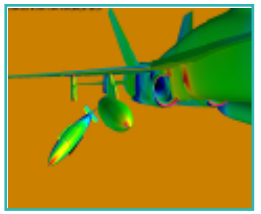
Ship Design



Climate Modeling



Nuclear Stockpile Stewardship



Weapons Integration

PERCS – Productive, Easy-to-use, Reliable Computing System is IBM's response to DARPA's HPCS Program

© 2007 IBM Corporation

IBM Hardware Innovations

- Next generation POWER processor with significant HPCS enhancements
  - Leveraged across IBM's server platforms
- Enhanced POWER Instruction Set Architecture
  - Significant extensions for HPCS
  - Leverage the existing POWER software eco-system
- Integrated high speed network (very low latency, high bandwidth)
- Multiple hardware innovations to enhance programmer productivity
- Balanced system design to enable extreme scalability
- Significant innovation in system packaging, footprint, power and cooling

© 2007 IBM Corporation

IBM Software Innovations

- **Productivity focus:** Tools for all aspects of human/system interaction; providing a 10x improvement in productivity
- **Operational Efficiency:** Advanced capabilities for resource management and scheduling including multi-cluster scheduling, checkpoint/restart, reservation in advance, backfill scheduling
- **Systems Management:** Significantly reduced complexity of managing petascale systems
  - Simplify the install, upgrade, and bring-up of large systems
  - Non-intrusive and efficient monitoring of the critical system components
  - Management framework for the networks, storage, and resources
  - OS level management: user ids, passwords, quotas, limits, …
- **Reliability, Availability and Serviceability:** Design for continuous operations

© 2007 IBM Corporation

IBM PERCS Software Innovations

- **Leverage proven software architecture and extend it to petascale systems**
  - **Robustness**: Design for continuous operation even in the event of multiple failures with minimal to no degradation
  - **Scaling**: All the required software will scale to tens of thousands of nodes while significantly reducing the memory footprint
  - **File System**: IBM Global Parallel File System (GPFS) will continue to drive toward unprecedented performance and scale.
  - **Application Enablement**: Significant innovation in protocols (LAPI), and programming models (MPI, OpenMP), new languages (X10), compilers (C, C++, FORTRAN, UPC), libraries (ESSL, PESSL) and tools for debugging (Rational$_R$, PTP) and performance tuning (HPCS toolkit)

© 2007 IBM Corporation

Blue Waters

- **National Science Foundation Track 1 Award**
- **Petascale Computing Environment for Science and Engineering**

- **Focus on Sustained Petascale Computing**
  - –Weather Modeling
  - –Biophysics
  - –Biochemistry
  - –Computer Science projects…
- **Technology: IBM POWER7 Architecture**
- **Location: NCSA**

© 2007 IBM Corporation

Roadrunner System Overview

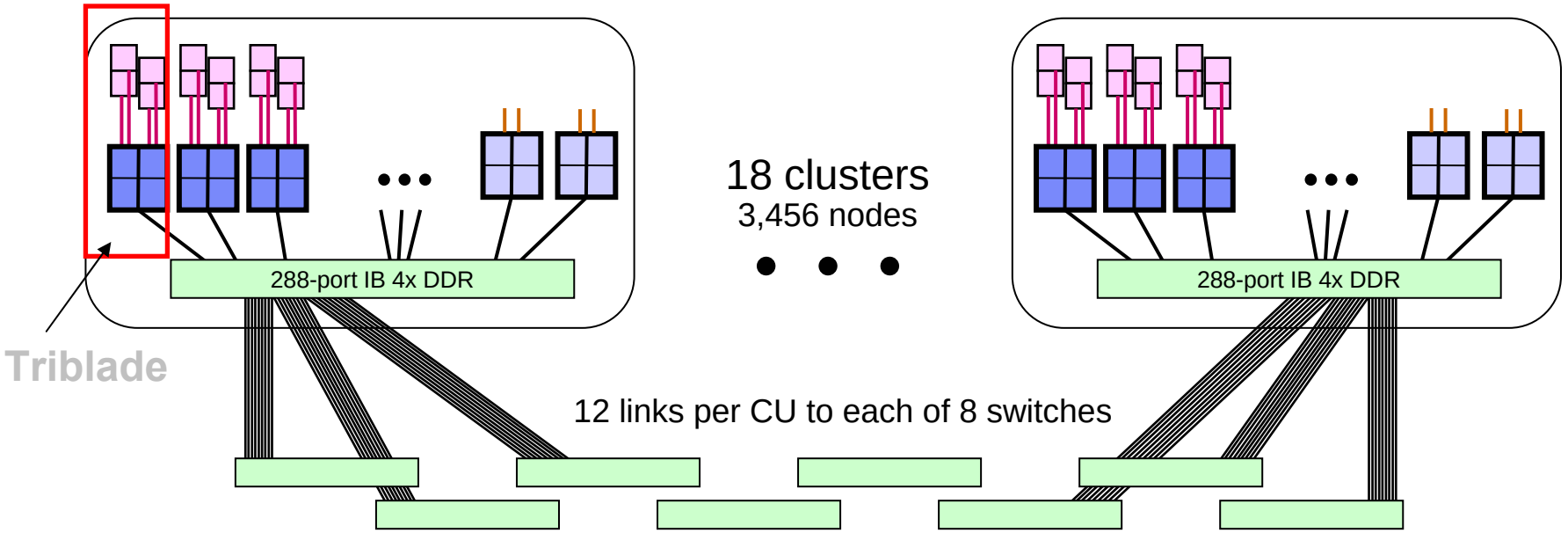
IBM Confidential

© 2007 IBM Corporation

Roadrunner is a hybrid cell accelerated petascale system delivered in 2008

Connected Unit cluster
180 compute nodes w/ Cells
12 I/O nodes

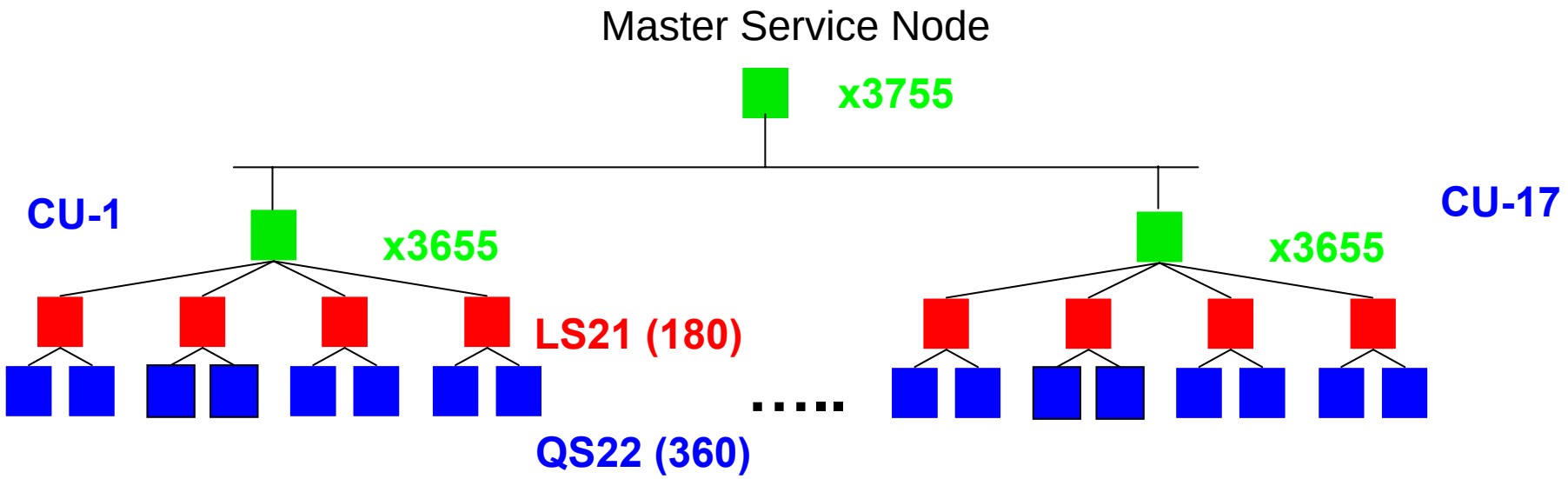
6,912 dual-core Opterons ⇒ 50 TF
12,960 Cell eDP chips ⇒ 1.3 PF



