Argonne NATIONAL LABORATORY

The LLNL/ANL/IBM Collaboration to Develop BG/P and BG/Q

Rick Stevens

Argonne National Laboratory

The University of Chicago

The Blue Gene Approach

- Focus on low power
- Focus on communications networks
- Stress Simplicity
  – Choose the right areas to innovate
  – Note: There is NO exotic or bleeding edge technology in Blue Gene
- Utilize PowerPC architecture and standard messaging interface (MPI)
  – Standard, familiar programming model and mature compiler support
- Use embedded system-on-a-chip (SOC) design
  – Advantages in utilizing SOC technique
    - Single-chip nodes result in significant reduction of complexity
  – Simplicity is critical, enough complexity already due to scale
    - Significant reduction of power
  – Critical to achieving a dense and inexpensive packaging solution
  – An absolute requirement for the future
    - Significant reduction in time to market, lower development cost and lower risk
  – Much of the technology is qualified
- Improve cost/performance (total cost/time to solution)
- Close attention to RAS (reliability, availability, and serviceability) at all system levels
  – One of the biggest challenges

Blue Gene Project

- **Project began in December 1999, With two initial goals:**
  – Advance the state of the art of biomolecular simulation
  – Advance the state of the art in computer design and software for extremely large scale systems
- **Goals, motivation and philosophy quickly evolved:**
- Address as broad as possible a set of applications while maintaining the cost/performance and power/performance of special purpose machines
  – **Many special purpose machines had been very successful despite the incredible technical barriers**
- Complexity and power are major driver for cost and reliability
  – **Optimal design point is very different from standard approach based on high-end superscalar nodes**
  – **Traditional supercomputer-processor design is hitting power/cost limits**
- Some applications for supercomputers do scale fairly well
  – **Growing volume of such applications**
  – **Physics is mostly local**
  – **Darwinian selection of applications is strong for supercomputers**
- Integration, power, and technology directions are driving toward multiple modest cores on a single chip rather than one high-performance processor
  – Watts/FLOP will not improve much from future technologies

Increasing Blue Gene Impact

- SC 2005 Gordon Bell Award, 101.7 TFs on real materials science simulation
  - Recently exceeding 150 TFs sustained
- Sweep of the all four HPC Challenge class 1 benchmarks
  - G-HPL (259 Tflop/s), G-RandomAccess (35 GUPS) ,
  - EP-STREAM (160 TB/s) and G-FFT (2.3 Tflop/s).
- Over 80 large-scale applications ported and running on BG/L

27.6 kW power consumption per rack (max)
7 kW power consumption (idle)

Blue Gene Installations (14 + 4 at IBM)

- **Lawrence Livermore National Laboratory – US**
- **TJ Watson Research Laboratory, IBM Almaden Res. Lab -- US**
- **ASTRON – Netherlands**
- **Advanced Industrial Science and Technology – Japan**
- **San Diego Supercomputer Center – US**
- **NIWS – Japan**
- **National Center for Atmospheric Research -- US**
- **Argonne National Laboratory – US**
- **University of Edinburgh – UK**
- **Boston University – US**
- **EPFL – Switzerland**
- **IBM Zurich Research Laboratory -- Switzerland**
- **Juelich -- Germany**
- **Princeton Plasma Physics Laboratory -- US**
- **Massachusetts Institute of Technology -- US**
- **Iowa State University -- US**

