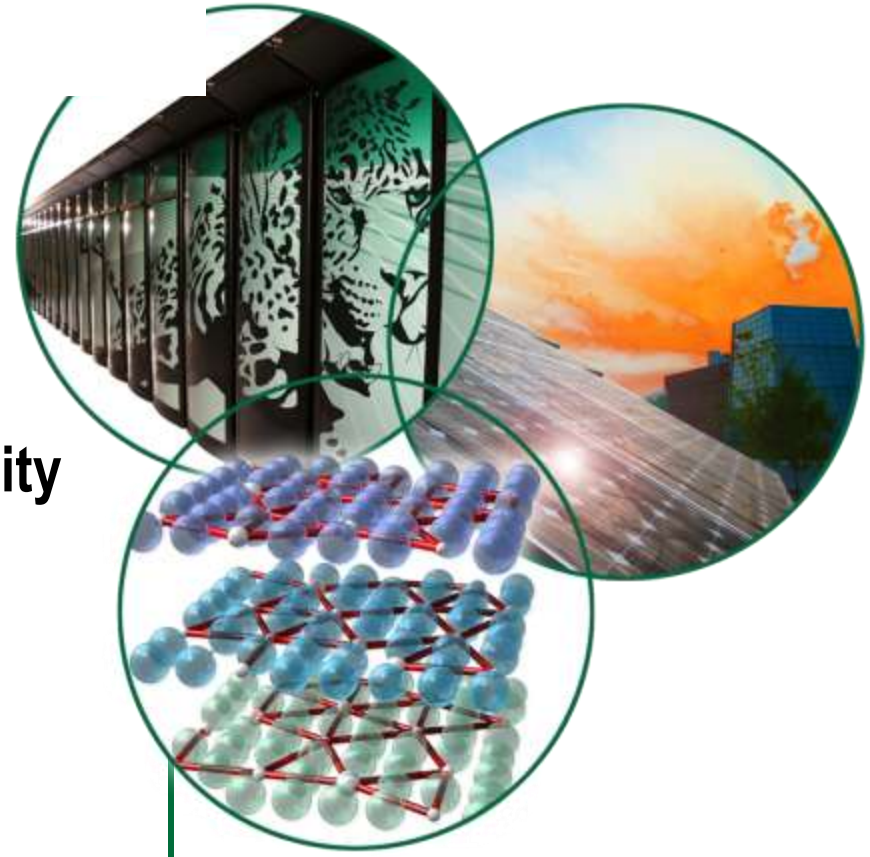
CO-DESIGN CHALLENGES GOING FROM PETASCALE TO EXASCALE



Al Geist

CTO Leadership Computing Facility
Oak Ridge National Laboratory

Boi-molecular Simulations on Future Computing Architectures

Oak Ridge, TN
November 4, 2009

U.S. DEPARTMENT OF ENERGY

Research supported by DOE ASCR

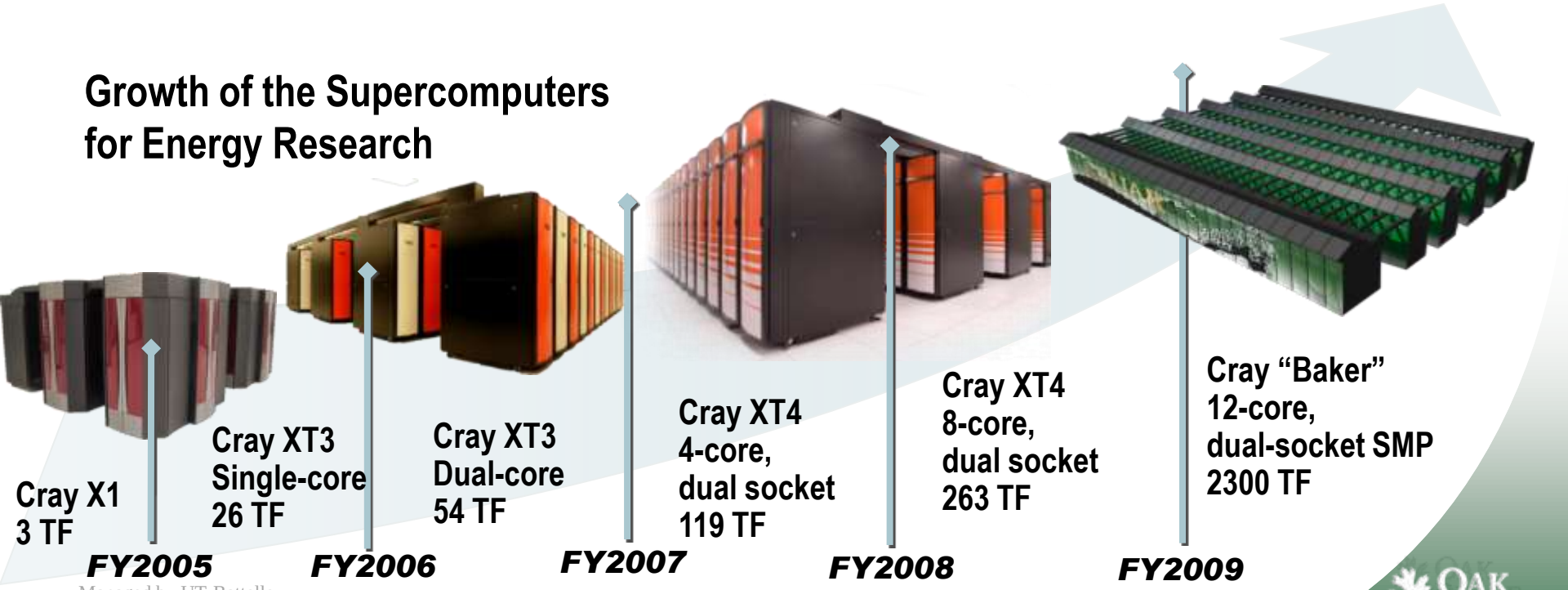
OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Petascale Roadmap