Eight 2nd-stage 288-port IB 4X DDR switches

© 2007 IBM Corporation

Roadrunner Organization



© 2007 IBM Corporation

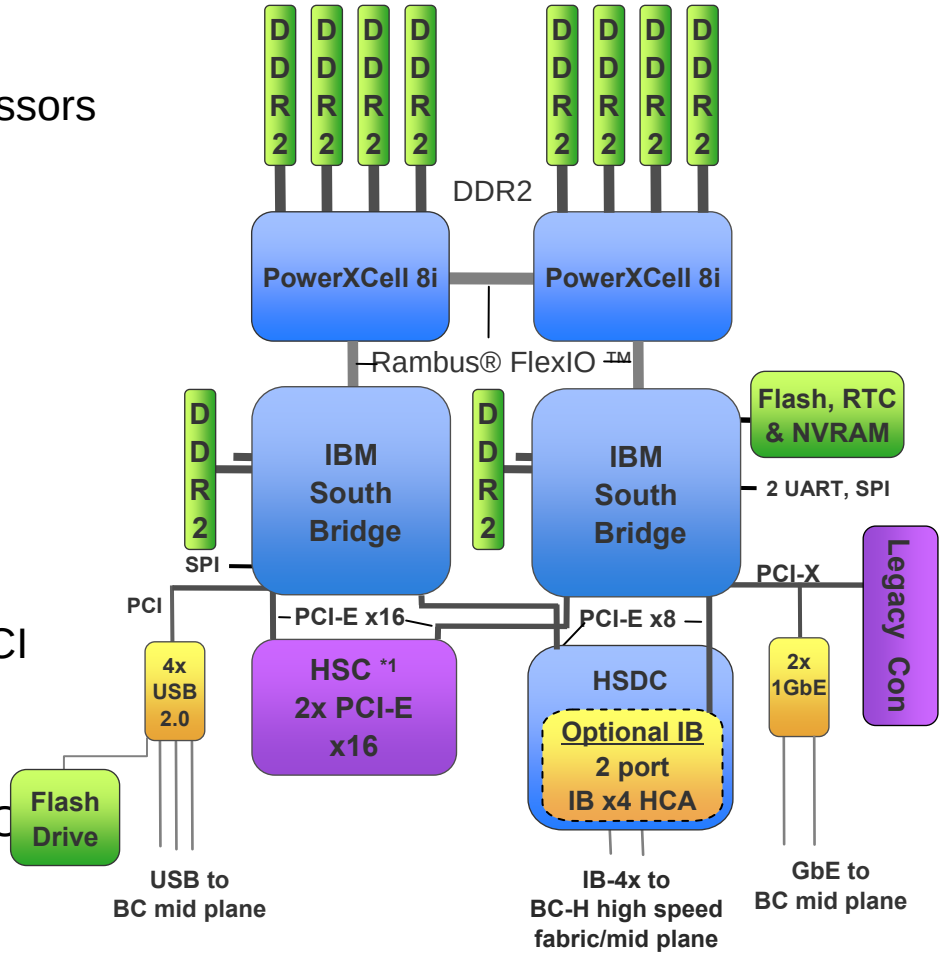
Roadrunner at a glance

- Cluster of 18 Connected Units
  - 12,960 IBM PowerXCell 8i accelerators
  - 6,912 AMD dual-core Opterons (comp)
  - 408 AMD dual-core Opterons (I/O)
  - 34 AMD dual-core Opterons (man)
  - 1.410 Petaflop/s peak (PowerXCell)
  - 46 Teraflop/s peak (Opteron-comp)
  - 1.105 Petaflop/s sustained Linpack
- InfiniBand 4x DDR fabric
  - 2-stage fat-tree; all-optical cables
  - Full bi-section BW within each CU
    - 384 GB/s
  - Half bi-section BW among CUs
    - 3.4 TB/s
  - Non-disruptive expansion to 24 CUs
- 104 TB memory
  - 52 TB Opteron
  - 52 TB Cell eDP
- 408 GB/s sustained File System I/O:
  - 204x2 10G Ethernets to Panasas
- RHEL & Fedora Linux
- SDK for Multicore Acceleration
- xCAT Cluster Management
  - System-wide GigEnet network
- 2.48 MW Power Linpack:
  - 0.445 MF/Watt
- Area:
  - 279 racks
  - 5200 ft^2
- Weight:
  - 500,000 lb
- IB Cables:
  - **55miles**

© 2007 IBM Corporation

BladeCenter® QS22 – PowerXCell 8i

- Core Electronics
  - Two 3.2GHz PowerXCell 8i Processors
  - SP: 460 GFlops peak per blade
  - DP: 217 GFlops peak per blade
  - Up to 32GB DDR2 800MHz
  - Standard blade form factor
  - Support BladeCenter H chassis
- Integrated features
  - Dual 1Gb Ethernet (BCM5704)
  - Serial/Console port, 4x USB on PCI
- Optional
  - Pair 1GB DDR2 VLP DIMMs as I/O buffer (2GB total) (46C0501)
  - 4x SDR InfiniBand adapter (32R1760)
  - SAS expansion card (39Y9190)

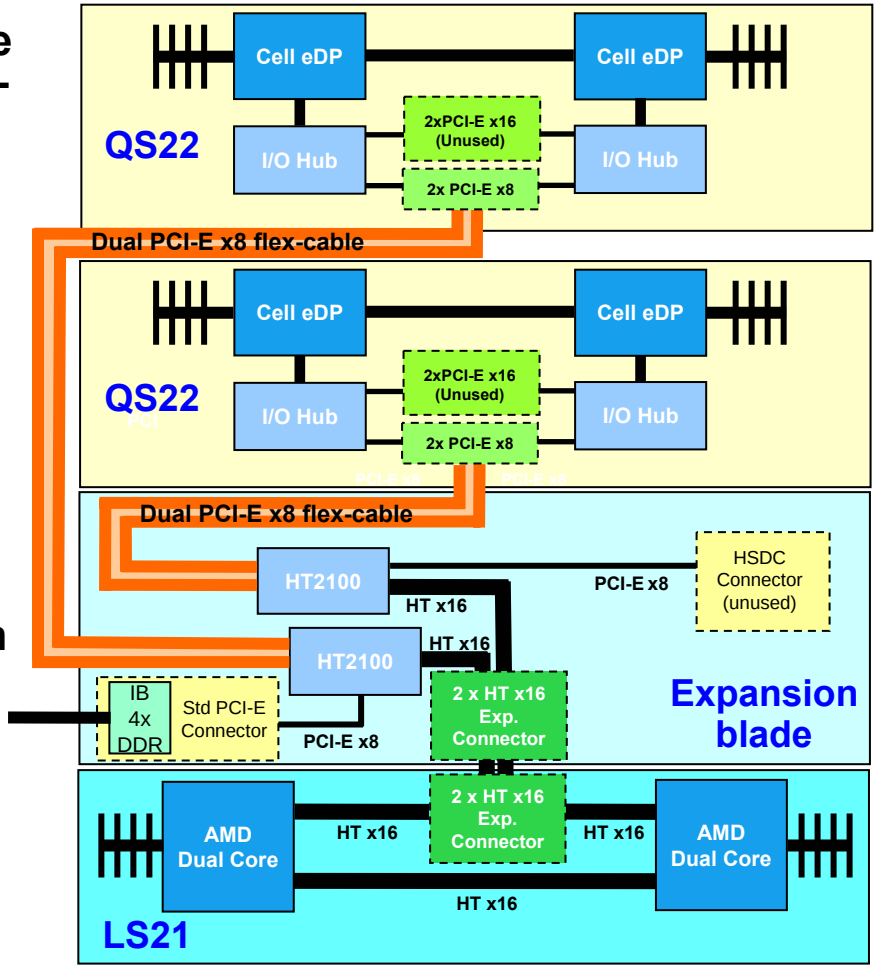


*The HSC interface is not enabled on the standard products. This interface can be enabled on "custom" system implementations for clients by working with the Cell services organization in IBM Industry Systems.

