T.J. Watson Research Center

Power Challenges in Extreme Scale Computing

Hans Jacobson

IBM T. J. Watson Research Center

hansj@us.ibm.com

ECTC Plenary Session - June 1, 2011

Extreme Scale Computing

What is extreme scale computing?

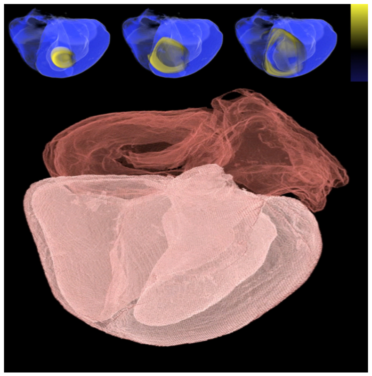
- Exascale computing identified by government agencies as critical need in 2018-2020 timeframe
- Exa refers to 10^{18} - *one million trillion* operations per second
- IBM top ranked system in "Top 500" 2008/2009 - first to reach a peak of 1 Petaflops
  - IBM's BlueGene product family have consistently been dominant players in the "Top500" and "Green500"
  - BlueGene won National Medal of Innovation & Technology for its breakthrough power/performance

Why is it a grand challenge?

- The constraints are multi-dimensional, interdependent and extremely hard to meet at affordable cost
  - 20 MWatt system power
  - 1 Exa-Flops sustained performance
  - MTBF of at least two weeks, preferably 1 month
- Exascale demands a ~1000x improvement in throughput in 10 years at a power increase of only ~10x
  - "Business as usual" scaling is not sufficient

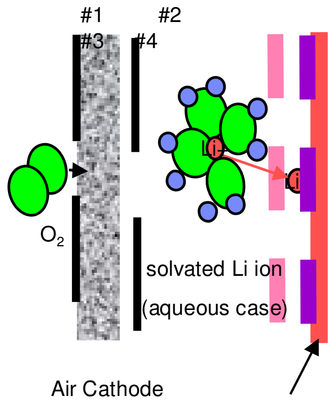
Ref: recent tutorial article by Josep Torrellas, "Architectures for Extreme Scale Computing," *IEEE Computer, Nov. 2009, pp. 28-35*

