ECMWF HPC Workshop 10/26/2004

IBM's High Performance Computing Strategy

Dr Don Grice
Distinguished Engineer, HPC Solutions

HPC Key to an Innovation Economy

Digital Media:

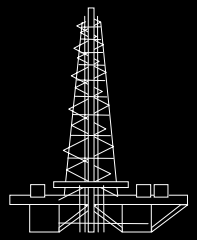
Increasing productivity for digital content creation and online gaming

Life Sciences:

pharmaceuticals and biotechs accelerating drug discovery and diagnostics

Petroleum:

accelerating rate of oil exploration and production



Industrial Sector:

accelerating CAE for Electronics, Automotive, and Aerospace

Financial Services:

optimizing IT infrastructure, risk analysis, portfolio management, and compliance

Government & Higher Ed.:

making scientific research more affordable

IBM's High Performance Computing Strategy

Solving Problems More Quickly at Lower Cost

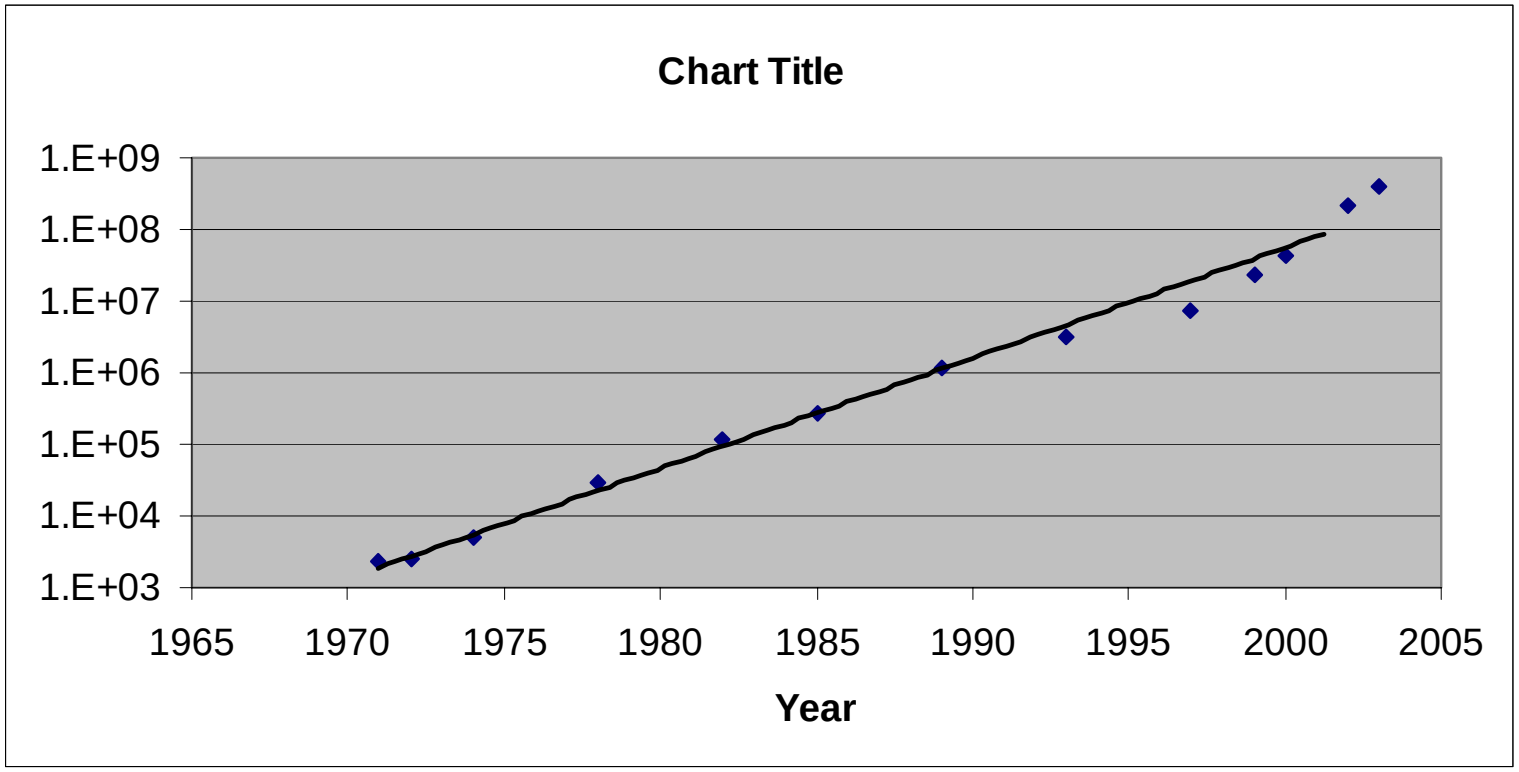
- Aggressively evolve the POWER-based Deep Computing product line
- Develop advanced systems based on loosely coupled clusters
- Deliver supercomputing capability with new access models and financial flexibility
- Research and overcome obstacles to parallelism and other revolutionary approaches to supercomputing

Technology Scaling; A Legacy Strategy

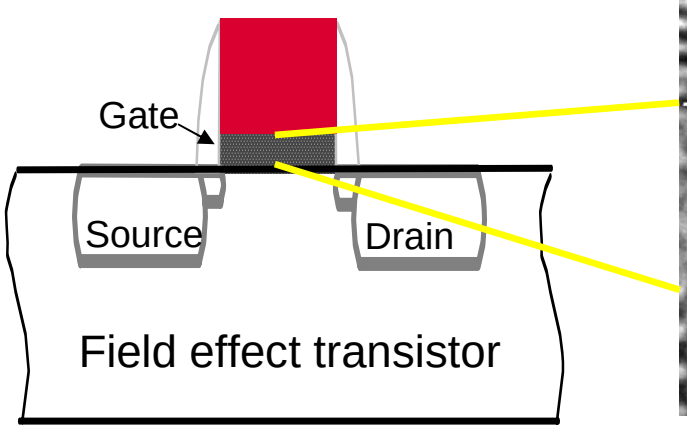
Moore's Law is a VERY simplistic subset of Semiconductor Scaling

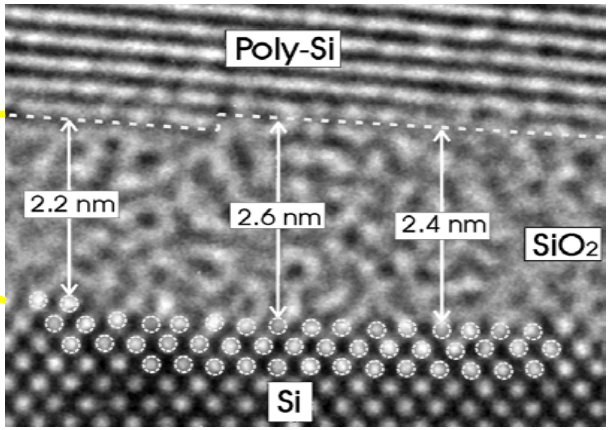
The Basic's of Moore's Law

The number of Devices on a chip of fixed size doubles every 12 to 18 months – This is accomplished by the scaling of technology