© 2007 IBM Corporation

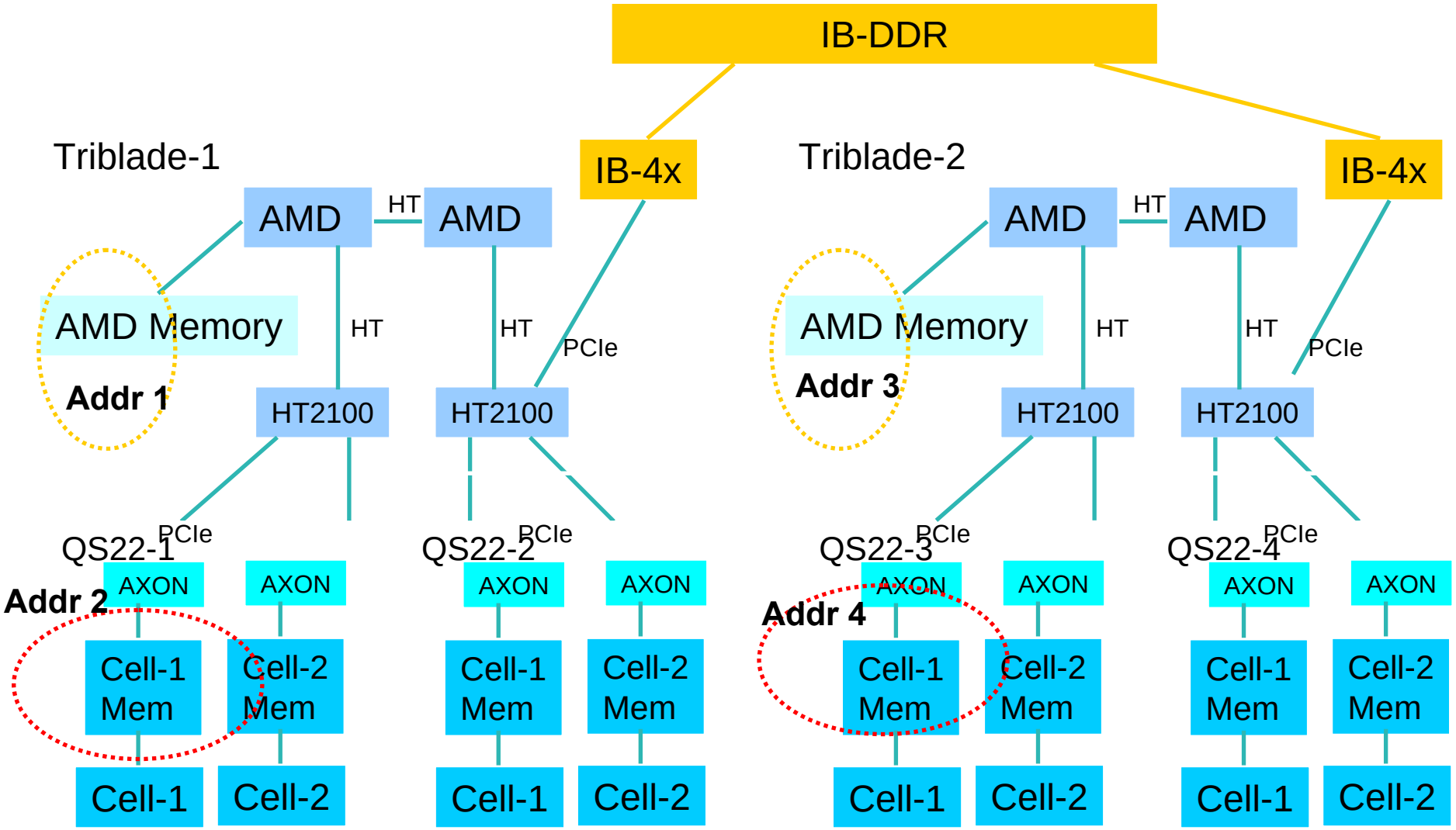
Roadrunner Triblade node integrates Cell and Opteron blades

- **QS22 is last generation IBM Cell blade containing two new enhanced double-precision (eDP/PowerXCell™) Cell chips**
- **Expansion blade connects two QS22 via four PCI-e x8 links to LS21 & provides the node's ConnectX IB 4X DDR cluster attachment**
- **LS21 is an IBM dual-socket Opteron blade**
- **4-wide IBM BladeCenter packaging**
- **Roadrunner Triblades are completely diskless and run from RAM disks with NFS & Panasas only to the LS21**
- **Node design points:**
  - One Cell chip per Opteron core
  - ~400 GF/s double-precision & ~800 GF/s single-precision
  - 16 GB Cell memory & 16 GB Opteron memory