Many Examples of BIG Applications that Need Extreme Scale Computing

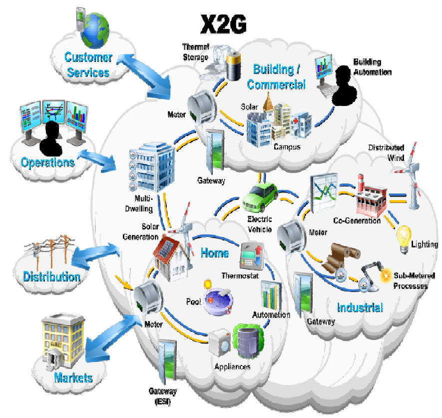


Whole Organ Simulation

Li/Air Batteries

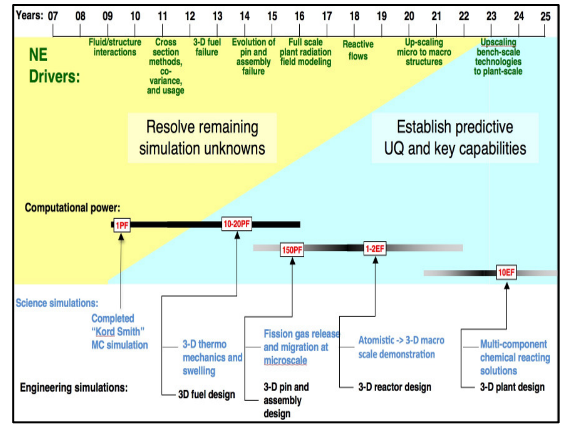


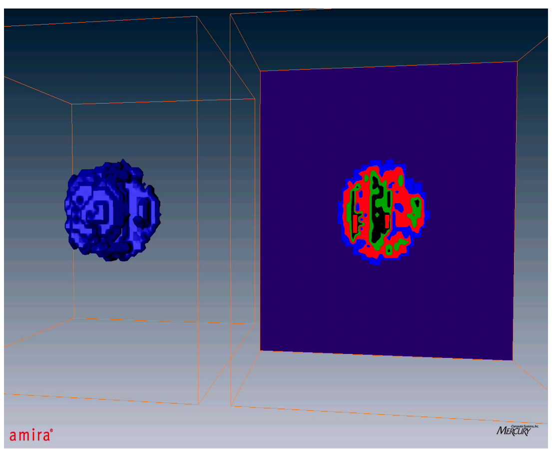
Smart Buildings



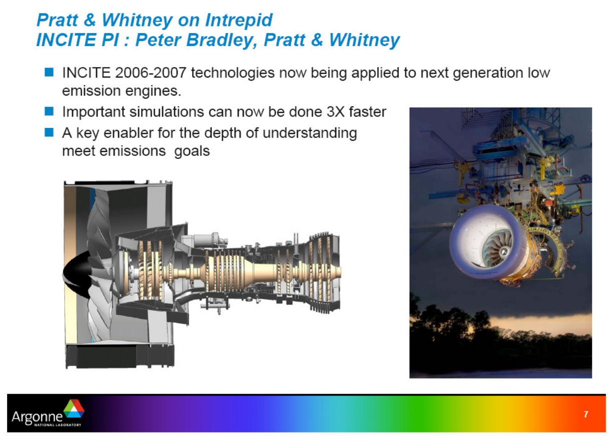
Smart Grid

Nuclear Energy



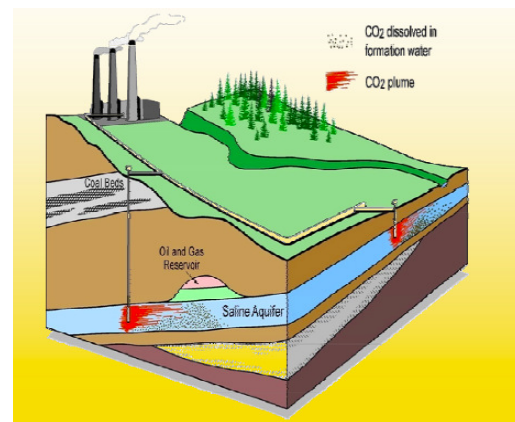


Tumor Modeling



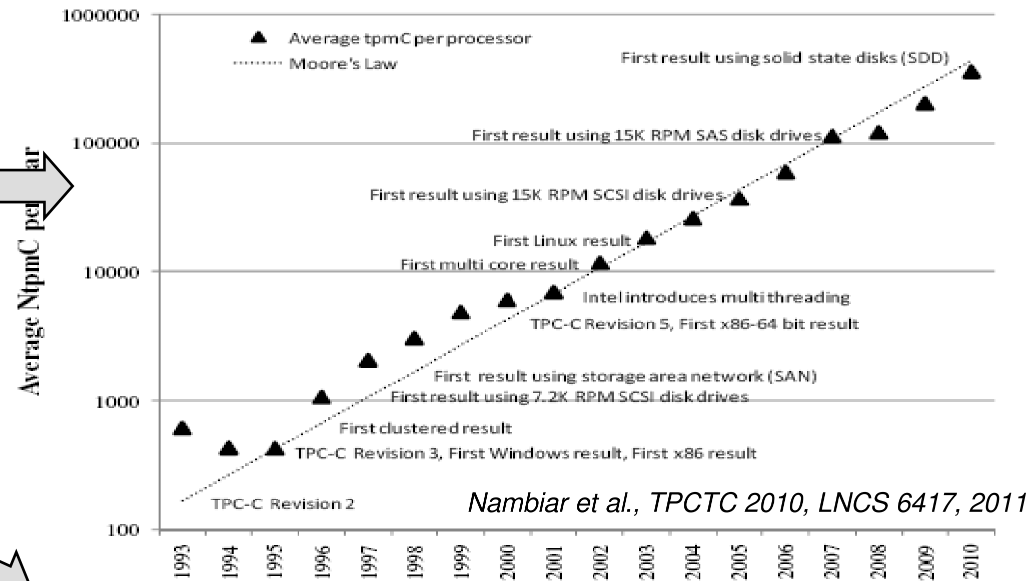
Low Emission Engine Design

CO_2 Sequestration

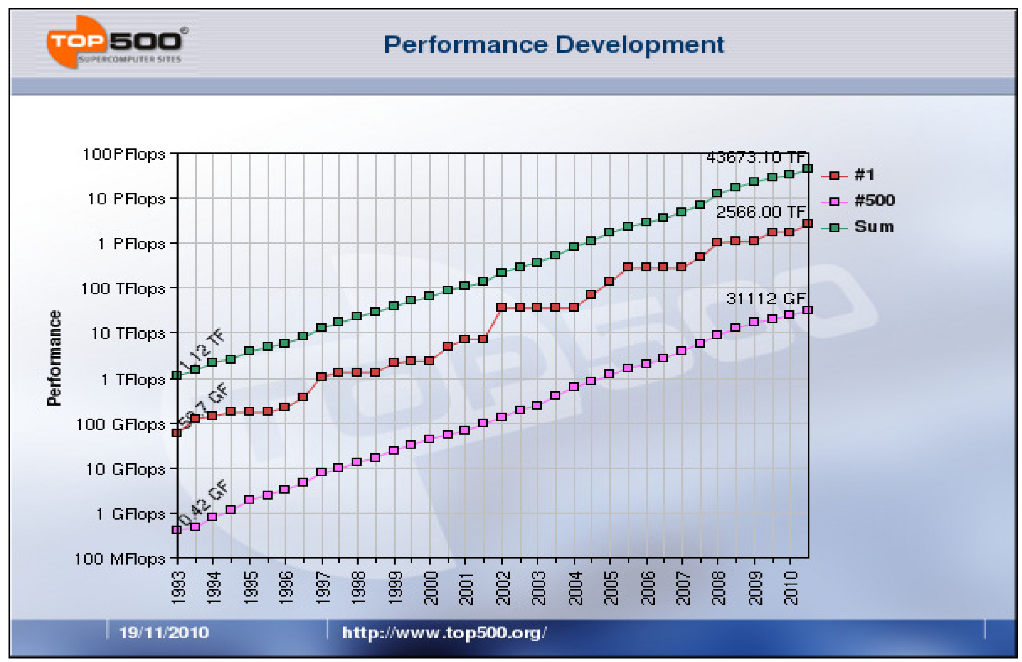


Performance Challenge of Extreme Scale Systems

- General purpose commercial servers have been on a *2X performance every 2 years* curve
- **Special-purpose** HPC supercomputers have been on a *4X performance every 2 years* curve
  - HPC expected to go from 1 PF at 2MW in 2008 to 1 EF at 20MW in 2018
  - Requires 1000x performance increase at only 10x power increase!
  - 1 EF would require about 80,000,000 2GHz 4-way DP SIMD cores for sustained ExaFlop performance!