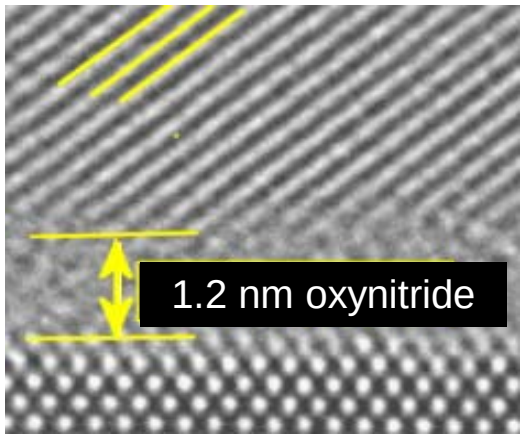


Why Scaling Breaks Down; We're down to atoms





"Thick" gate oxide

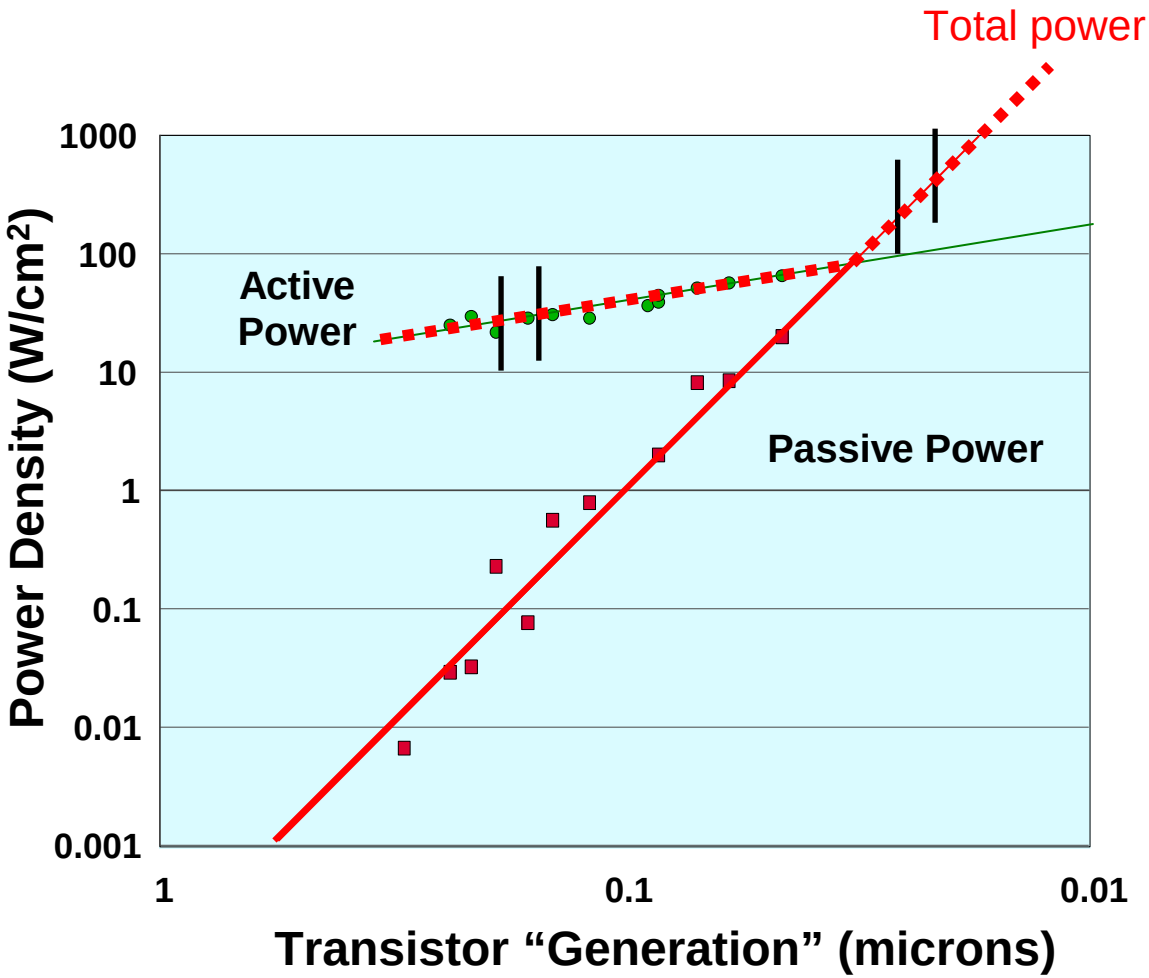


Scaled gate oxide

- Consider the gate oxide in a CMOS transistor (the smallest dimensions today)
  - Assume only 1 atom high "defects" on each surrounding silicon layer
    - For a modern "scaled" oxide, 6 atoms thick, 33% variability is induced.
  - The bad news
    - Single atom defects can cause local current leakage 10-100x higher than average
  - The really bad news
    - Such "non-statistical behaviors" are appearing elsewhere in technology

Consider the Issue of Chip Power

- **Fundamental Changes**
  - **“Stopping” the chip no longer reduces chip power.**
  - **One must develop means to literally “unplug” unused circuits.**
  - **Software must become much more sophisticated to cope with selective shutdowns of processor assets.**
  - **Scaling produces profoundly different results when attempting to “push“ chip speeds**



What Constitutes Future Processor *Technology*?

- **Circa 2004-2044**
  - **Materials, Devices, Scalable and Integrable Cores (IP), System Architecture, System Integration, and System Software**
    - **The word "Technology" now encompasses far more than just semiconductors going into the future**
    - **Integration, the creation of systems rather than just "chips", will become the means by which past trajectories for computing performance are maintained**

Innovation via Holistic Design, from Atoms to Software

- **Only the simultaneous optimization of materials, devices, circuits, cores, chips, system architecture, and system software, provides an effective means to optimize for both performance and power.**

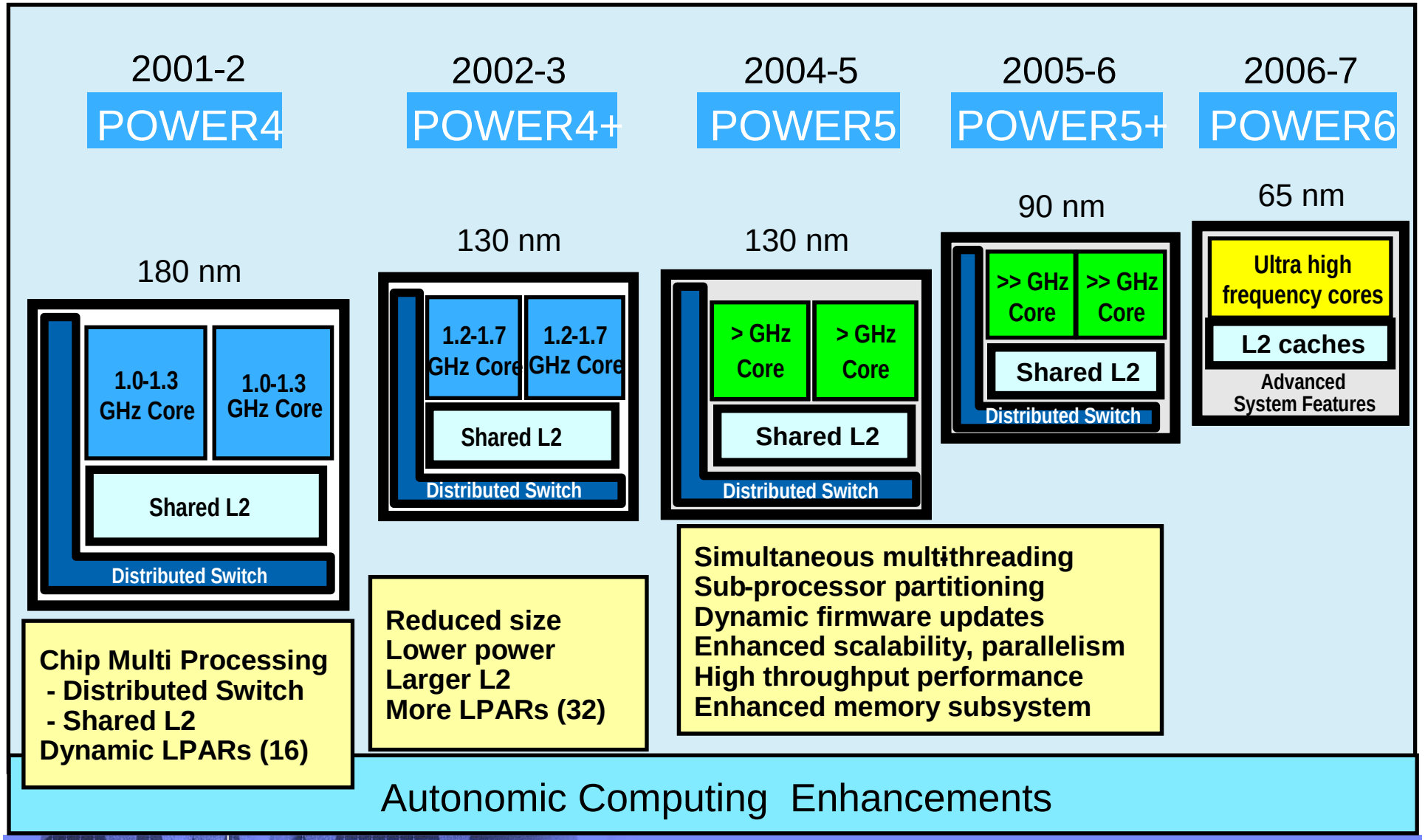
- **IBM's Power Architecture is taking a major step towards creating an open ecosystem of highly scalable Multi-Core chips having power control and performance characteristics required for future Processor Technology.**
  - **Asset virtualization**
  - **Fine grained clock gating**
  - **Dynamically optimized multi-threading capability**
  - **Open (accessible) architecture for system optimization/compatibility**
  - **Scalability enabling IP re-use in a broad range of systems and products**