Oak Ridge increased computational capability by almost 1000X in half a decade.

ORNL Leadership Computing Facility successfully executed its petascale roadmap plan on schedule and budget.	Mission: Delivering resources for science breakthroughs. Multiple science applications now running at over a sustained petaflop

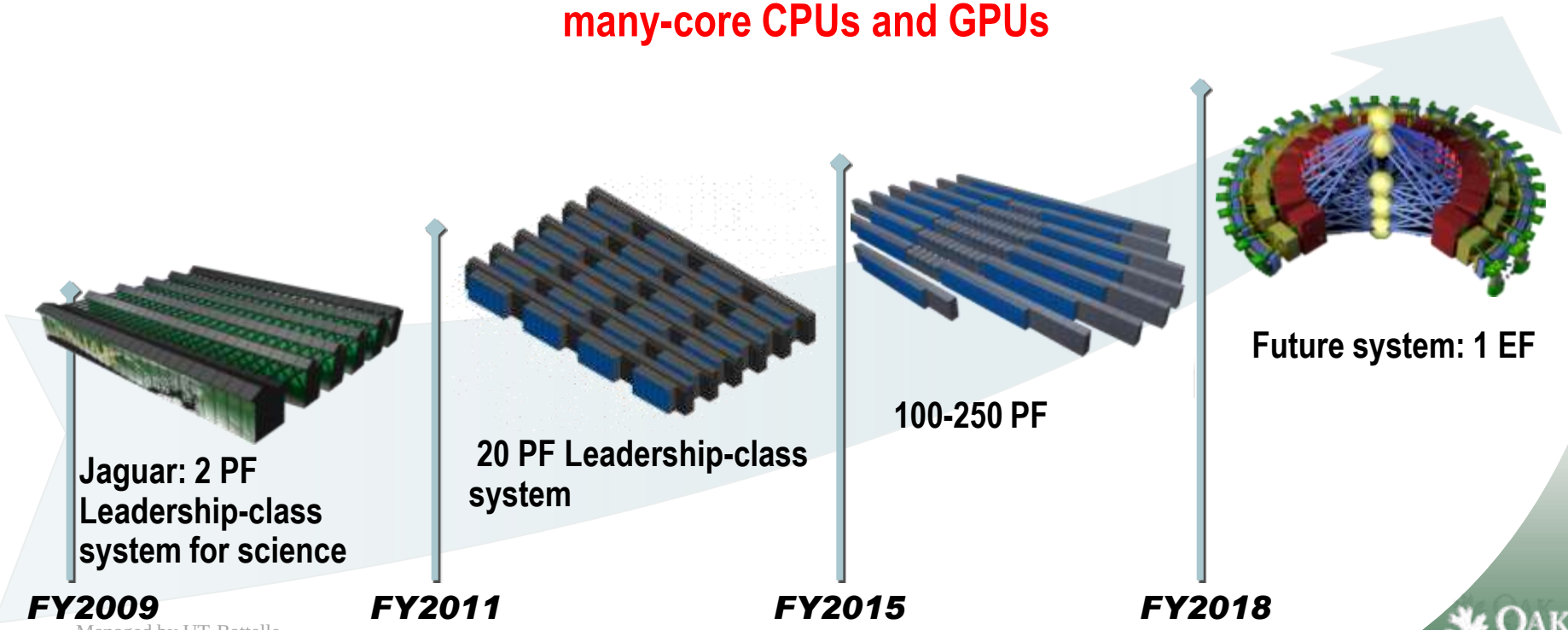
Growth was driven by multi-core sockets and increase in the number of cores per node