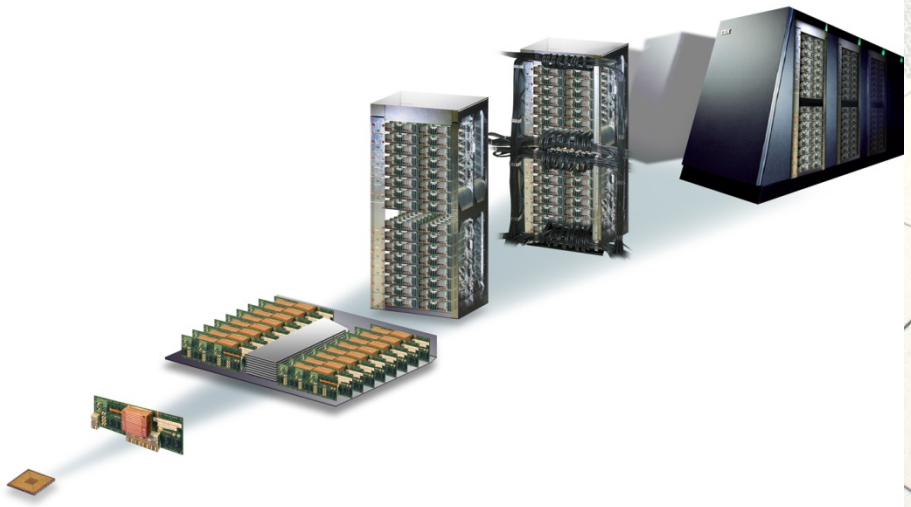
© 2007 IBM Corporation

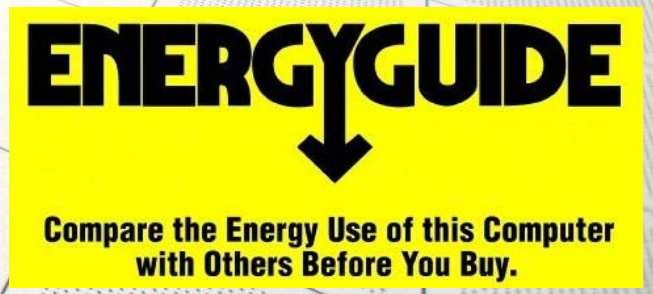
System Configuration



© 2007 IBM Corporation

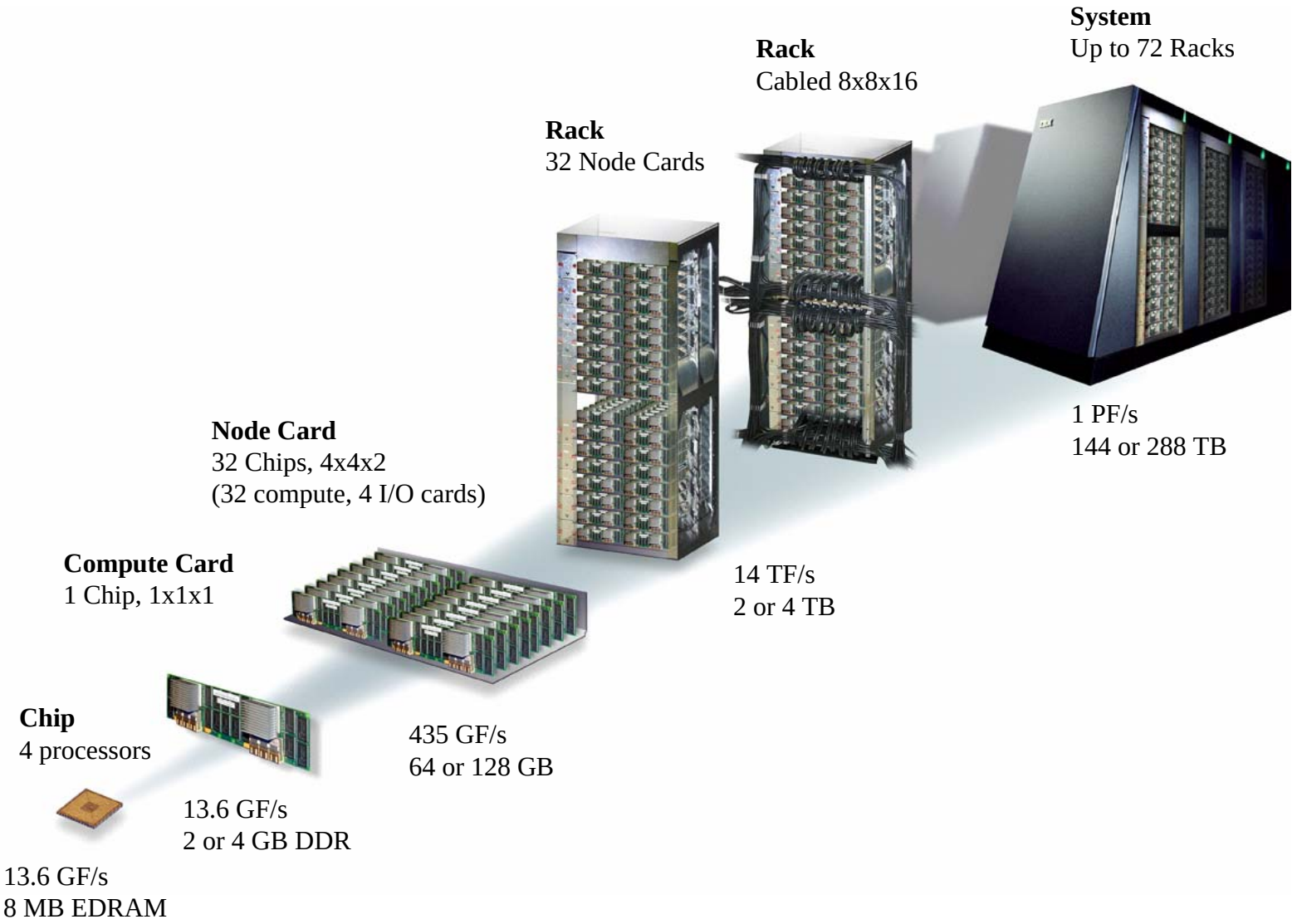
IBM Blue Gene





© 2007 IBM Corporation

BlueGene/P



© 2007 IBM Corporation

BGP vs BGL

Property		BG/L	BG/P
Node Properties	Node Processors	2* 440 PowerPC	4* 450 PowerPC
	Processor Frequency	0.7GHz	0.85GHz (target)
	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/1GB	2GB
	Main Store Bandwidth	5.6GB/s (16B wide)	13.9 GB/s (2*16B wide)
	Peak Performance	5.6GF/node	13.9 GF/node
Torus Network	Bandwidth	6*2*175MB/s=2.1GB/s	6*2*435MB/s=5.2GB/s
	Hardware Latency (Nearest Neighbor)	200ns (32B packet) 1.6us(256B packet)	160ns (32B packet) 1.3us(256B packet)
	Hardware Latency (Worst Case)	6.4us (64 hops)	5.5us(64 hops)
Collective Network	Bandwidth	2*350MB/s=700MB/s	2*0.87GB/s=1.74GB/s
	Hardware Latency (round trip worst case)	5.0us	4.5us
System Properties	Peak Performance 72k nodes	410TF	1PF
	Total Power	1.7MW	2.5-3.0MW

© 2007 IBM Corporation

Blue Gene/P Architectural Highlights

- **Scaled performance through density and frequency bump**
  - 2x performance from BlueGene/L through doubling the processors/node
  - 1.2x from frequency bump due to technology (target 850 MHz)
- **Enhanced function from BlueGene/L**
  - 4 way SMP
  - DMA, remote put-get, user programmable memory prefetch
  - Greatly enhanced 64 bit performance counters (including 450 core)
- **Hold BlueGene/L packaging as much as possible:**
  - Improve networks through higher speed signaling
  - Modest power increase through aggressive power management
- **Higher signaling rate**
  - 2.4x higher bandwidth, lower latency for Torus and Tree networks
  - 10x higher bandwidth for Ethernet IO
- **72ki nodes in 72 racks for 1.00 PF/s peak**
  - Could be as much as 750 TF/s on Linpack
  - Amazing ASC application performance of around 250 TF/s

© 2007 IBM Corporation

Sequoia Announcement

- On Feb 3, IBM announced a deal to sell a new supercomputer, SEQUOIA, to DOE
- 20 PFlops system in 2011
- 20 x faster than RoadRunner
- 10 times less energy per calculation than Jaguar
- 18 cores per chip, interconnect on the chip, 1.6 millions cores tota