Multi-Core Options

- Homogeneous Symmetric Multi-Core General Purpose CPUs
- Homogeneous Symmetric Multi-Cores with Specialized Instructions
- Heterogeneous Cores with Specialized Accelerators

Science Driven Design

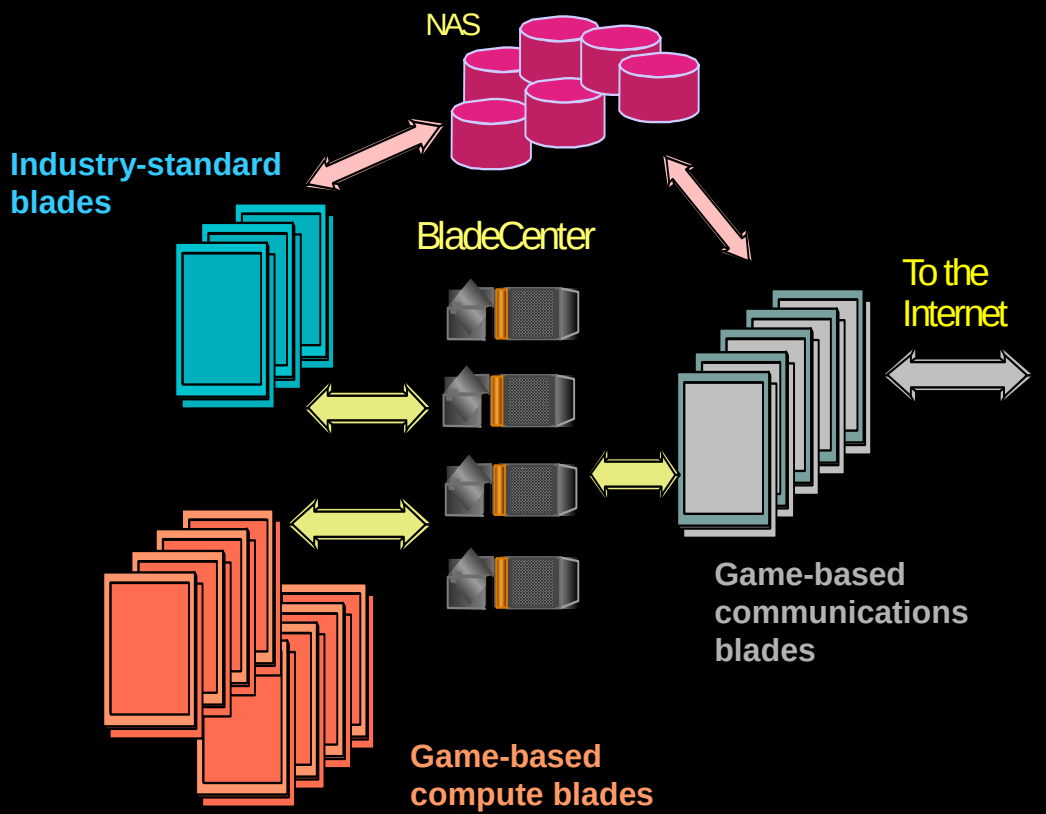
Multiple Core Processors and Optimizations



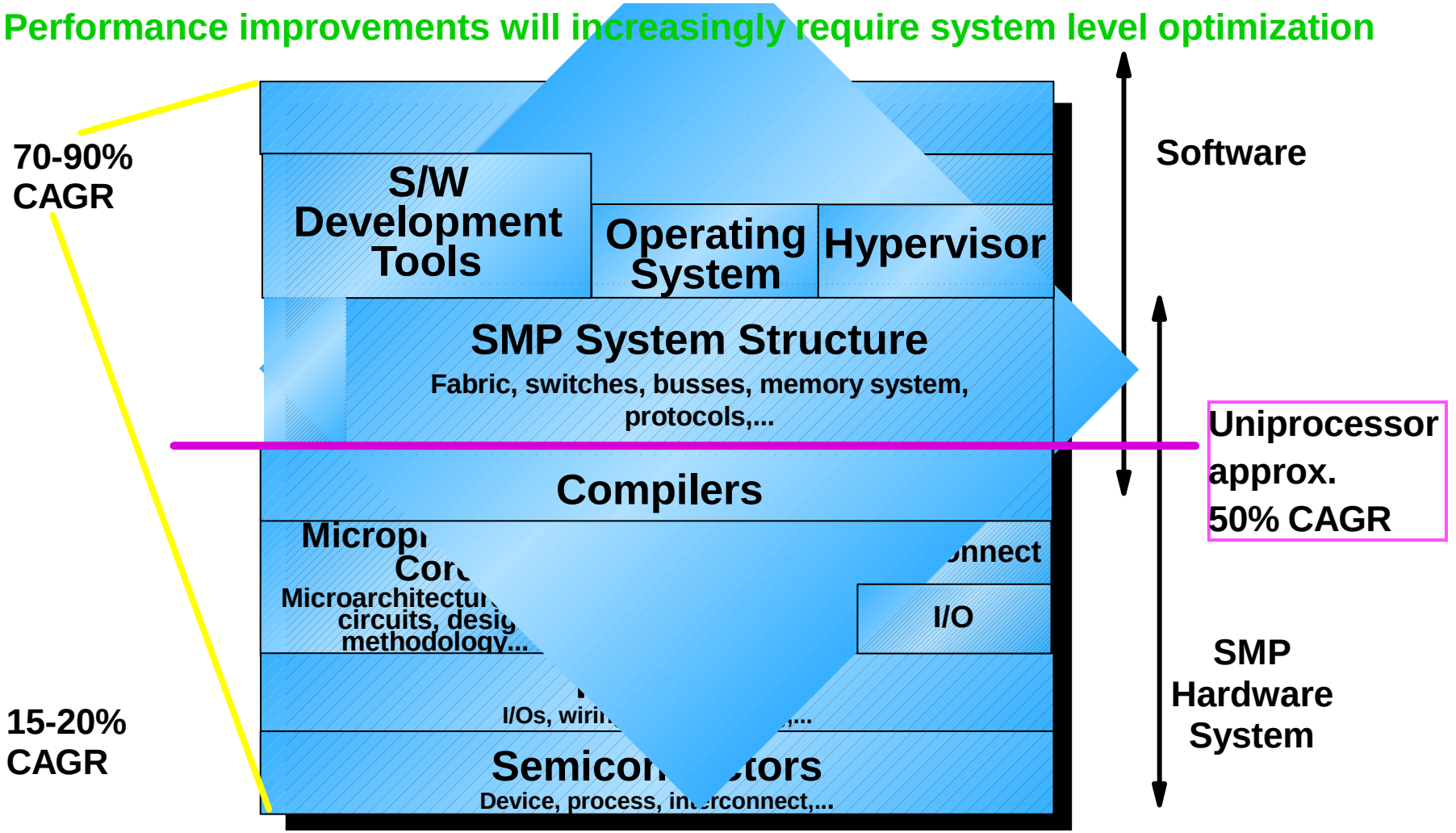
Game Processor Technologies

- The game console market is driving the development of microprocessors with high numeric processing capabilities, high bandwidths, and features to tolerate memory latency