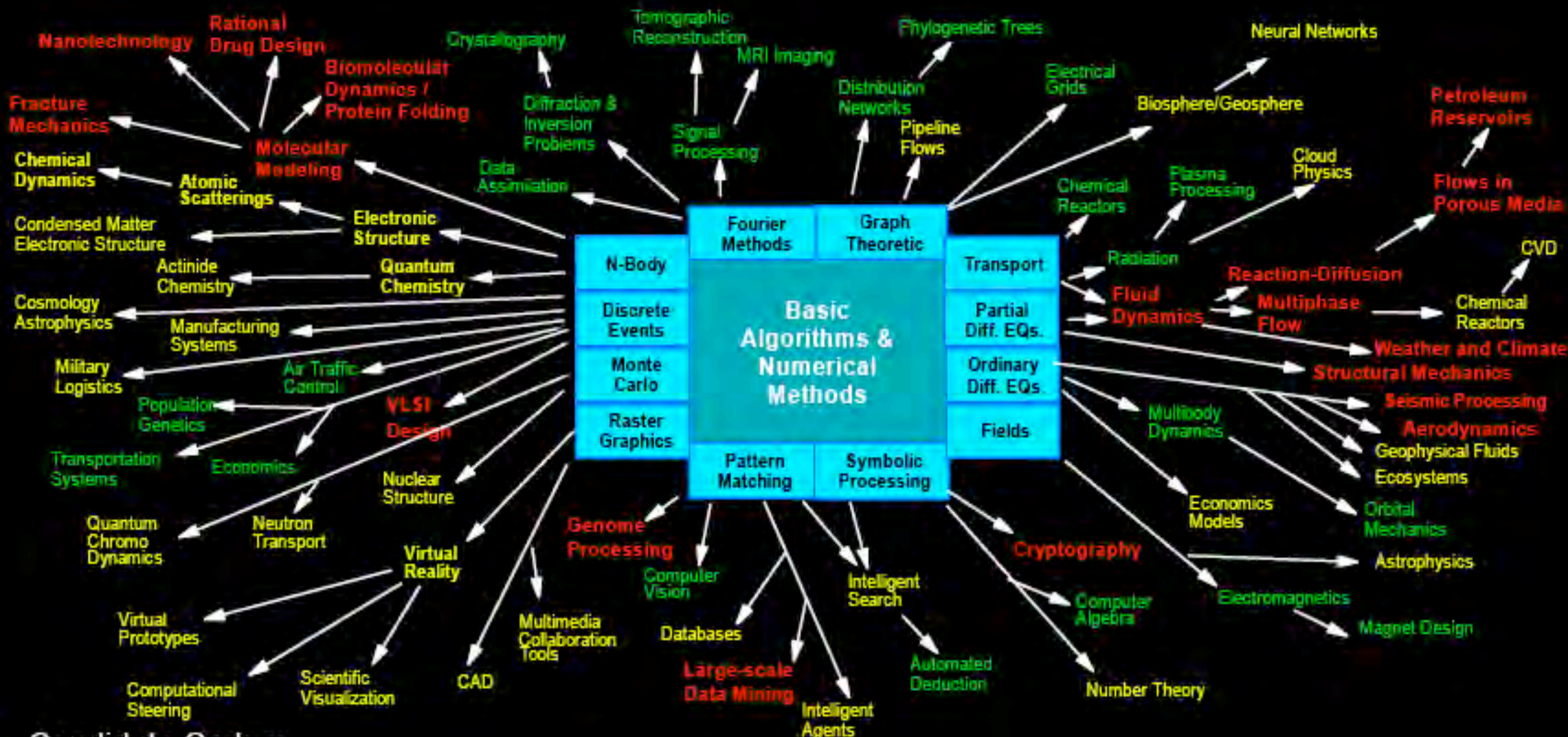
Blue Gene Consortium (53 Institutions, 213 pp)

Ames National Laboratory/Iowa State University
Argonne National Laboratory
Brookhaven National Laboratory
Fermi National Laboratory
Jefferson Laboratory
Lawrence Berkeley National Laboratory
Lawrence Livermore National Laboratory
Oak Ridge National Laboratory
Pacific Northwest National Laboratory
Princeton Plasma Physics Laboratory
UNIVERSITIES:
Boston University
California Institute of Technology
Columbia University
Harvard University
Indiana University
Louisiana State University
Massachusetts Institute of Technology
National Center for Atmospheric Research
New York University/Courant Institute
Northern Illinois University
Northwestern University
Ohio State University
Pittsburgh Super Computing Center
Princeton University

Purdue University
Rutgers University
Stony Brook University
Texas A&M University
University of California - Irvine
University of California - San Francisco
University of California - San Diego/San Diego Super Computing Center
University of Chicago
University of Colorado - JILA
University of Delaware
University of Illinois Urbana-Champaign
University of Minnesota
University of North Carolina
University of Southern California/Information Sciences Institute
University of Texas at Austin/Texas Advanced Computing Center
University of Utah
University of Wisconsin
INDUSTRY:
Engineered Intelligence Corporation
IBM
INTERNATIONAL MEMBERS:
Trinity College, Trinity Centre for High Performance Computing, O'Reilly Institute, Dublin, Ireland
John von Neumann Institute for Computing, Juelich, Germany
NIWS Co., Ltd., Tokyo Japan