© 2007 IBM Corporation

Programming Models

And

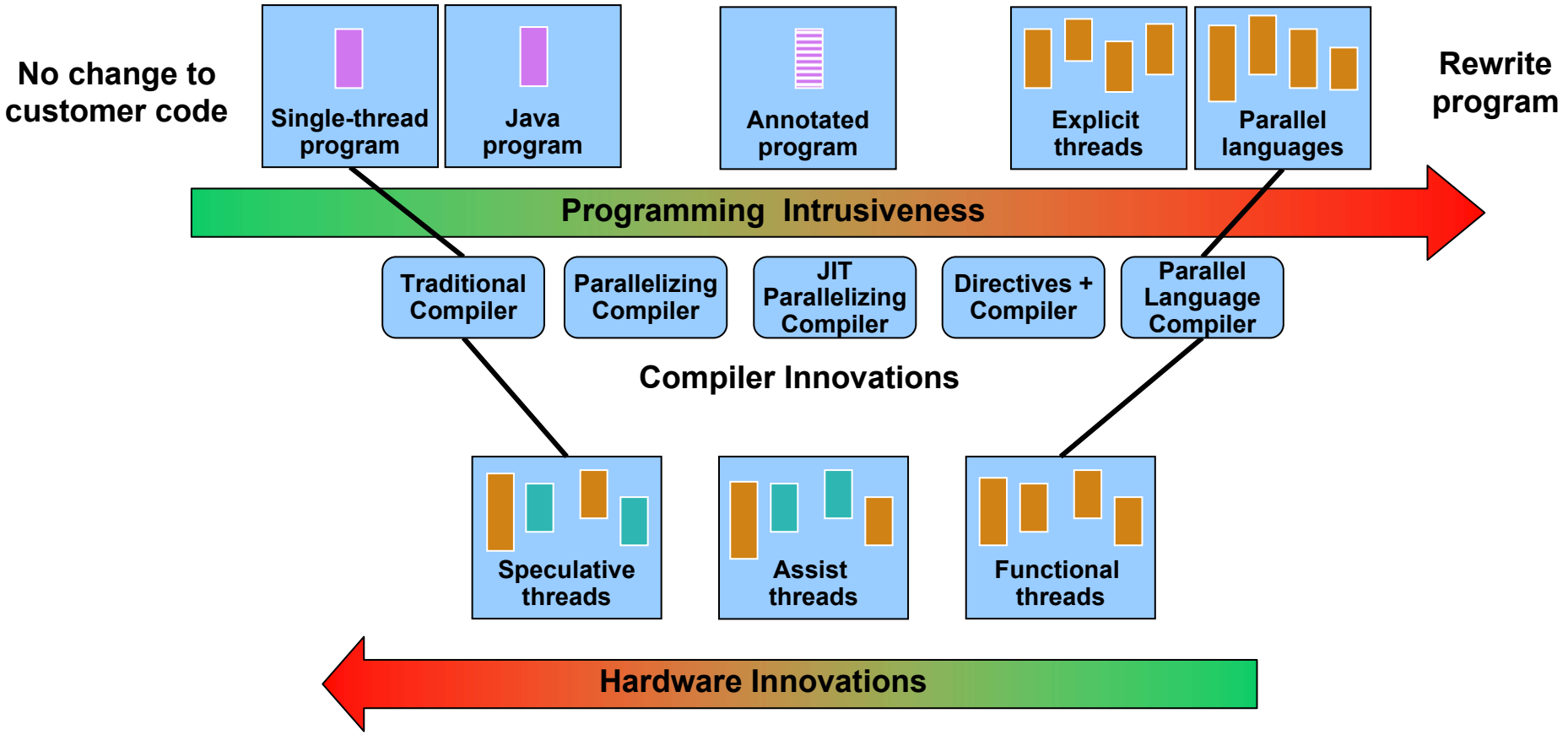
Languages Support for

Petascale Computing

© 2007 IBM Corporation

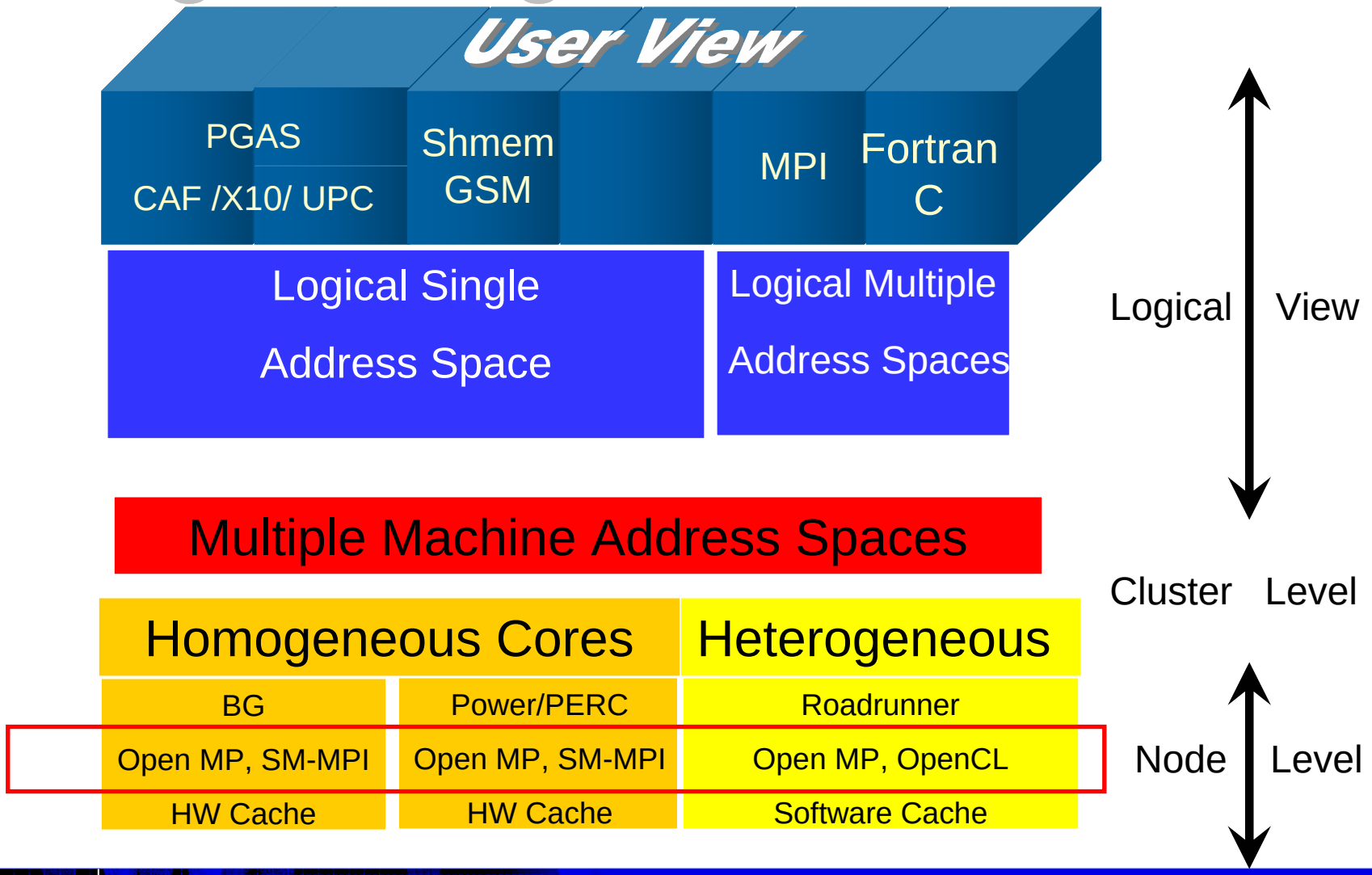
Different Approaches to Exploit Multi-Core Multi-Function Chips

Systems built around multi-core processor chips are driving the development of new techniques for automatic exploitation by applications