- Research is exploring the use of game-processor technologies to boost performance in areas including:
  - High-performance scientific computing
  - On-line client-server games
  - Video applications including surveillance
  - Secure communications acceleration

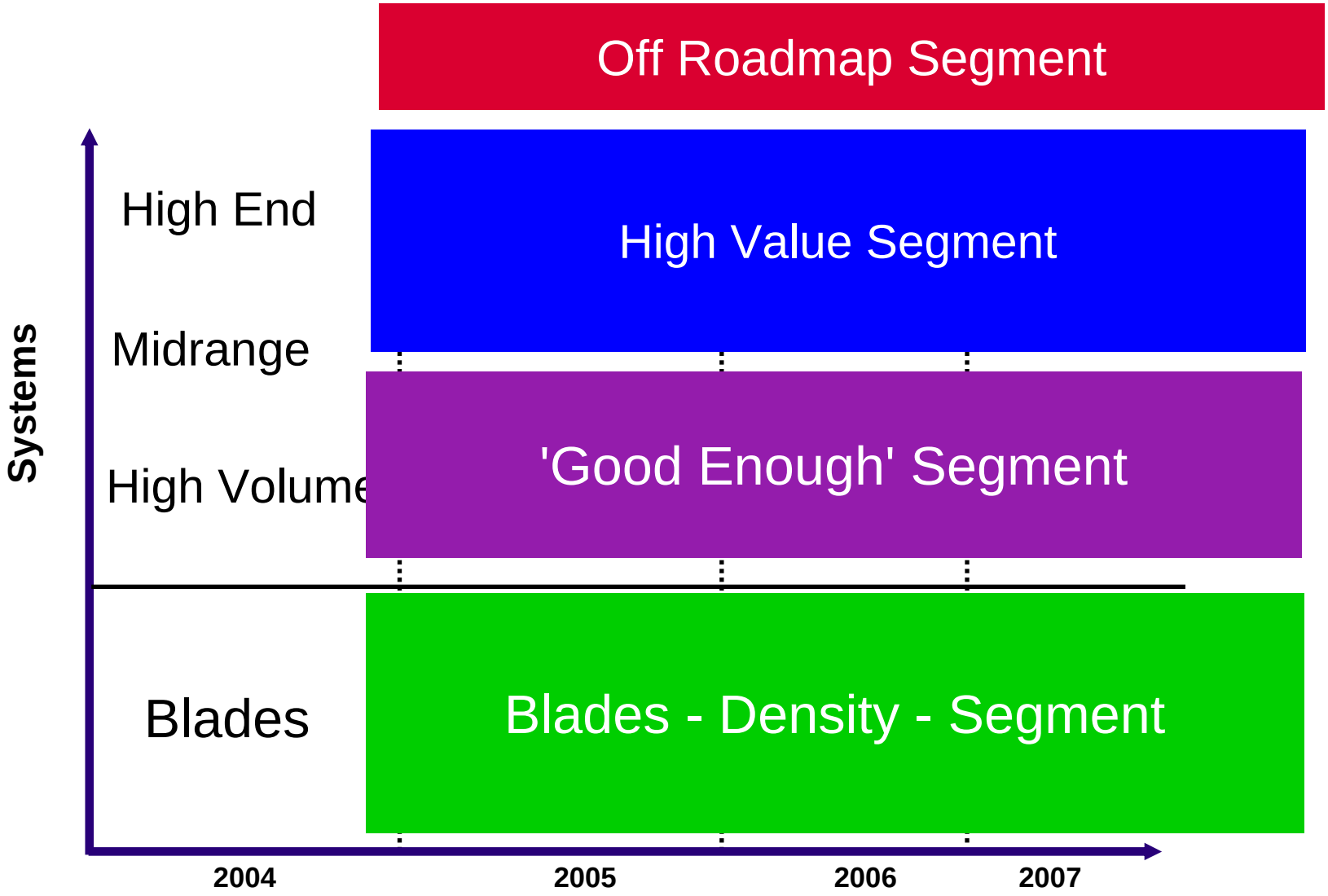


System performance gains of 70-90% CAGR derive from far more than semiconductor technology alone