Nambiar et al., TPCTC 2010, LNCS 6417, 2011



Power Challenge of Extreme Scale Systems

Oxide thickness is near the limit in late CMOS design era

- **Density improvements will continue but… power efficiency from technology will only improve very slowly.**
- **Historic trend of power efficiency improvement will slow**

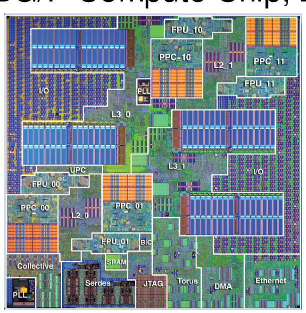
Rank	MFLOPS per Watt	KiloWatts	Supercomputer Location	Brand
1	722.98	59.49	Germany	IBM [QPACE]
1	722.98	59.49	Germany	IBM [QPACE]
1	722.98	59.49	Germany	IBM [QPACE]
4	458.33	276	DOE/NNSA/LANL (USA)	IBM [BladeCenter QS22]
4	458.33	138	IBM Poughkeepsie (USA)	IBM [BladeCenter QS22]
6	444.25	2345.5	DOE/NNSA/LANL (USA)	IBM [BladeCenter QS22]
7	428.91	51.2	Japan	
8	379.24	1484.8	China	
9	378.77	504	United Arab Emirates	IBM [BlueGene/P]
9	378.77	252	France	IBM [BlueGene/P]

Data from: http://www.green500.org

June 2009 Green 500 List:
If the world's most power efficient supercomputer is extrapolated to a sustained Exaflop, power would be …

~ **2 GigaWatts**

BG/P Compute Chip, 2007



System-on-a-Chip (SoC)

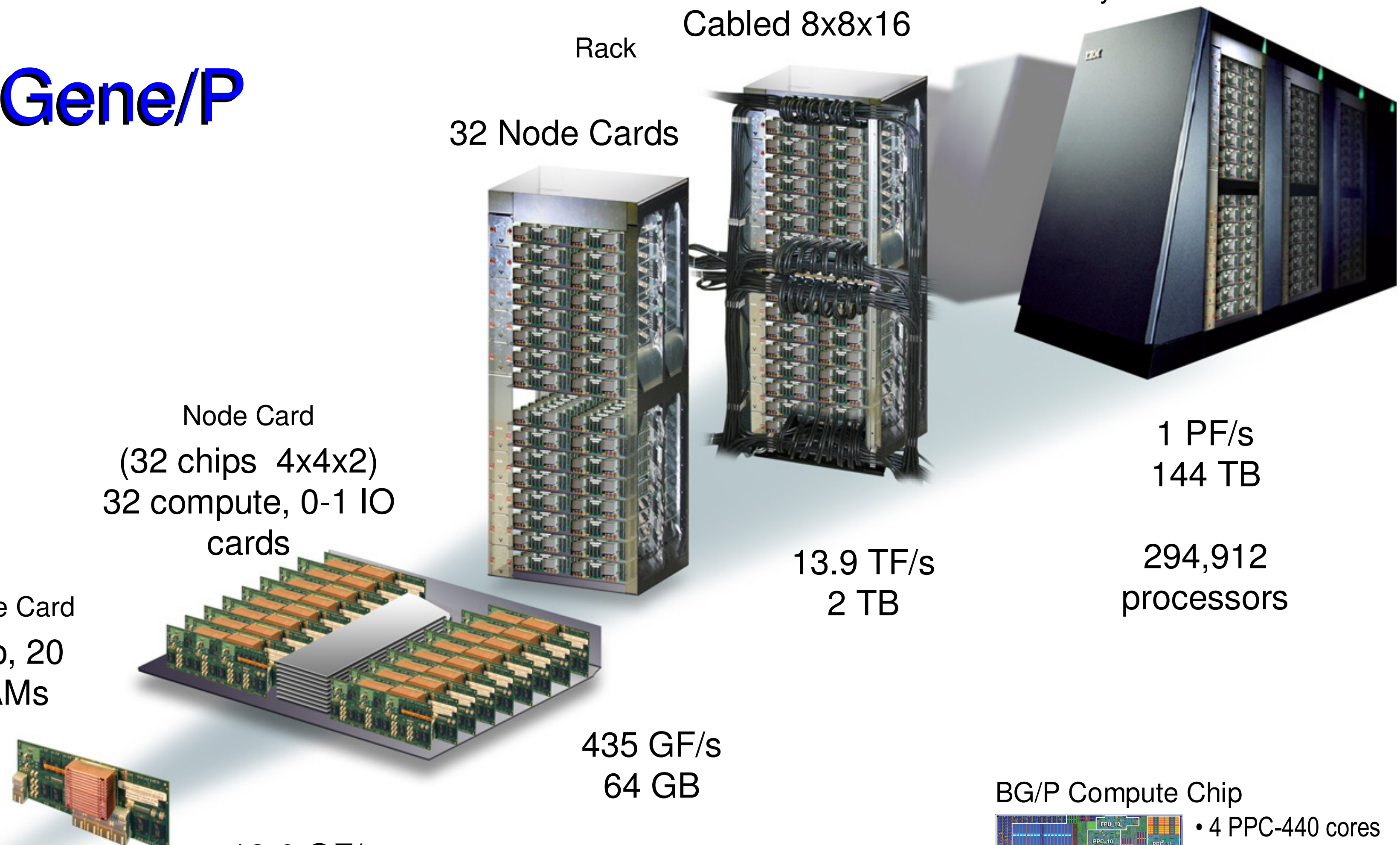
- 4 PPC-440 cores, 850 MHz
- IBM 90nm CMOS ASIC
- 173 sq. mm.
- 208 million transistors
- 16 W

IBM Blue Gene Supercomputers

National Medal of Technology & Innovation
October 2009

IBM has been a leader in large systems energy efficiency, but meeting the Exascale goals is nothing short of a *very* grand challenge!