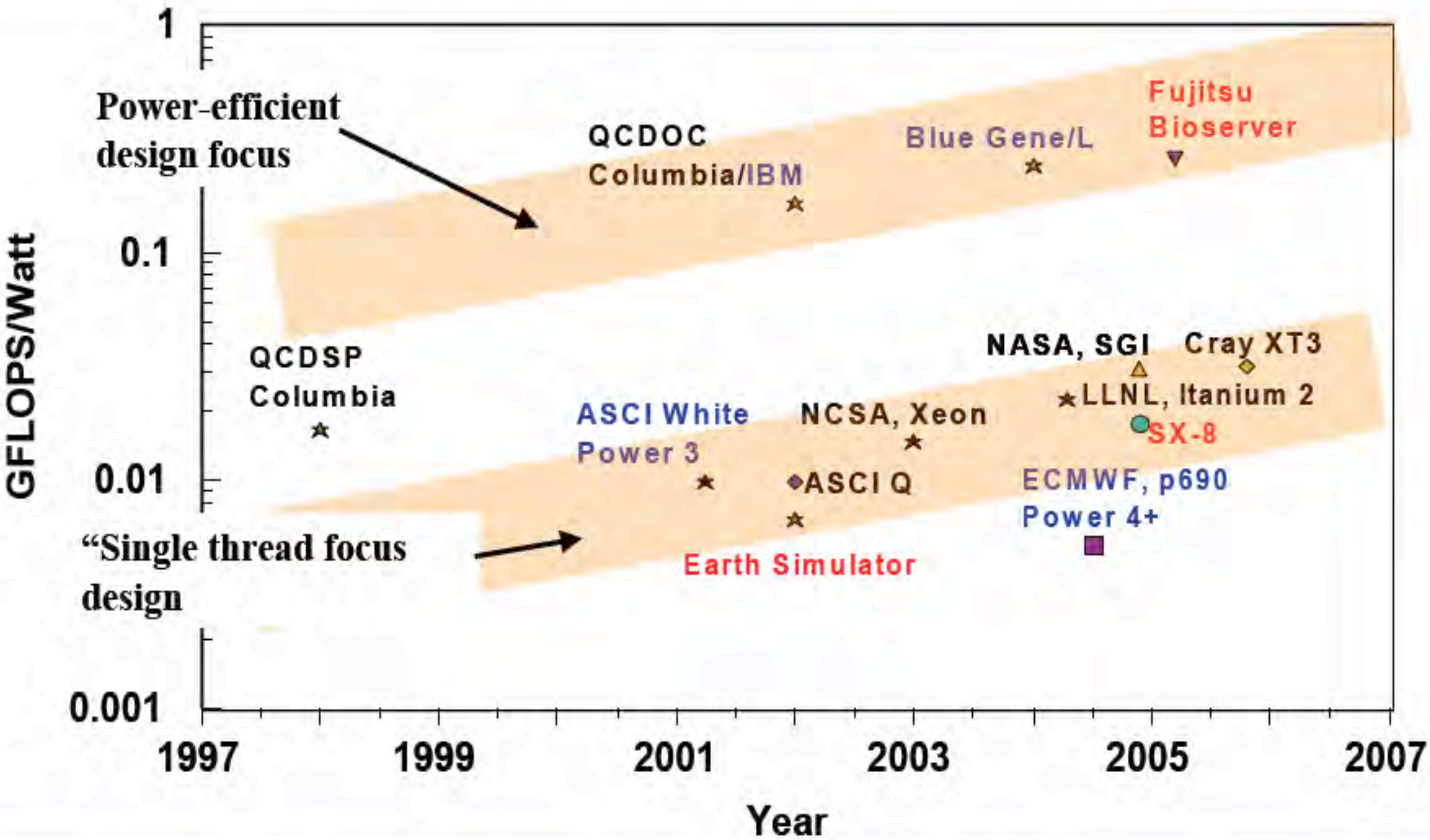
Good Better Best

Many Classes of Applications are Massively Parallel



- Candidate Codes:
  - Inherently parallel; written using MPI
  - Memory required per MPI task is less than that available on a BG/L node
  - Dominated by collective communication across all nodes
  - Locality of communications within 3D mapping
- Non-Candidate Codes:
  - Large memory footprints required on individual nodes
  - Client/server structures
  - Dominated by disk I/O