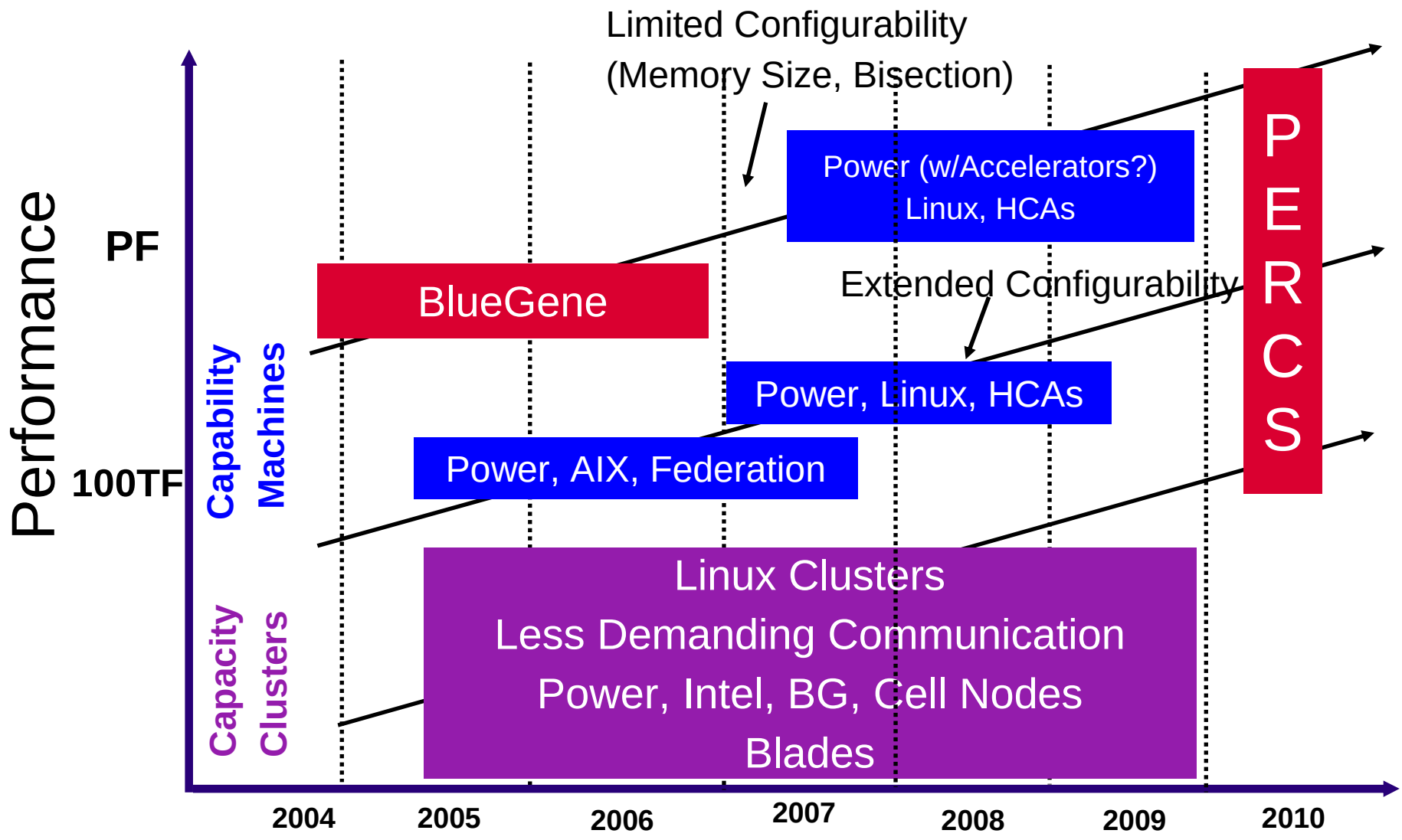


HPC Cluster System Direction

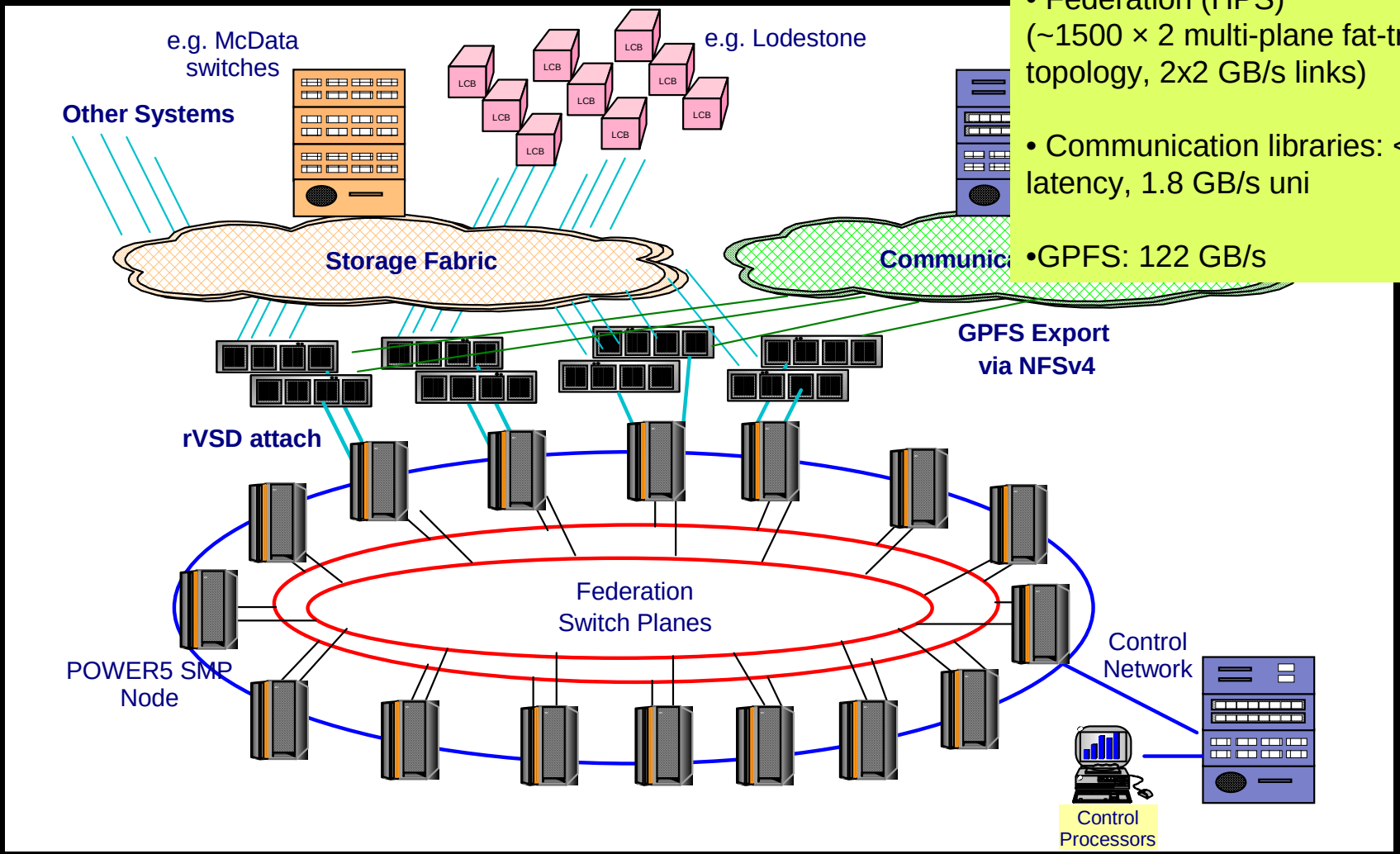
Segmentation Based on Implementation



HPC Cluster Directions



ASCI Purple Architecture



- 100TF Machine

- ~1500 8-way Power5 Nodes
- Federation (HPS)
(~1500 × 2 multi-plane fat-tree topology, 2x2 GB/s links)

- Communication libraries: < 5 µs latency, 1.8 GB/s uni

- GPFS: 122 GB/s

Interconnect Adapter Types

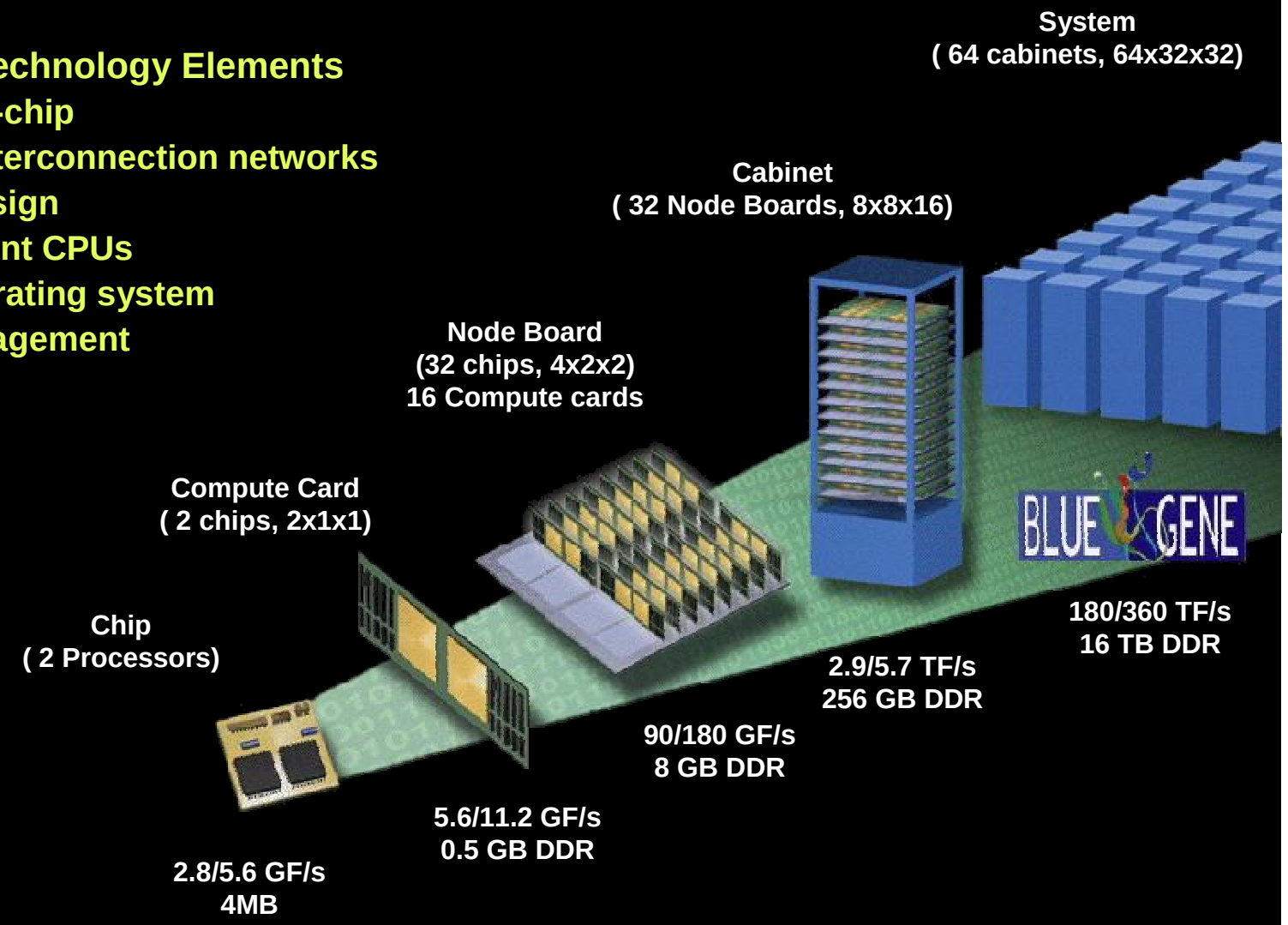
- **Adapter (HCA) Server Attachment Method**
  - **Internal 'Proprietary' Bus Attachment**
    - **Optimizing Performance for the Server**
  - **Open/Multi-Vendor Slot Attachment**
    - **Facilitates Heterogeneous System Solutions**
- **Interconnect Fabric Type**
  - **'Proprietary' Protocol and/or Network**
    - **Value Add over current Industry Standards**
  - **Industry Standard Carrier and APIs**