BlueGene/P



System 72 Racks

Rack Cabled 8x8x16

32 Node Cards

Node Card

(32 chips 4x4x2)
32 compute, 0-1 IO cards

Compute Card

1 chip, 20 DRAMs

Chip

4 processors

13.6 GF/s
8 MB EDRAM

13.6 GF/s
2.0 GB DDR2
(4.0GB is an option)

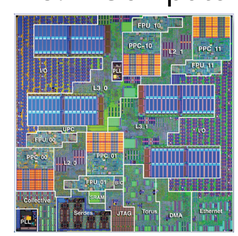
435 GF/s
64 GB

13.9 TF/s
2 TB

1 PF/s
144 TB

294,912 processors

BG/P Compute Chip

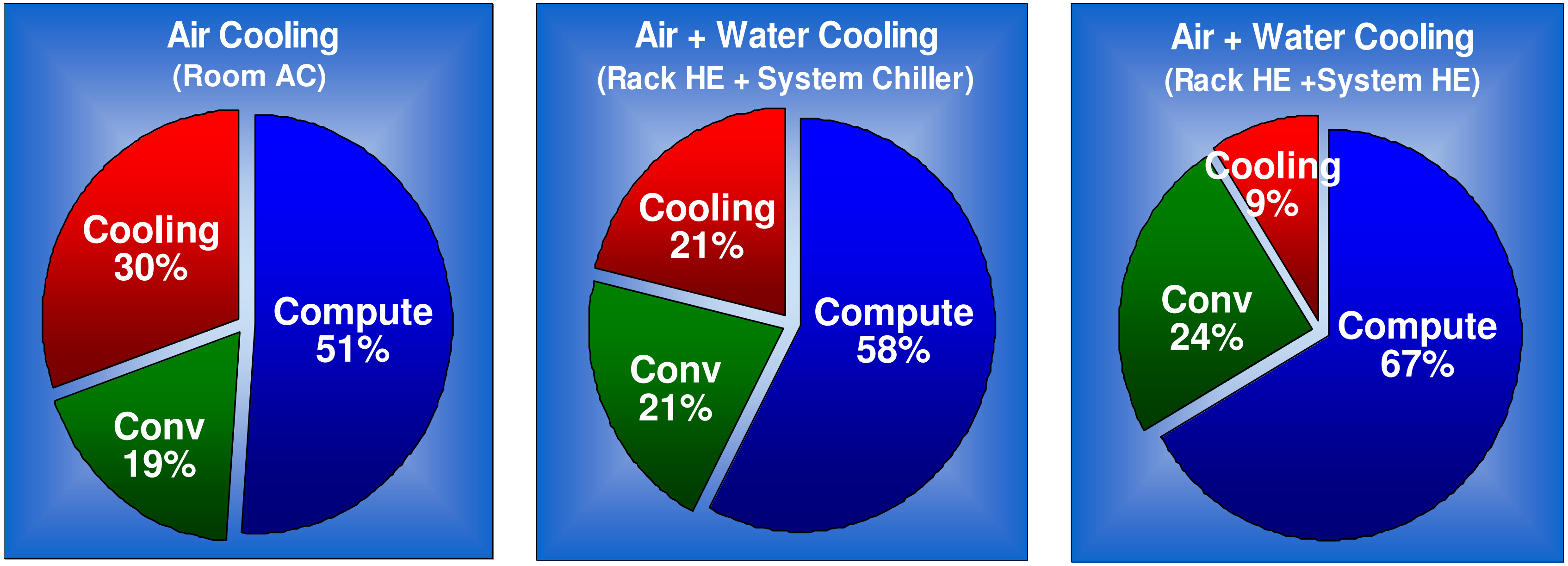


- 4 PPC-440 cores
- 850 MHz
- IBM 90nm ASIC
- 173 sq. mm.
- 208 M transistors
- 16 Watt

BG/P System Power Breakdown

- **System power components**
  - Cooling fans, water pumps and compressors
  - Voltage conversion and distribution loss
  - Computation, communication and storage