© 2007 IBM Corporation

Programming Models: Architecture

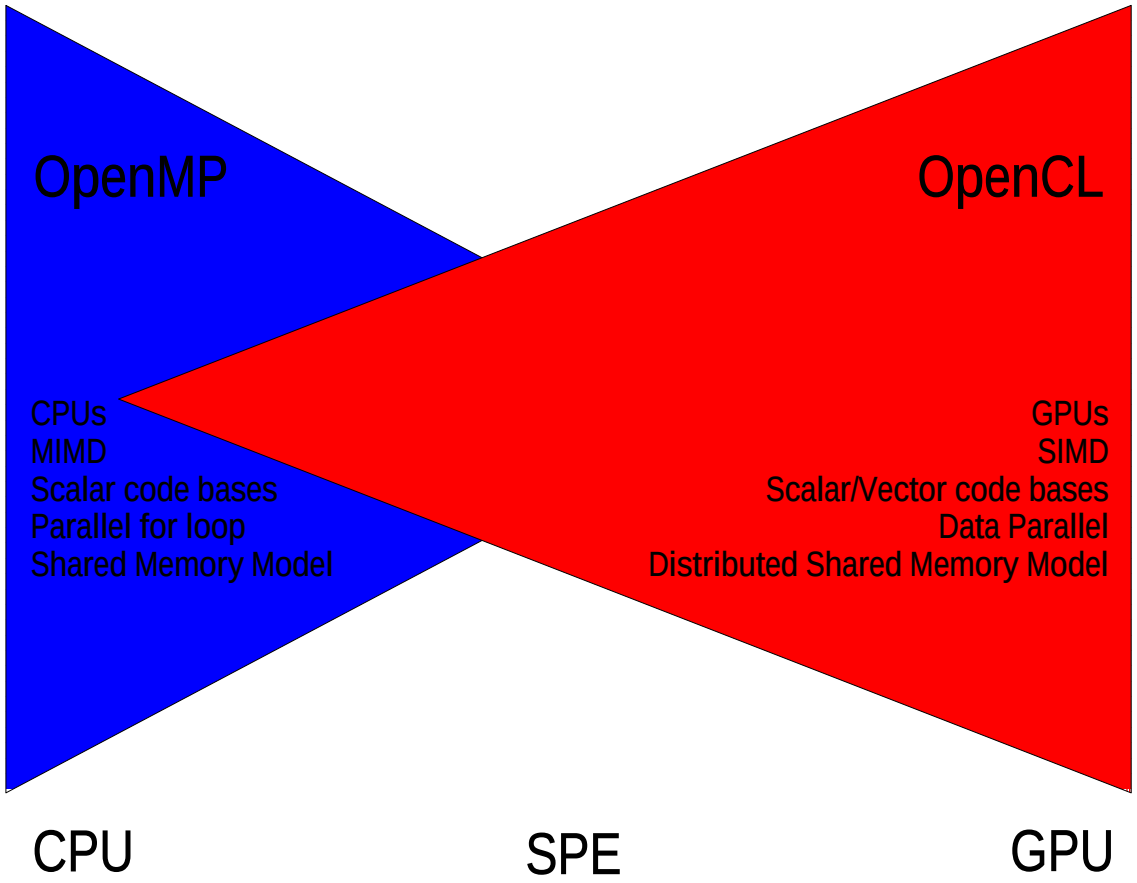


All statements regarding IBM future directions and intent are subject to change or withdrawal without notice and represent goals and objectives only.
Any reliance on these Statements of General Direction is at the relying party's sole risk and will not create liability or obligation for IBM.

© 2007 IBM Corporation

Two Standards

- **Two standards evolving from different sides of the market**



© 2007 IBM Corporation

OpenMP program computing PI

```
int i;
float x, Pi, sum;
float step = 1.0f / (float)NSET;

sum = 0.0f;

#pragma omp parallel for private(x) reduction(+:sum)
for( i = 0; i < NSET; i++ )
{
      x = (i+0.5)*step;
      sum = sum + 4.0/(1.0+x*x);
}
Pi = step * sum;
```

© 2007 IBM Corporation

CUDA: computing PI

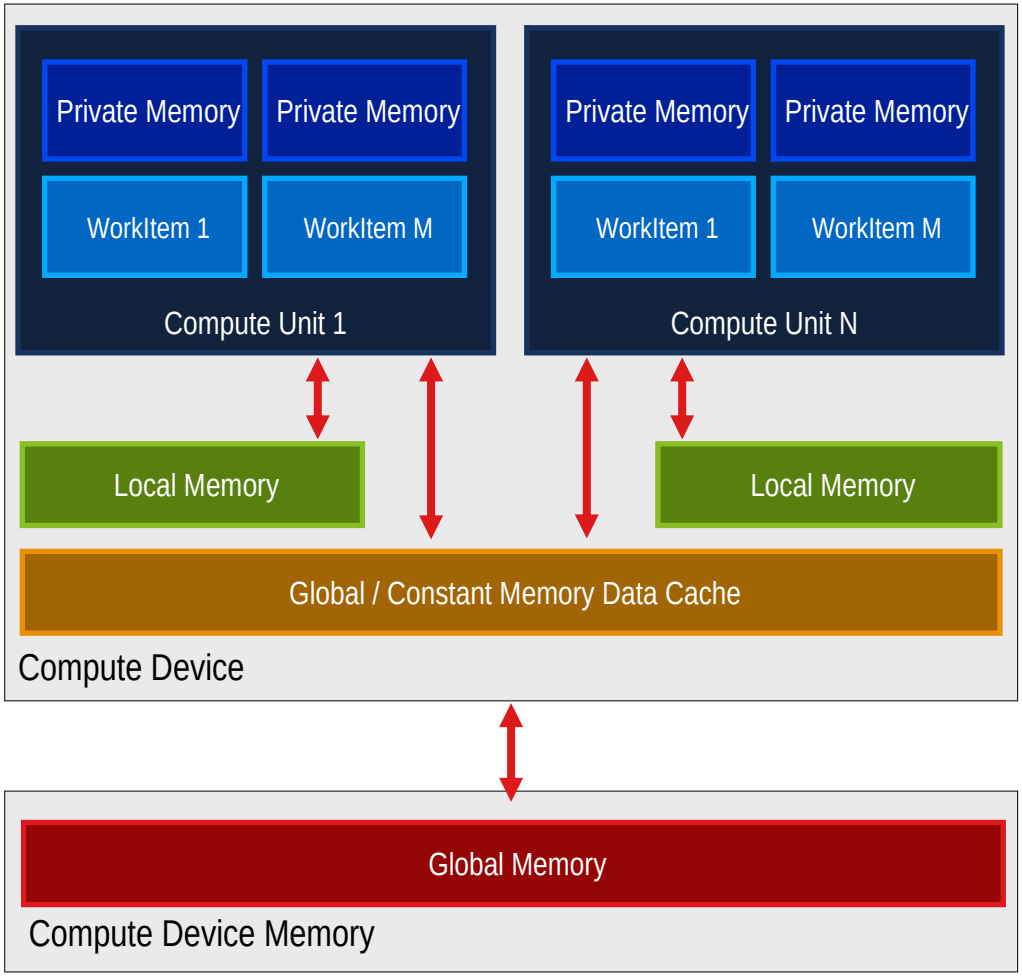
```
float step = 1.0f / (float)NSET;
float sum = 0.0f;

PiSimple2<<<GRIDDIM, BLOCKDIM>>>
    (d_partials, step, NSET);
CUT_CHECK_ERROR("***PiSimple2
    execution failed!!!***");



CUDA_SAFE_CALL(cudaMemcpy(h_partia
    ls, d_partials,
    fSmallArraySize,
    cudaMemcpyDeviceToHost) );

for (j = 0; j < GRIDDIM; j++)
{
    sum += h_partials[j];
}
Pi = step * sum;
```

```
__global__ void
PiSimple2( float* g_partialOut, float step,
    int NSamples)
{
  const int tid = blockDim.x * blockIdx.x +
   threadIdx.x;
  const int blocksize = blockDim.x;
  const int THREAD_N = blockDim.x * gridDim.x;
  float x, partialsum = 0.0f;

for(int i = tid; i < NSamples; i += THREAD_N){
     x = (i * 0.5f)*step;
     partialsum = partialsum + 4.0f / (1.0f
                  + x*x);
  }

 __shared__  float threadsum[BLOCKDIM];
 threadsum[threadIdx.x] = partialsum;

  __syncthreads();
  float blocksum = 0;
  if (threadIdx.x == 0)   {
    const int blockindex = blockIdx.x;
    for (int i = 0; i < blocksize; i++)
        blocksum += threadsum[i];
    g_partialOut[blockindex] = blocksum;
  }
}
```