Interconnect Type Evolution

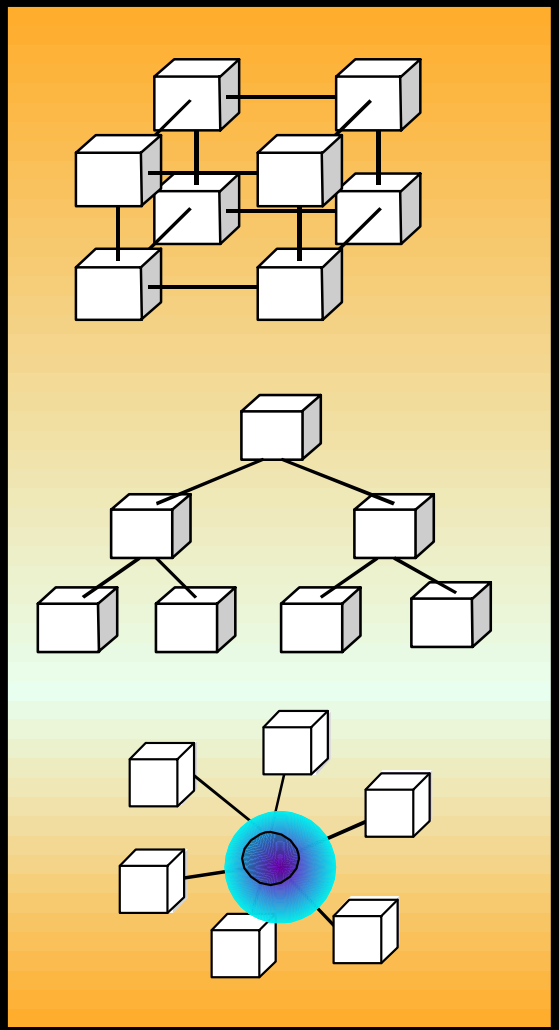
- P-series
  - High End - Focus on advancing Standard Networks Following HPS (Federation)
    - IB-DDR/QDR
    - Low Latency (User Space) Ethernet
    - Collective Offload/Acceleration
    - Combination of Internal Attachment Point and Open Slot HCAs
  - Standard Networks
    - IP: Ethernet; 1Gb, 10Gb, 40-100Gb, Lower Latency
    - MPI: Myrinet, Myrinet-10G, (IB and LL-Ethernet as they evolve)
    - Parallel DB and ‘commercial’: IB4x/12x
  - Work with Industry to extend PCI-Express to DDR (QDR?)
    - To match Internal I/O Bus capability
- Other Series Deep Computing
  - Use Standard Networks and Standard Attachment points
- Research
  - Continue to look at new ways to use networks especially for large Scale-out Solutions

BlueGene/L

- **Key Blue Gene Technology Elements**
  - **System-on-a-chip**
  - **Integrated interconnection networks**
  - **Balanced design**
  - **Power-efficient CPUs**
  - **Efficient operating system**
  - **System management**



BlueGene/L Interconnection Networks



3 Dimensional Torus

- Interconnects all compute nodes (65,536)
- Virtual cut-through hardware routing
- 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- Communications backbone for computations
- 0.7/1.4 TB/s bisection bandwidth, 68TB/s total bandwidth

Global Tree

- One-to-all broadcast functionality
- Reduction operations functionality
- 2.8 Gb/s of bandwidth per link
- Latency of tree traversal 2.5 µs
- ~23TB/s total binary tree bandwidth (64k machine)
- Interconnects all compute and I/O nodes (1024)

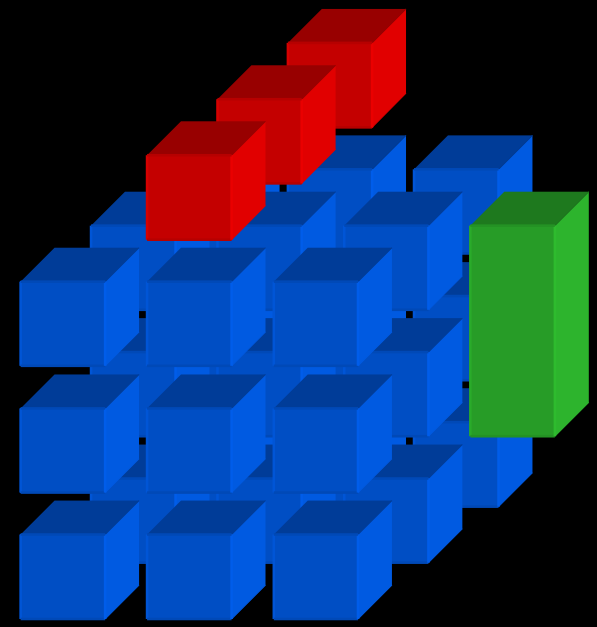
Ethernet

- Incorporated into every node ASIC
- Active in the I/O nodes (1:64)
- All external comm. (file I/O, control, user interaction, etc.)

Low Latency Global Barrier and Interrupt

Control Network

BlueGene/L System Software



compute node
application volume

I/O node
operational surface

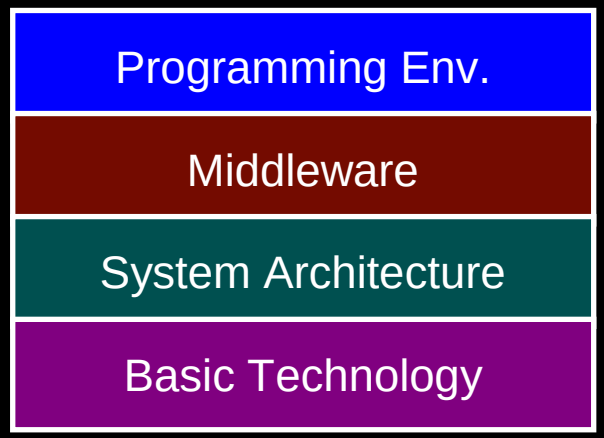
service node
control surface

- Hierarchical organization
  - 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, debugging, and termination
  - Service node performs system management services (e.g., heart beating, monitoring errors) – largely transparent to application/system software
- Looks like a 1024-node cluster to outside world

  Job scheduling through LoadLeveler extensions
- File system: GPFS
- Libraries: ESSL, MPI

2010: **H**igh **P**roductivity **C**omputing **S**ystems Research PERCS (Productive, Easy-to-use, Reliable Computer System)

Balanced attack across all system layers



Main theme: A system that adapts to the application, not the other way around

- Continuous program optimization
- System performance evaluation methodology and infrastructure

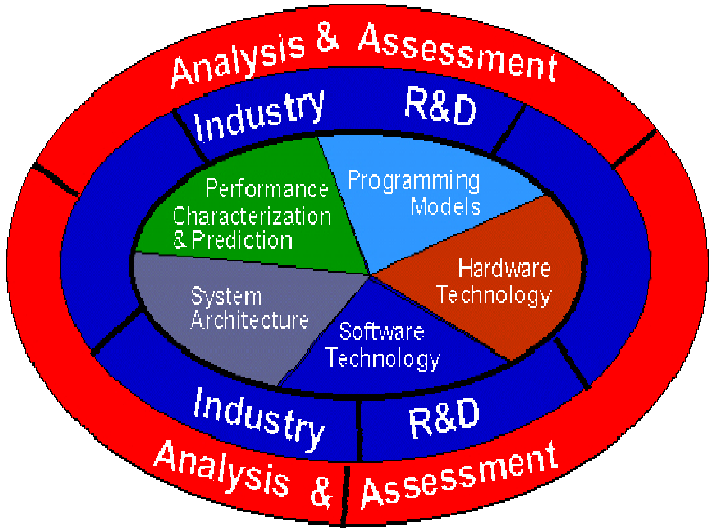
- Programming environments
  - Focus on simplifying programming tasks and reducing development cycle
- Scalable OS and middleware
  - Support for on-demand computing
- Compilers
  - Tolerating memory latency
- Development of new systems analysis tools
  - An execution-driven evaluation infrastructure
- Systems architecture
  - Application dependent morphing architectures under software control, addressing the memory wall
- Circuits/power/technology
  - High performance, lower power circuits, system level power analysis, advanced packaging