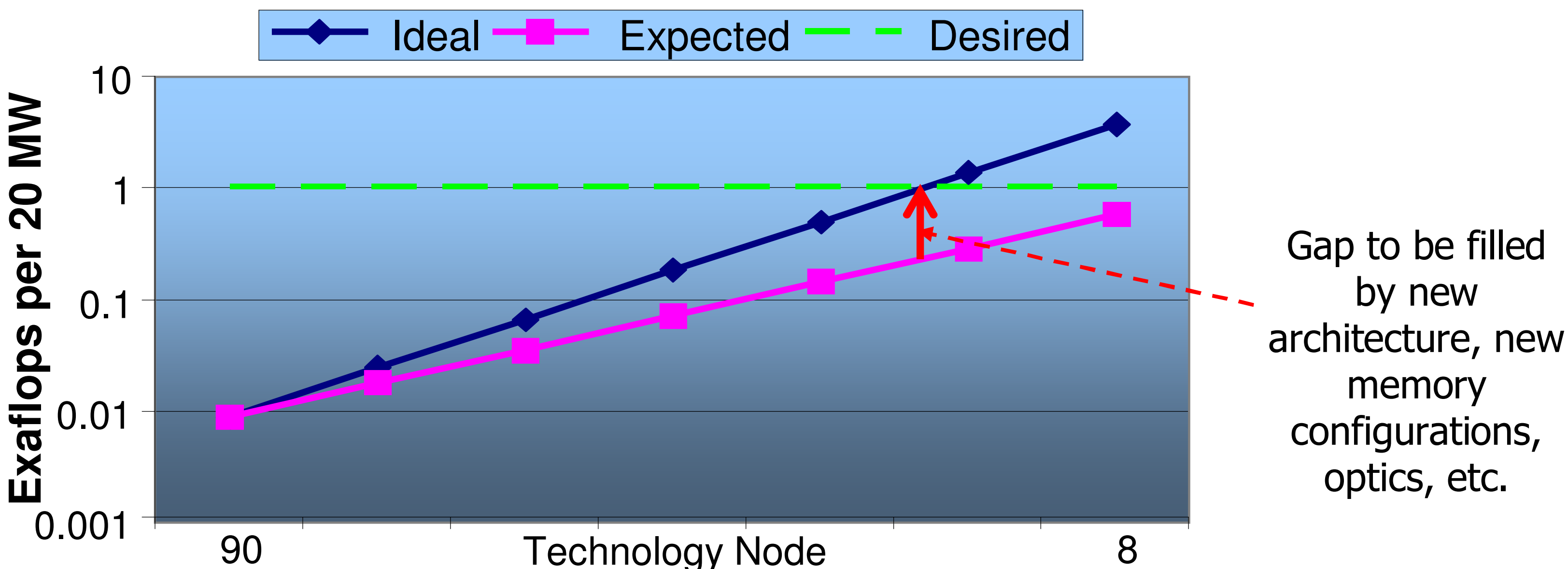
BG/P System Power Breakdown Running Linpack



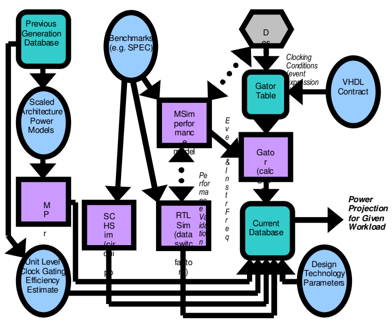
Challenging Road to Exascale

- **Technology improvements slowing down**
  - Significant gap between ideal and expected
  - Will not take us to Exascale in 2018 timeframe
- **Significant design innovation required to reach Exascale**
  - New processor architectures and memory configurations
  - Improved modeling and design optimizations
  - New power management techniques
  - Optics pervasive on board/module/chip

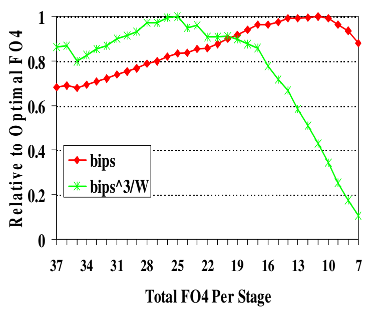
Must be cost-effective!