Supercomputer Power Efficiencies

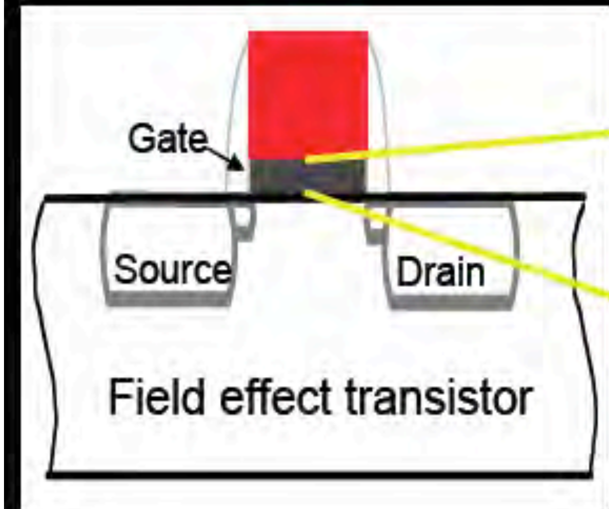


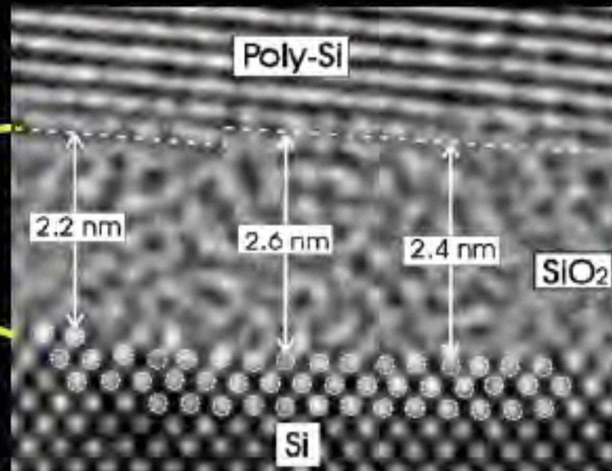
Future Scaling without innovation

If we scale current peak performance numbers for various architectures and allowing system peak doubling every 18 months. Trouble ahead

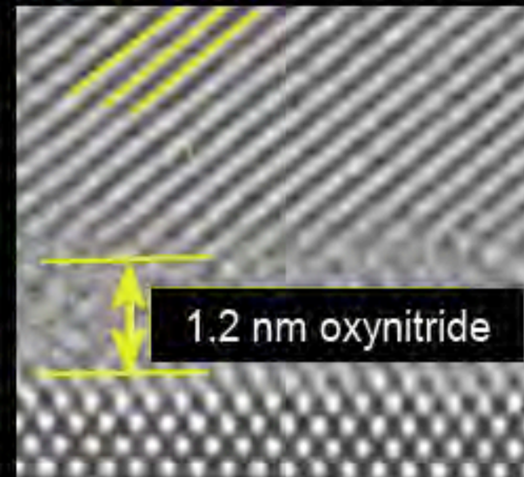
	Projected Year	BlueGene/L	Earth Simulator	MareNostrum
250 TF	2005	1.0 MWatt	100 MWatt	5 MWatt
1 PF	2008	2.5 MWatt	200 MWatt	15 MWatt
10 PF	2013	25 MWatt	2000 MWatt	150 MWatt
100 PF	2020	250 MWatt	20,000 MWatt	1500 MWatt