High Productivity Computing Systems

Provide a new generation of economically viable high productivity computing systems for the national security and industrial user community)

Goals:

- **Performance** (time-to-solution): speedup critical national security applications by a factor of 10X to 40X
- **Programmability** (time-for-idea-to-first-solution): reduce cost and time of developing application solutions
- **Portability**
- **Robustness** (reliability): includes security in addition to traditional RAS



HPCS Program Focus Areas

Applications:

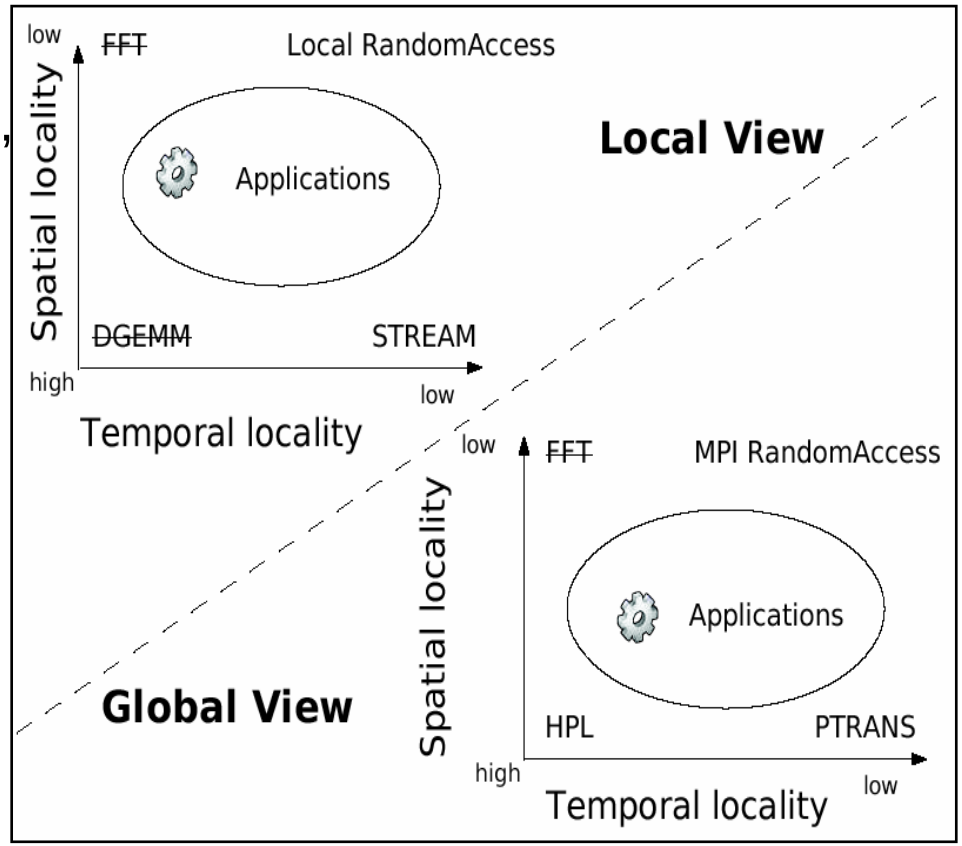
- Intelligence/surveillance, reconnaissance, cryptanalysis, weapons analysis, airborne contaminant modeling and biotechnology

Ref: R. Graybill, DARPA 1/27/03

IBM

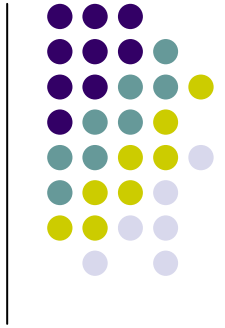
Customer Expectations

- High performance on HPC Challenge, e.g.
  - GUPS (security application, network bound)
  - STREAM (data streaming, memory bound)
  - Linpack (traditional HPC, CPU bound)
- High productivity:
  - Ease of programming, administration & general use
  - Robustness
- Actual applications:
  - UMT-2K, …



All delivered within a commercially-viable, mainstream product

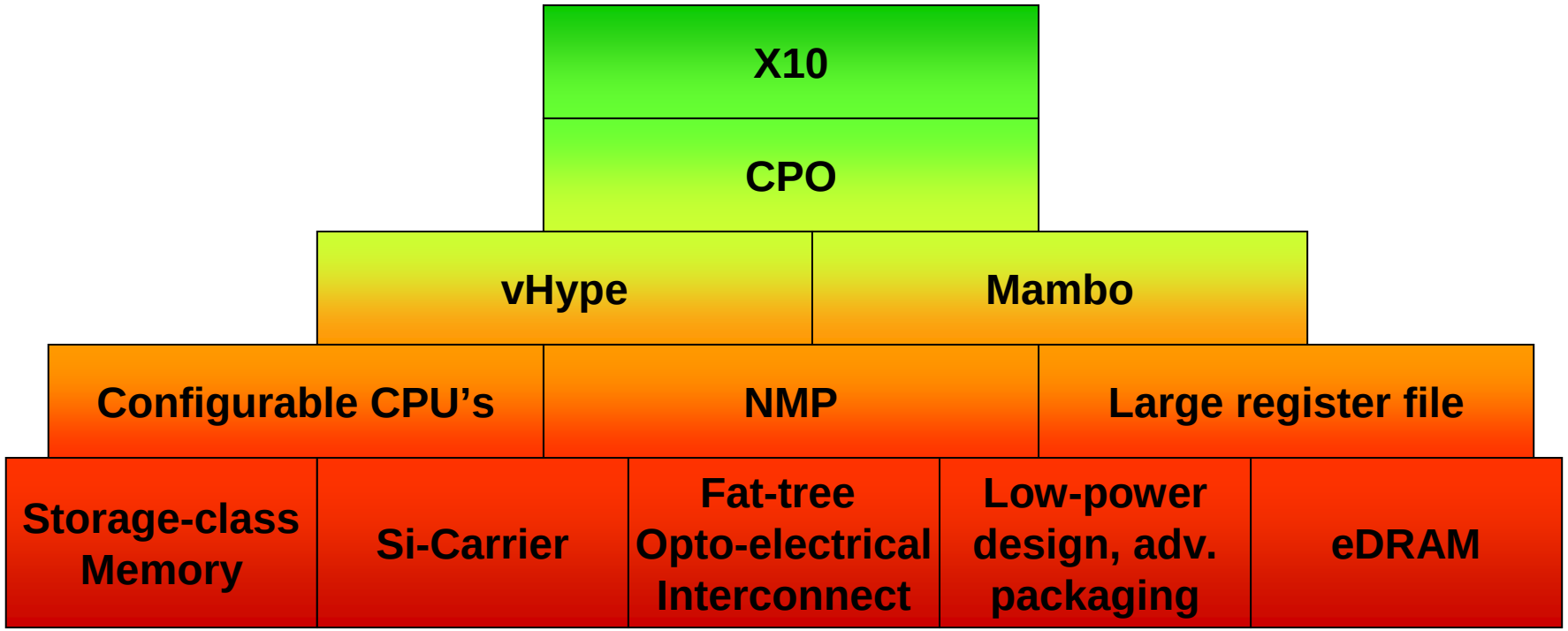
The IBM Team



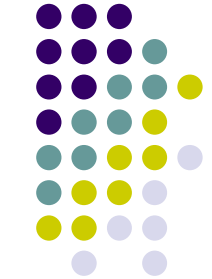
- IBM Research
- Software Group
- Server and Technology Group
- Universities
  - Systems and architecture
    - MIT, UT Austin, Illinois, RPI
  - Programming environments & languages.
    - UC Berkeley, Purdue, Vanderbilt, U. of Delaware
  - Usability and applications
    - Cornell, U of New Mexico, Pittsburgh, Dartmouth, Los Alamos National Lab