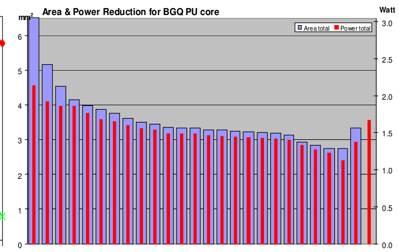


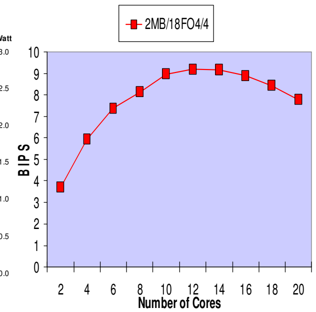
Broad Attack on the Power Front

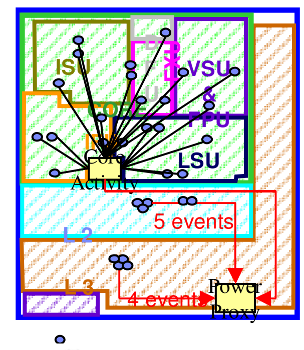


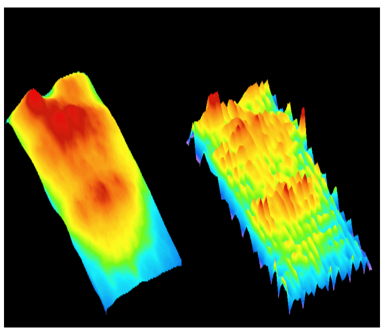
H. Jacobson et al., HPCA-17, 2011

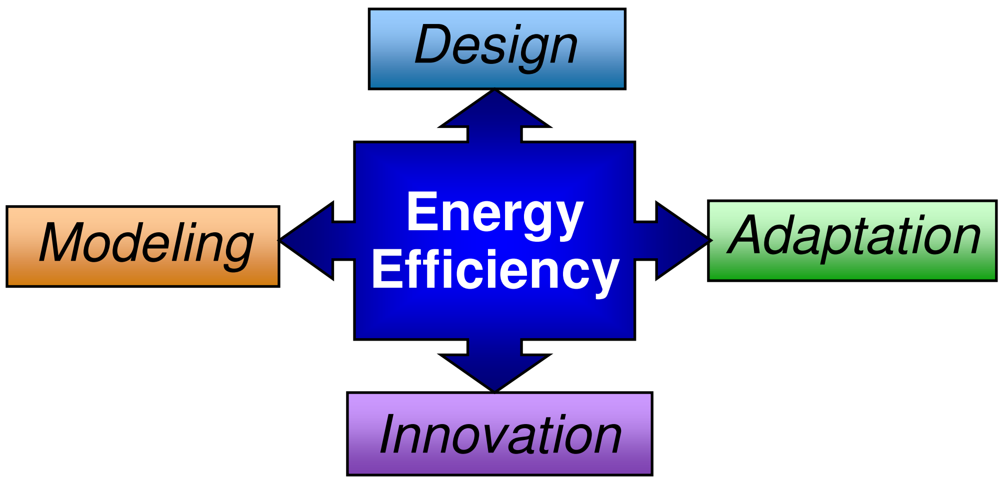


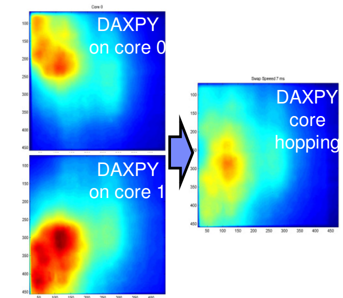


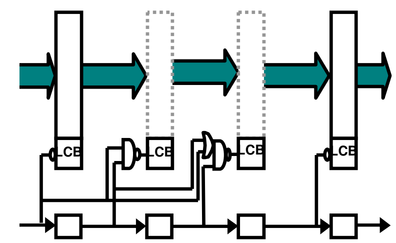




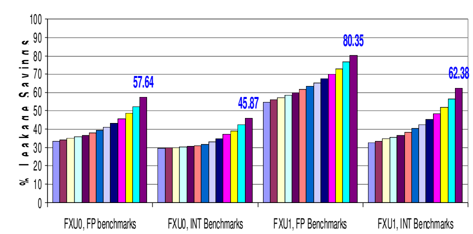