Why CMOS scaling breaks down-We're down to atoms





"Thick" gate oxide

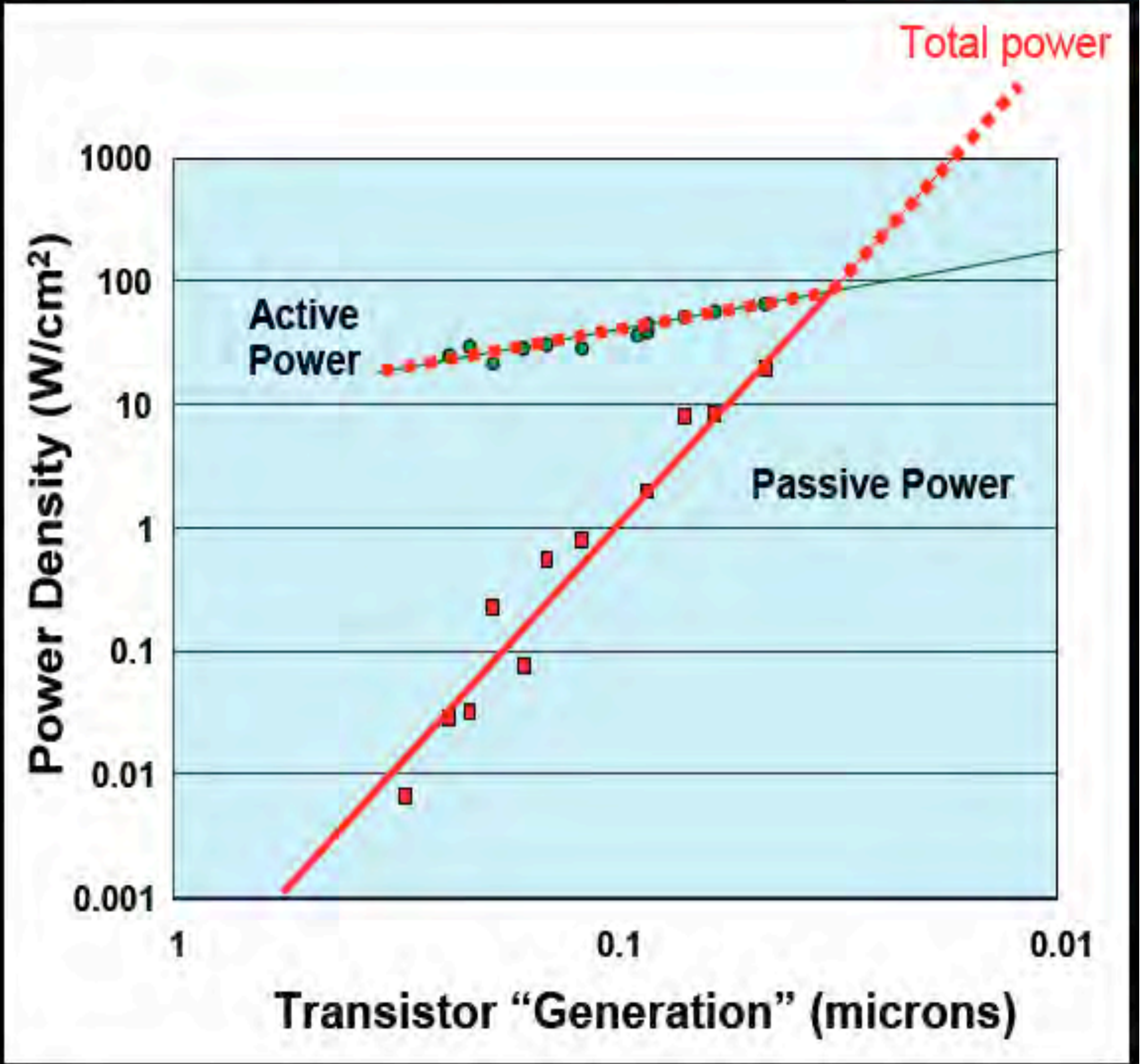


Scaled gate oxide

- 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
      - Probably not a positive for reliability
    - Oxides scaled below ~9 angstroms are too "leaky" and thus unreliable
- Industry is currently at the limitation imposed by physics