© 2007 IBM Corporation

OpenCL Memory Model

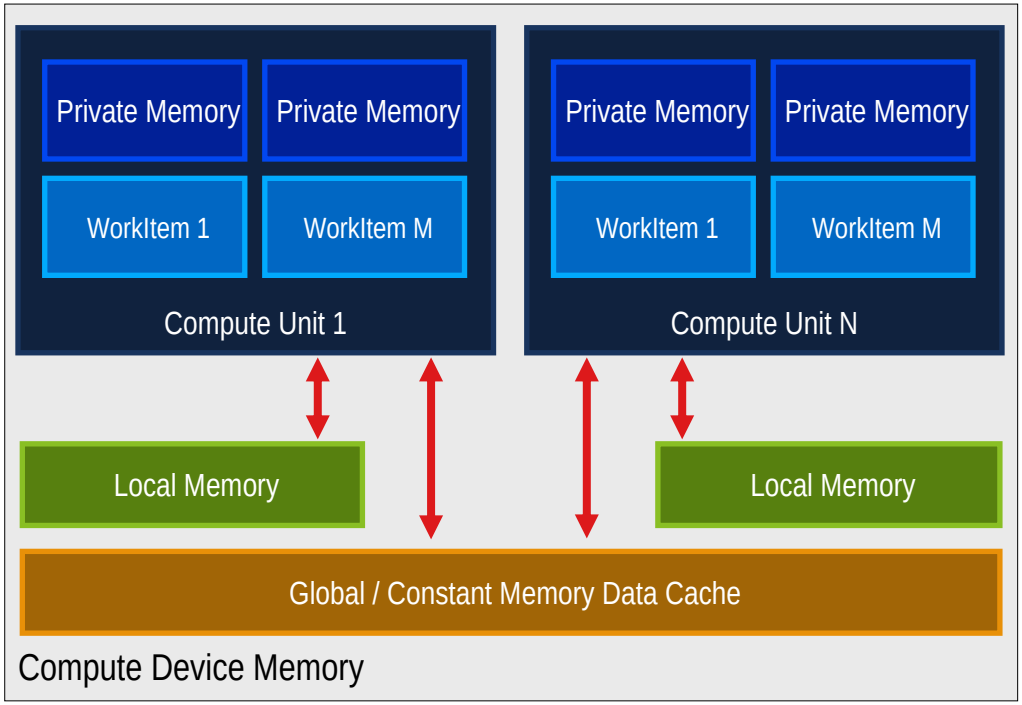
- **Shared memory model**
  - Release consistency
- **Multiple distinct address spaces**
  - Address spaces can be collapsed depending on the device's memory subsystem
- **Address Qualifiers**
  - __private
  - __local
  - __constant and __global
- **Example:**
  - __global float4 *p;



© 2007 IBM Corporation

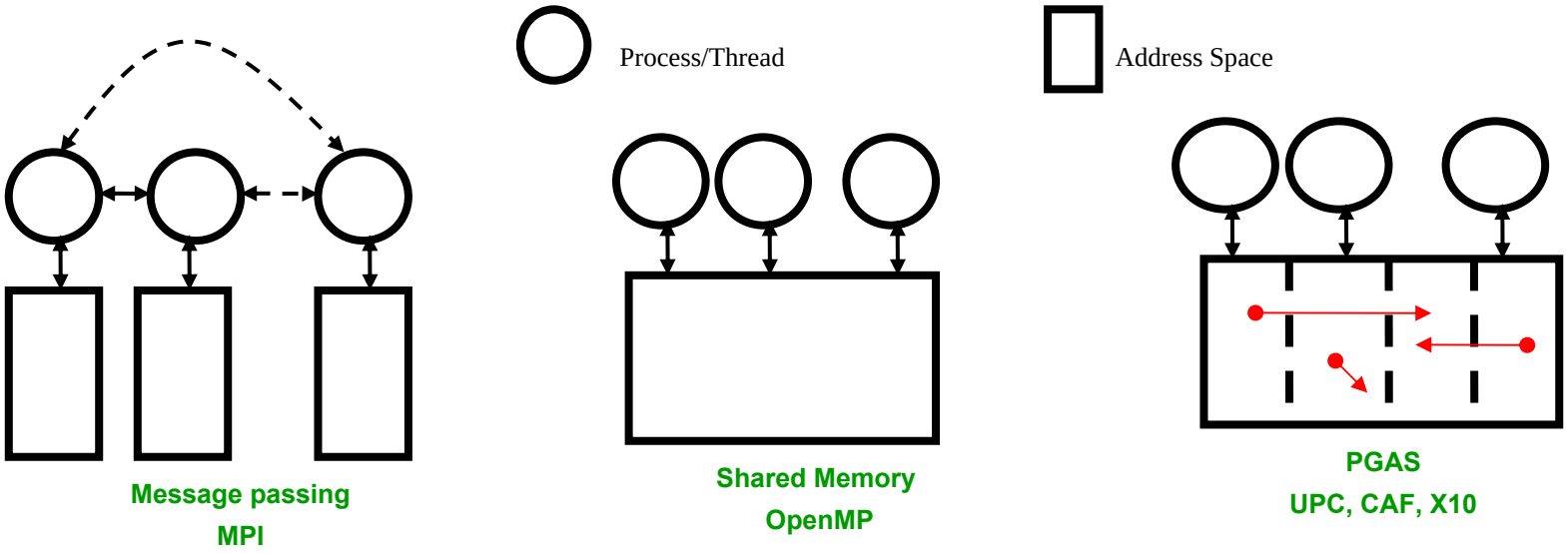
OpenMP Memory Model

- **Shared memory model**
  - Strong consistency
- **Single address space**
  - Uniform address space
- **Pragmas**
  - V2.5 : omp parallel
  - V3.0 : omp tasks
- **Example:**
  - #pragma OMP parallel



© 2007 IBM Corporation

What is Partitioned Global Address Space (PGAS)?

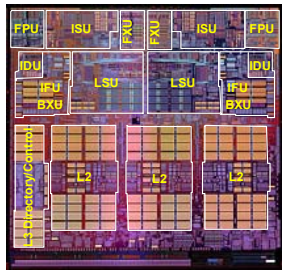


- Computation is performed in multiple places.
- A place contains data that can be operated on remotely.
- Data lives in the place it was created, for its lifetime.
- A datum in one place may reference a datum in another place.
- Data-structures (e.g. arrays) may be distributed across many places.
- Places may have different computational properties

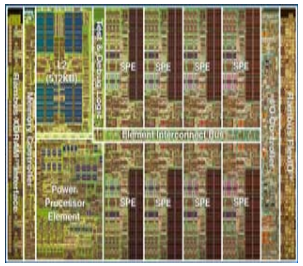
© 2007 IBM Corporation

Current Language Landscape

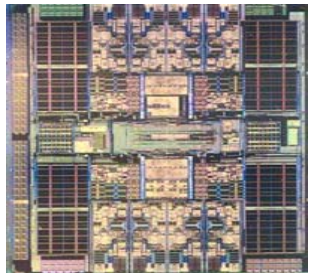
OpenMP 1997	MPI 1994	PGAS (199x) CAF/UPC/Titanium	Scripting	Domain specific
Fortran 1954	C/C++ 1973/1983	Java 1994		