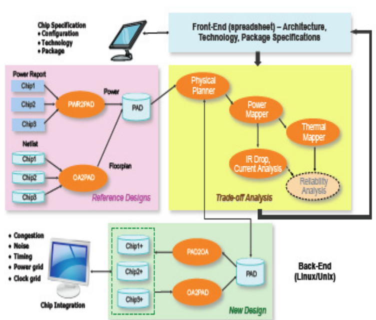


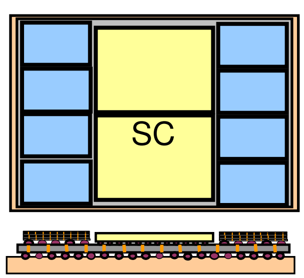


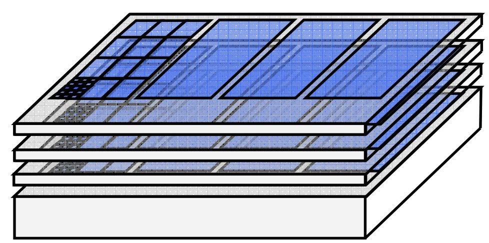


Power gate potential function of break-even point for FXU0 and FXU1 units

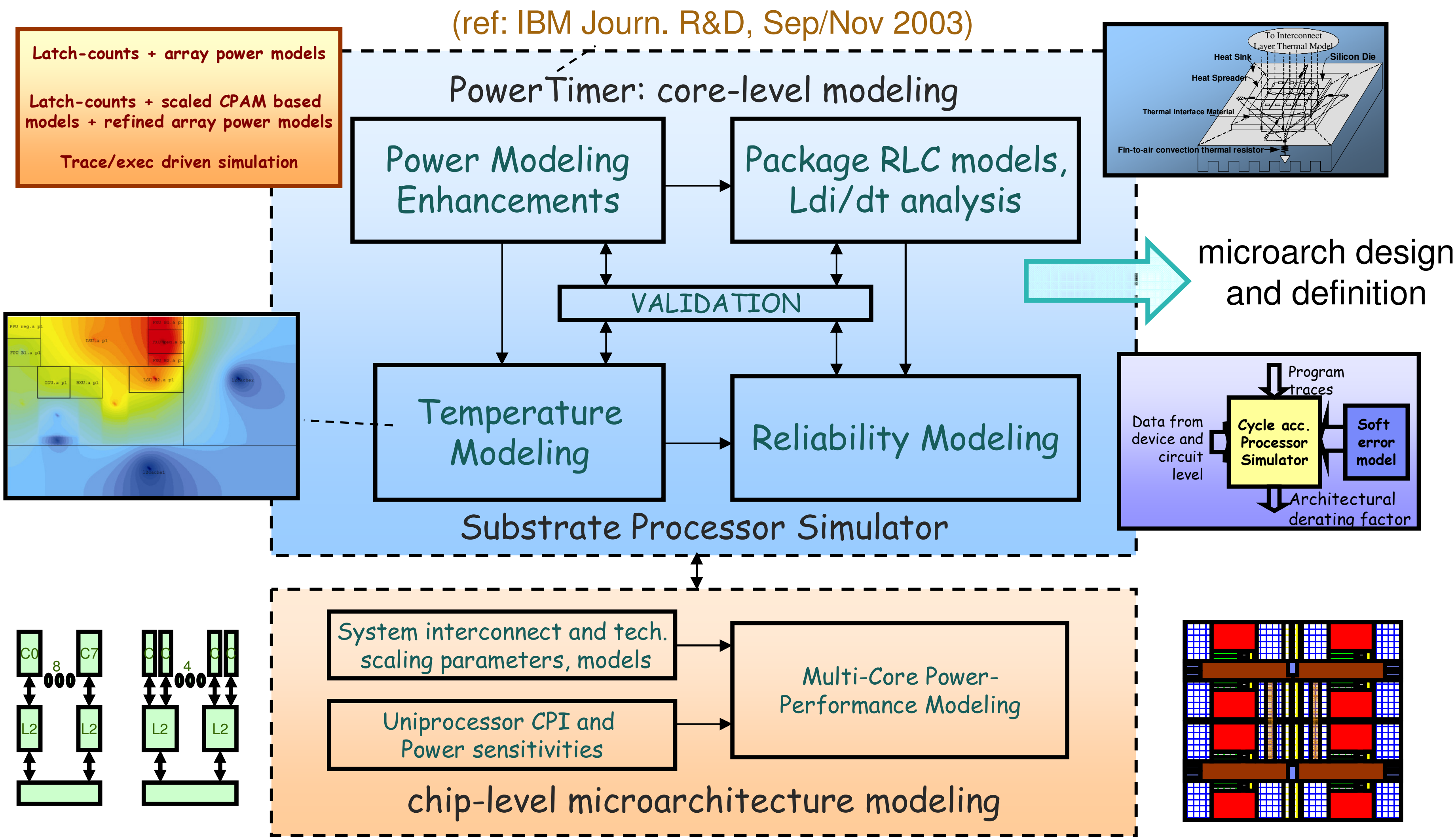




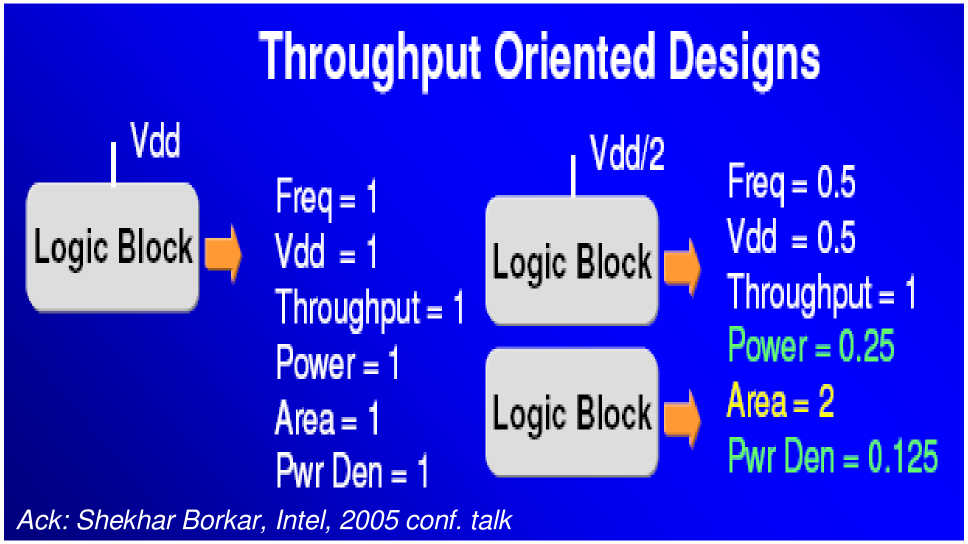




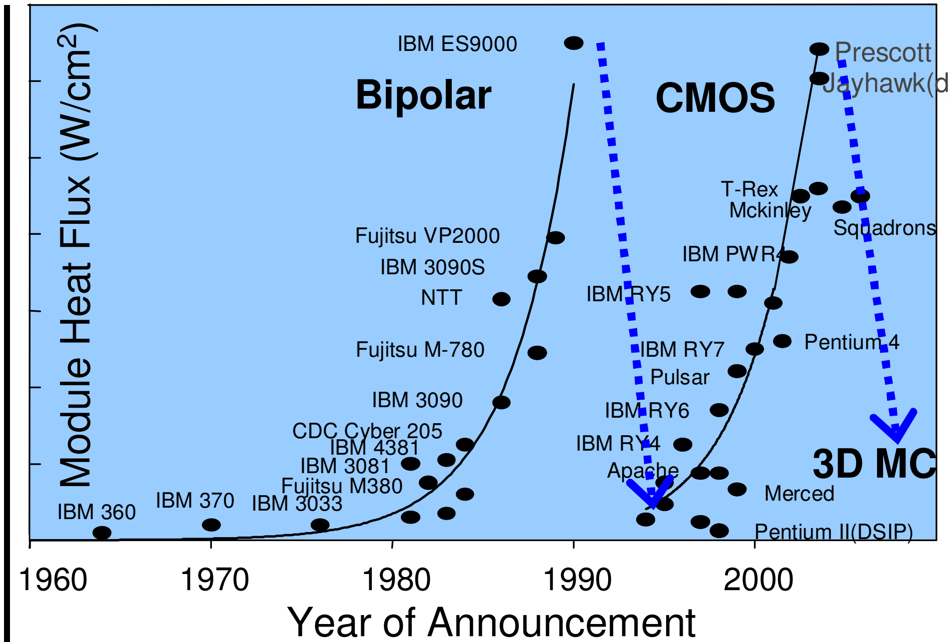
Power Reduction via Design Time Modeling