Impact on design: “Close” is no longer remotely good enough

- “Challenges”
  - “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

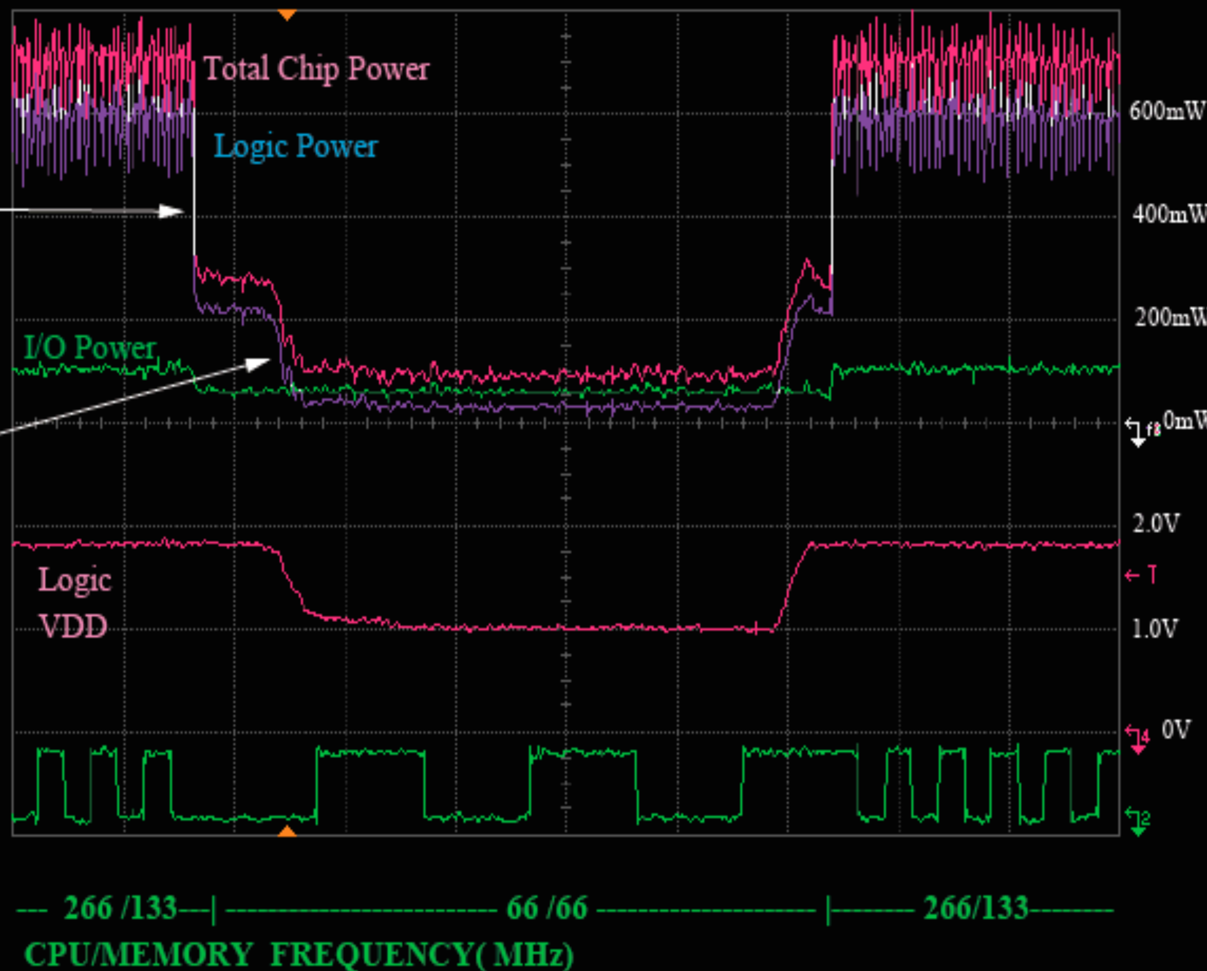


Active Power Management

Dynamic Frequency Scaling
266Mhz CPU to 66MHz CPU

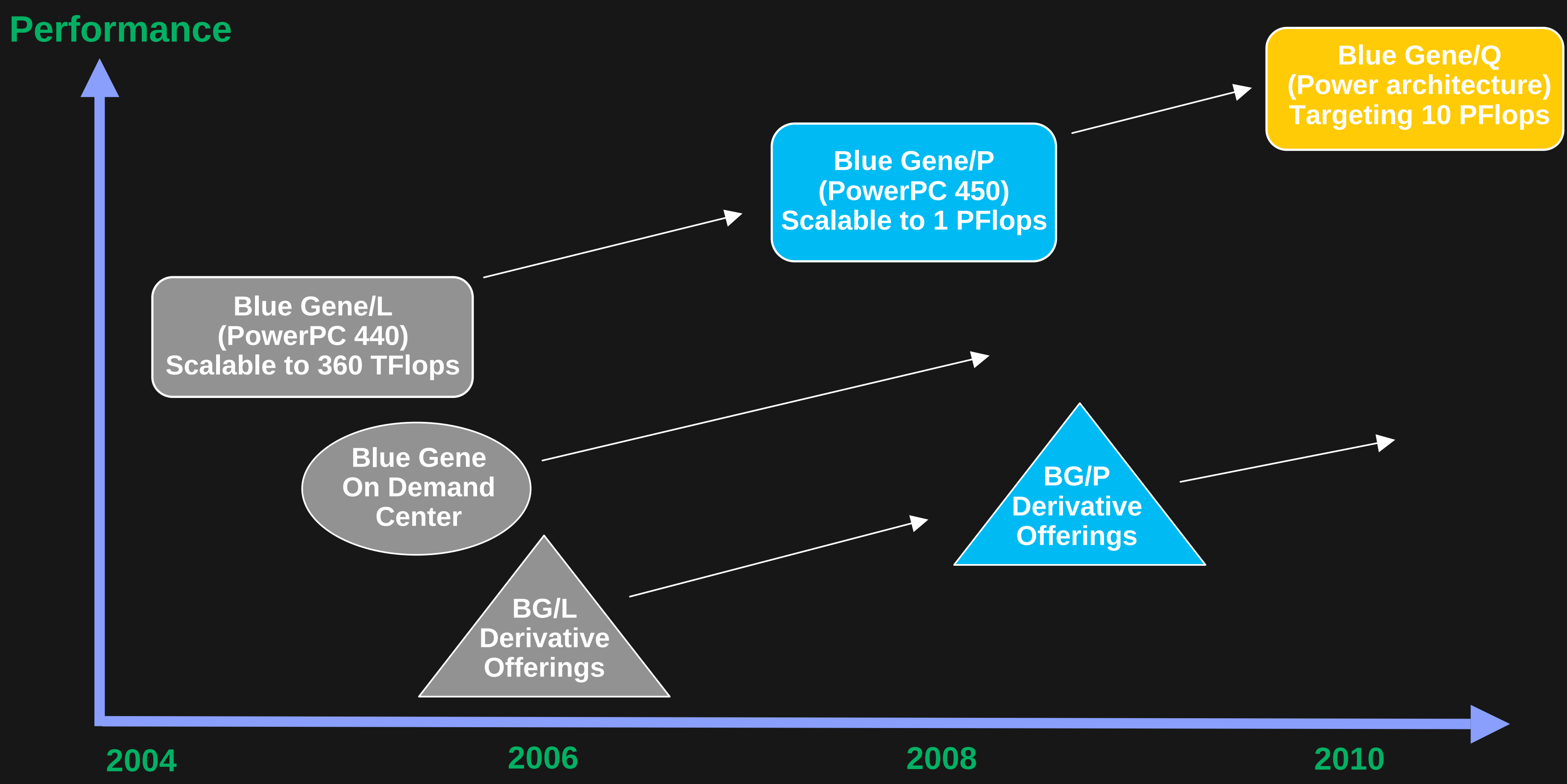
Dynamic Voltage Scaling
1.8V --> 1.0V at upto 1V/100us

Uninterrupted Operation
Linux 2.3.17 Running
Dhrystone 2.1 code
400 loops per cycle .