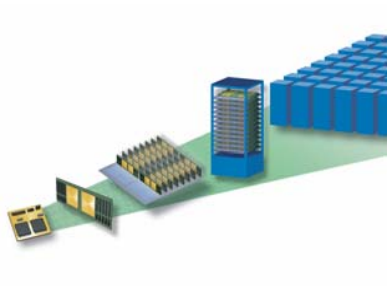
Power 4



Cell BE


Niagara

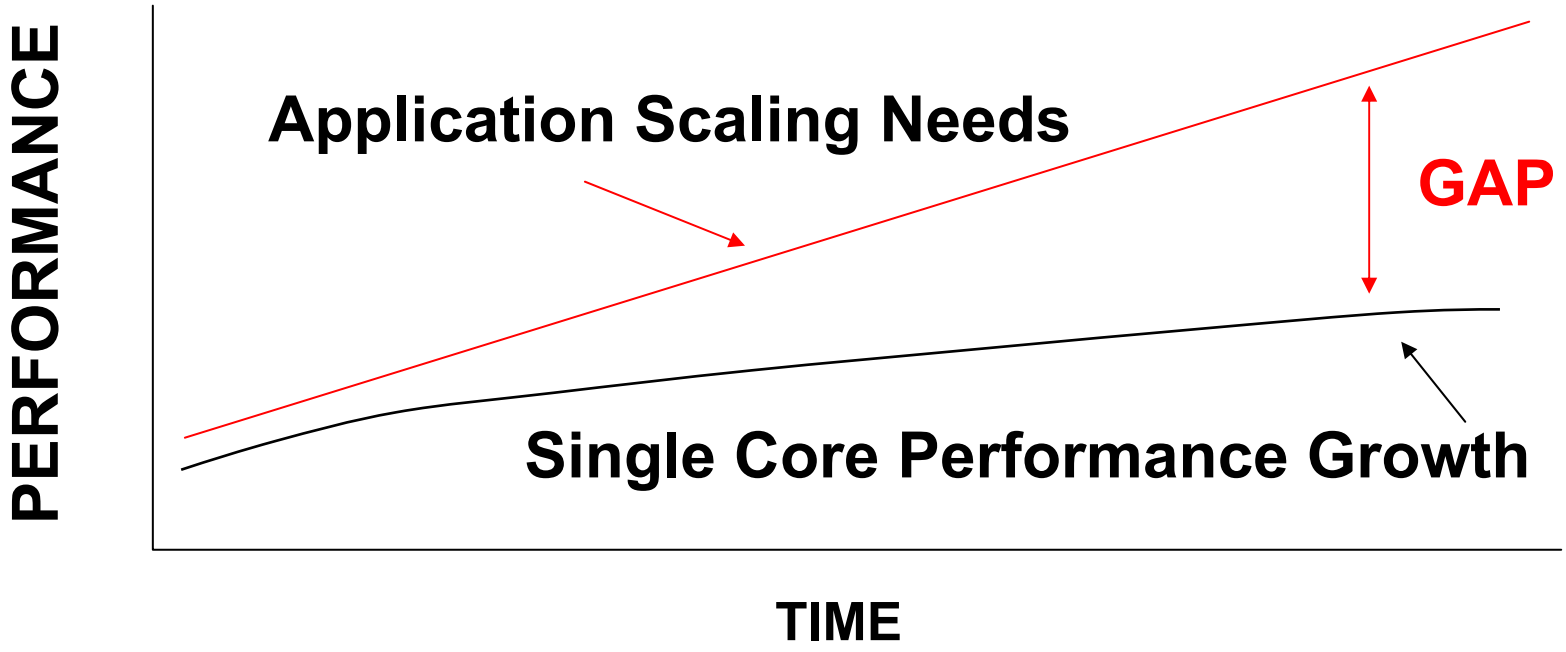

Clusters


BlueGene/L

© 2007 IBM Corporation

Technology Trends and Challenges

Programmer Productivity



Key Problem: Frequency Improvements Do Not Match App Needs

Increasing Burden On The Application Design

Objective: Provide Tools to allow Scientists to Bridge the Gap

© 2007 IBM Corporation

Next stop, exaflop?

www.lanl.gov/roadrunner

www.lanl.gov/news

www-03.ibm.com/press/us/en/pressrelease/24405.wss

www.ibm.com/deepcomputing

© 2007 IBM Corporation

IBM Ultra Scale Approaches

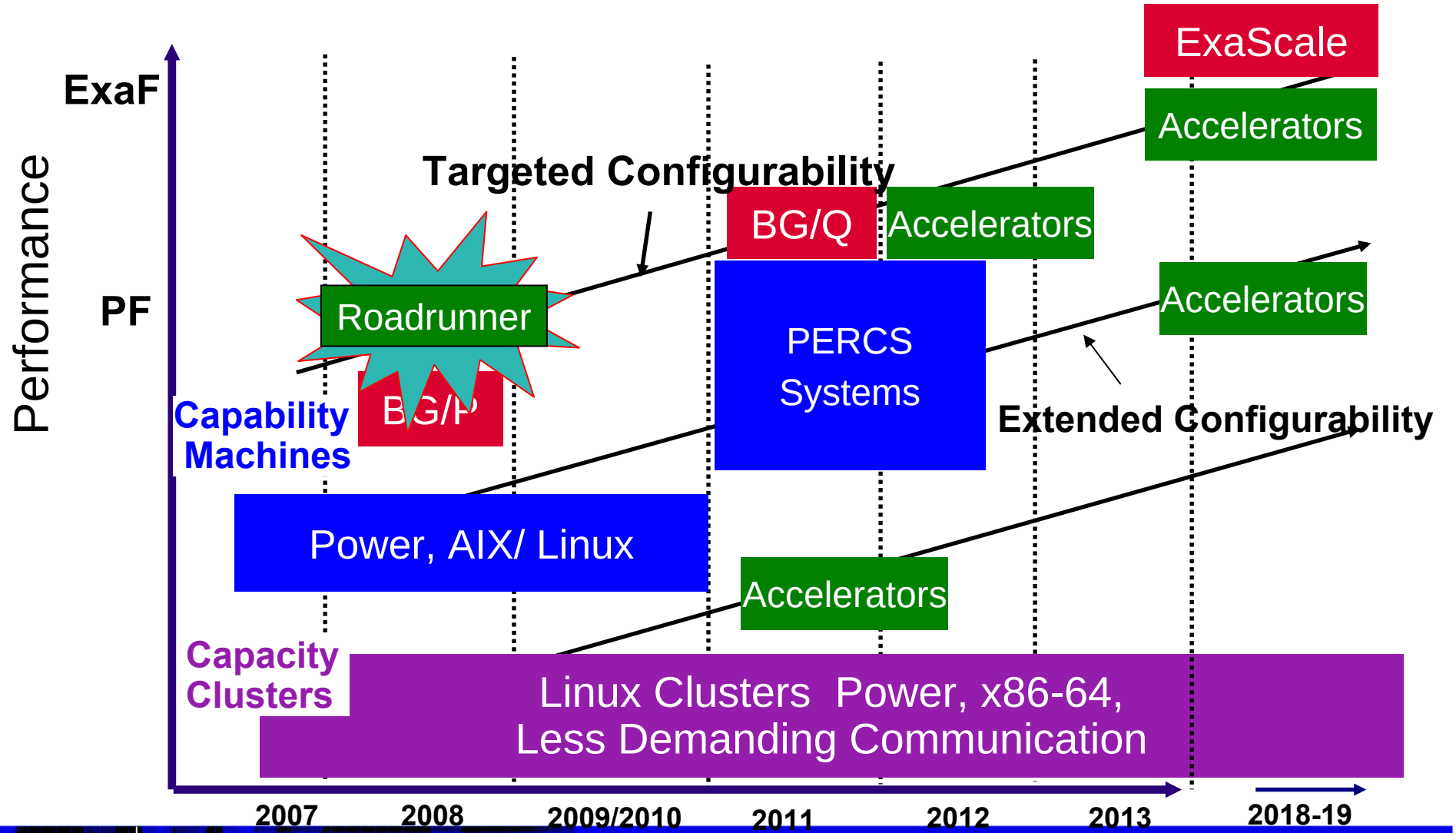
-
- Blue Gene – Maximize Flops Per Watt
  with Homogeneous Cores
  by reducing Single Thread Performance

- Power/PERCS – Maximize Productivity and Single Thread
  Performance with Homogeneous Cores

- Roadrunner – Use Heterogeneous Cores and
  an Accelerator Software Model
  to Maximize Flops Per Watt
  and keep High Single Thread Performance

© 2007 IBM Corporation

HPC Cluster Directions



© 2007 IBM Corporation