Active Power Reduction via Concurrency



- **Power = C * V^2 * F**
- **A key principle for high performance in large-scale parallel HPC systems**
- **Cost constraint for exascale-regime systems implies**
  - Manageable number of compute nodes → dozens of cores/chip
- **Also, must not forget the serial** *(Amdahl)* **component of HPC codes!**

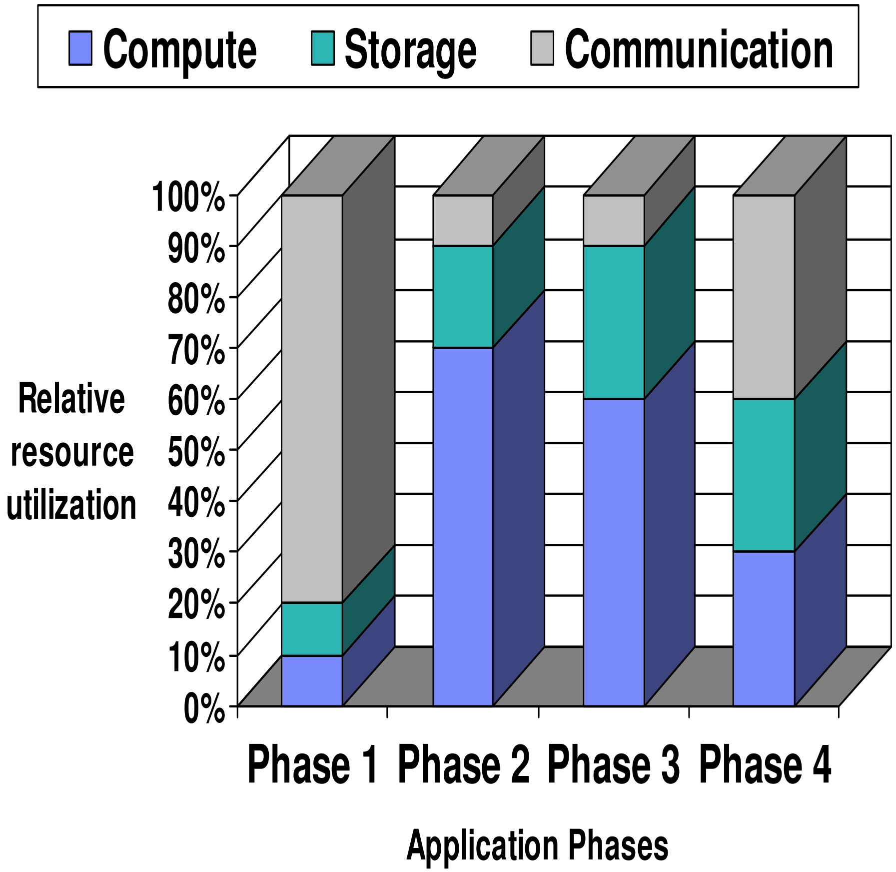


	Bipolar→CMOS	**CMOS→3D**
Power	**0.07x**	**0.1x**
Frequency	**0.3x**	**0.3x**
Density	**50x**	**4-10x**

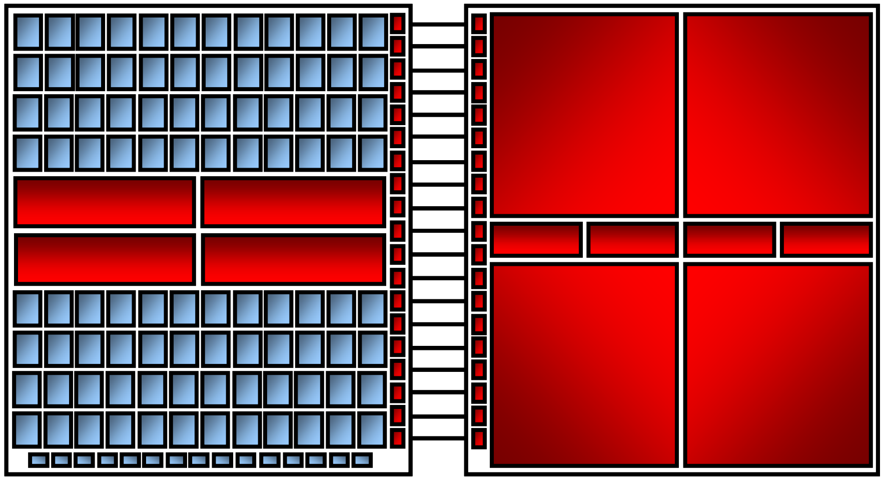
- **3D Many-Core solutions key to extreme concurrency**
  - 4-10 high chip stacks possible with advanced packaging solutions

Power Reduction via Dynamic Resource Management

- **Workloads operate in phases that utilize system resources differently**



- **High power systems require dynamic management capability**
  - *Power Shifting* across compute, communication and storage to avoid power overrun
  - Also provides energy efficiency by powering down unused components



Concluding Remarks

- **The Power Wall is a key impediment to realization of extreme scale computing targets of the future**
  - Extreme scale computing challenged by diminishing performance and power benefits from technology scaling
  - Significant innovation in low power design and dynamic power management required throughout the system
  - Modeling accuracy is more stringent than ever because of the implications of the huge scale of the system
  - Innovation required also in cooling and voltage regulators