Power consumption for the CPU and logic was reduced by 13X dynamically under the control of the Linux kernel
(NO PLL Relock and NO stopping of the application)

Blue Gene® Supercomputer– Product Directions



BG[PQ] Research and Development Goals

- Utilize critical applications from NNSA and SC as program drivers to
  - Enable a broader class of applications reach at each generation
  - Improve performance of these critical applications
  - Reduce level overall complexity and rigidity of system administration, RAS and operations
- BlueGene/P
  - Improved SMP performance and compute node kernel functionality
  - Open source, collaborative development model
  - Improved compiler generated code
  - 10Gb/s Ethernet external IO infrastructure
  - Enable 100TF/s by 2007, 1PF by 2008
- BlueGene/Q
  - Microprocessor core
  - Node architecture
  - Interconnect architecture
  - Next generation software environment
  - Targeting 10PF/s demonstration 2010/2011

BlueGene/P – Architectural Highlights

- BG/P is an update of BG/L
  - CPU architecture from PPC440’ to PPC450’
  - Network topology is the same
  - Systems organization is the same
- Scaled performance through higher density and frequency bump
  - 2x performance through doubling the processors/node
  - 1.2x from frequency bump due to technology
- Enhanced functionality of the compute node
  - 4 way SMP
  - Greatly enhanced 64 bit performance counters (including 450 core)
- Hold BlueGene/L packaging as much as possible
  - Improve networks through higher speed signaling on same wires
- External I/O network from 1 Gbps to 10 Gbps per node