IBM

PERCS Technology Bets



In Summary



- Cross-stack technologies for 2010, integrated SW-HW
- Goal is to influence IBM’s main products and compete successfully for building a peta-scale machine by 2011
- Large team effort, government partnership
- Cost realism, software inertia, and other issues may limit the reach of our effort, and it’s important to adjust expectations accordingly

IBM

POWER Everywhere

Nintendo®

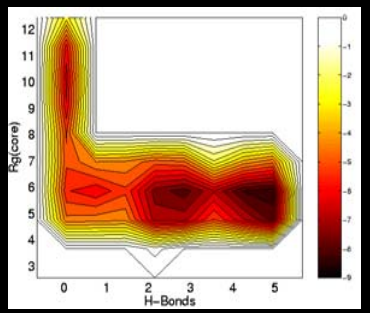
IBM eServer
BladeCenter
Total Storage

BLUE GENE

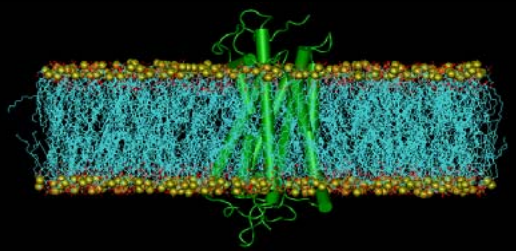
ASCI PURPLE

Blue Gene Science

- Advance our understanding of biologically important processes via simulation, in particular the mechanisms behind protein folding
  - Thermodynamic & kinetic studies of model peptide systems



  - Structural and dynamical studies of membrane and membrane/protein systems

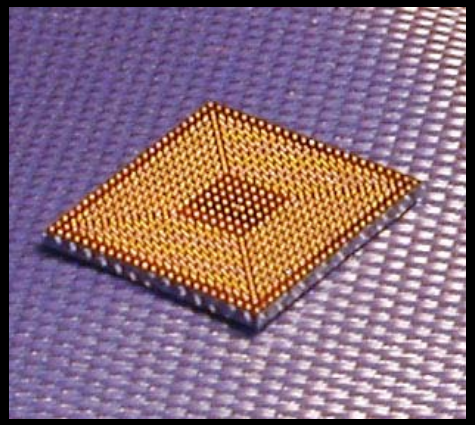


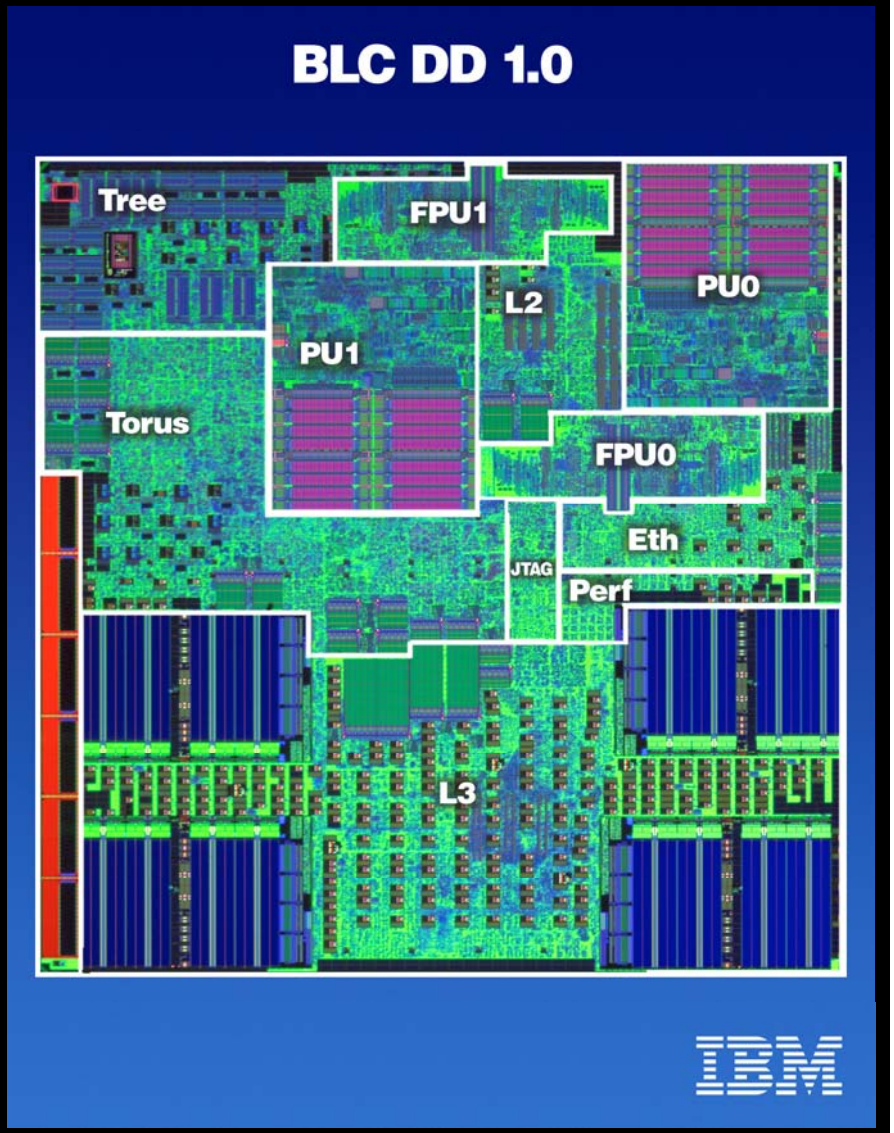
Innovation – Holistic Design

It’s the SYSTEM, !#$!@#$%!!!!!!

BlueGene/L System-on-a-Chip ASIC

- 130nm
- 11 × 11 mm die size
- 25 × 32 mm CBGA
- 474 pins, 328 signal
- 1.5/2.5 Volt





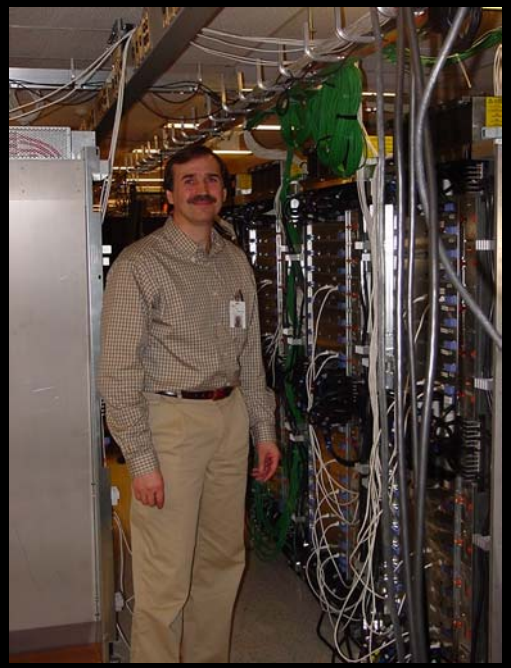
Integrated functionality

- Two PPC 440 cores
- Two “double FPUs”
- L2 and L3 caches
- Torus network
- Tree network
- JTAG
- Performance counters
- EDRAM

BlueGene/L on TOP500

I.B.M. Decides to Market a Blue Streak of a Computer
NY Times, 21 June 2004

From the June 2004 TOP500 List



#4: BlueGene/L Prototype (500 MHz, 256 MB/node)

- 8192 processors
- **11.68 TF/s** (73% of peak)
- 72 kW
- 0.162 GF / W

#8: BlueGene/L Prototype (700 MHz, 512 MB/node)

- 4096 processors
- 8.655 TF/s (75% of peak)
- 46.8 kW
- 0.185 GF / W

The Discontinuity

Then (2002)

- Scaling drove performance
- Scaling drives down cost
- Performance constrained
- Active power dominates
- Line tailoring in manufacturing
- Focus on technology performance

Now (2004)

- Innovation drives performance
- Scaling drives down cost
- Power constrained
- Standby power dominates
- Performance tailoring in design
- Focus on system performance

HPC Cluster Directions