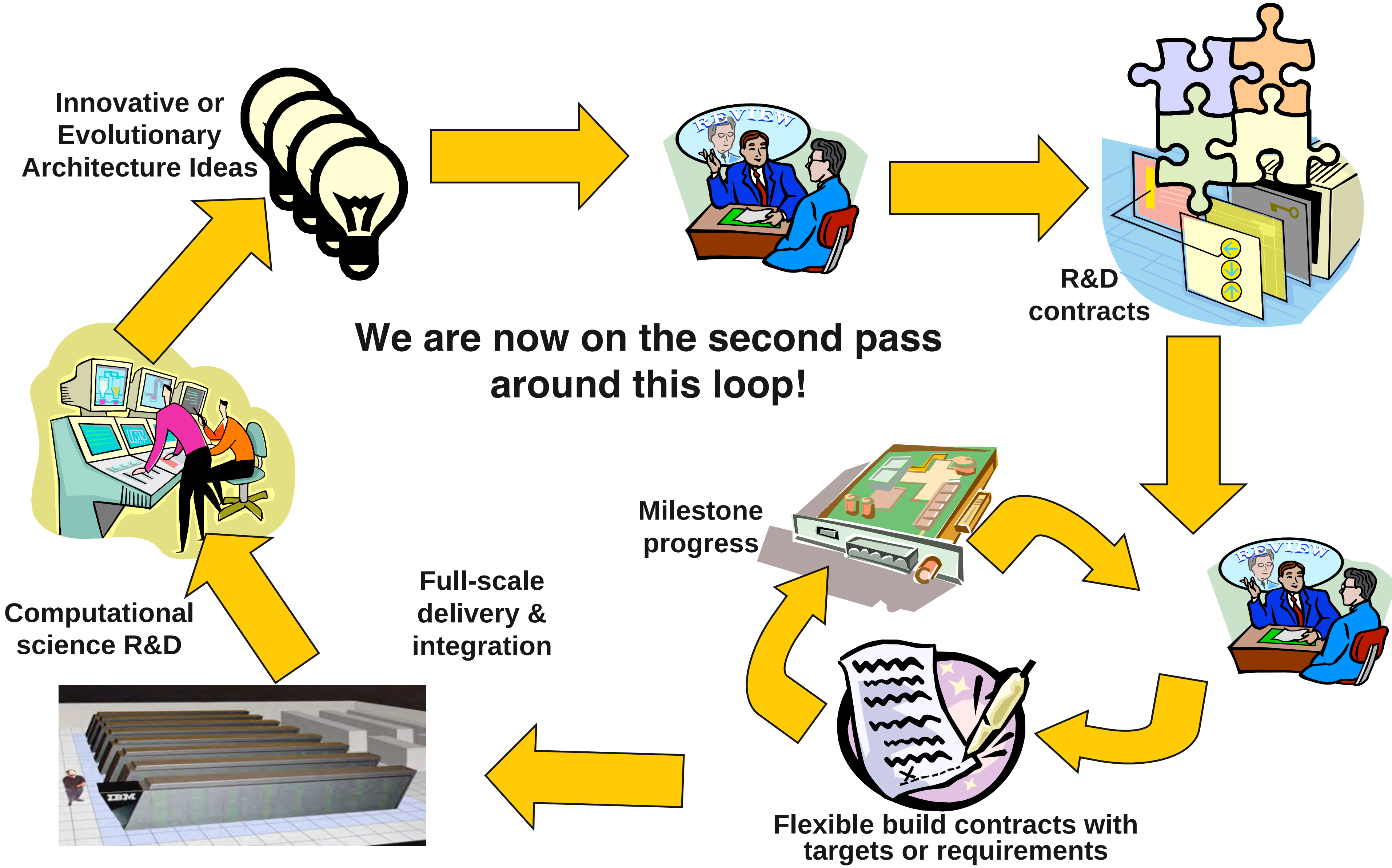
BG/P Software Directions

- Programming Models
  - MPI only – virtual node mode
  - MPI + OpenMP
  - Global address space – Global Arrays, UPC, CAF
- Compute Node Kernels
  - Extensions to mini kernel: threads, limited dynamic linking (Python, PERL) – must scale to 72K nodes (1 PF)
  - Linux – for smaller systems, partitions in a large BG/P system
- XL Compilers – OpenMP support, better SIMD support
- MPI support
  - Full MPI2, except for dynamic process management
  - Exploitation of DMA – overlap between computation and communication
- Control System
  - Internals re-architected for scalability – simplicity and robustness
  - Most external interfaces will be preserved – RAS database
- Complete HPC software stack from beginning
  - GPFS, Loadleveler, ESSL, HPC Toolkit

Two Phase Plan Based on Experience w/ BGL

- Research and Development Phase
  - R+D culminates in platform demo (~beta) with actual application performance metrics
  - Focus on both hardware and software
- Build Phase
  - Standard fixed price build contract with targets, not requirements
    - *Targets set so they can be met, on average*
    - *Add flexibility on memory pricing*
  - Once full speed demo complete, modify build contract
    - *Targets → Achievable Requirements (minimums)*
- Review Process
  - Technical working groups for joint development projects
  - Monthly technical reviews
  - Quarterly NNSA tri-lab, DOE SC lab and IBM executive reviews
  - External reviews at critical junctures → no-fault termination

BlueGene Development Model Minimizes Risk and Maximizes Flexibility to Innovate



Scope of the Development Program

- Addresses both BG/P and BG/Q development
  - BG/P has same level of specificity as BG/L R&D contract
  - BG/Q is more evaluation type deliverables
    - *Planned refinement of later BG/Q milestone deliverables as more about the deliverables is known*
    - *Hardware refined after BG/Q CPU choice review*
    - *Software refined after BG/Q System Architecture Review*
- Each set of BG/P and BG/Q milestones are divided between hardware and software
- BG/P and BG/Q overlap
  - Long lead time architectural evaluations of BG/Q going on during BG/P development
  - As effort ramps down on BG/P, effort ramps up on BG/Q

BG [PQ] Software Strategy

- Build off BG/L experience
  - Improve compilers
  - Improve communications
  - Improve compute node kernel SMP and apps reach
- Change development, deployment and support model
  - Pursuing Open Source, Collaborative Development Model
  - Defined work groups in the following areas
    - *Job scheduling, partition management & system management (including RAS)*
    - *System OS functions*
    - *Applications development environment*
    - *Collaborative development model, coding and doc standards*
    - *Service and support (*difficult since this spans R&D and build contract boundaries*)*
- Build BG/Q software off BG/P experience

Conclusions

- DOE Office of Science (SC) and National Nuclear Security Agency (NNSA) have agreed to jointly support the continued development of Blue Gene P and Q systems
- LLNL and ANL have formed a partnership to co-manage the development contract with IBM
- The program will provide for technology demonstrations of BG/P in FY07 and BG/Q in FY10
- Risk is managed though incremental milestones, reviews, progress payments and multiple go/nogo decision points
- Deployment of systems will be via separate “build” contracts focused on specific systems, perhaps with comment terms and conditions for DOE based procurements
- The DOE-SC/NNSA BG[PQ] development program complements the DARPA HPCS program