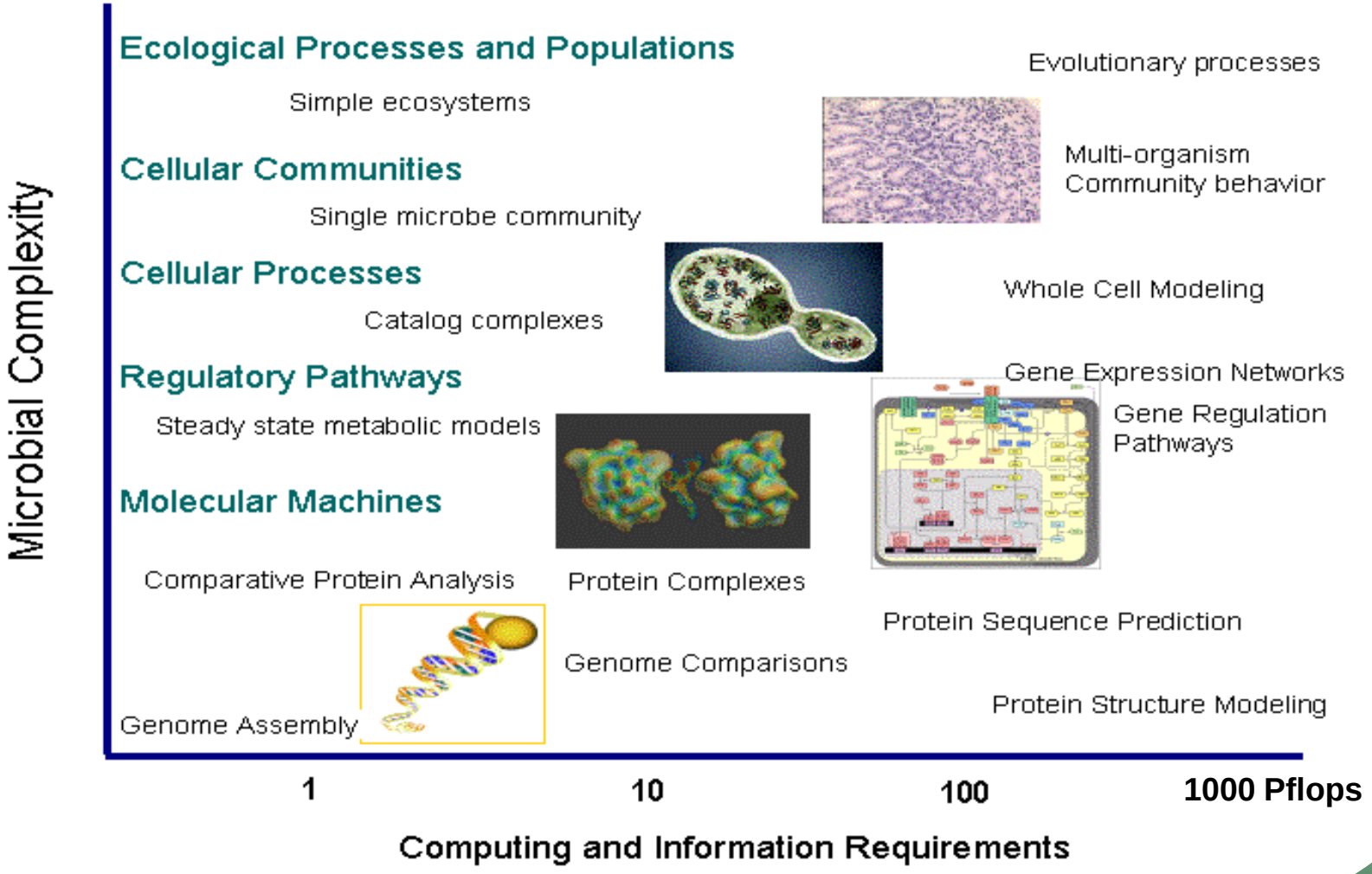
Exascale Roadmap
Delivering the next 1000x capability in a decade

Mission need: Provide the computational resources required to tackle critical national problems	Must also provide the expertise and tools to enable science teams to productively utilize exascale systems

Expectation is that systems will be heterogeneous with nodes composed of many-core CPUs and GPUs



Multiple scales of Biological Complexity and Computational Requirements



Impediments to Useful Exascale Computing

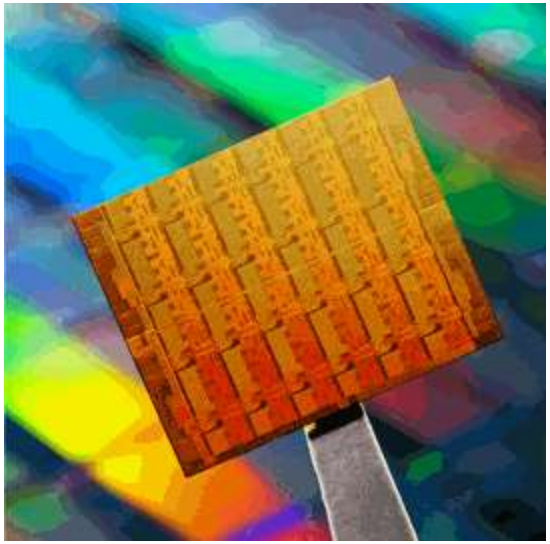
Danger curves ahead

- **Power Consumption**
  - Do Nothing: 100 to 140 MW
- **Scalability**
  - 10,000,000 nodes
  - 100,000,000 cores
  - 1,000,000,000 threads
- **Resilience**
  - Perhaps a harder problem than all the others
  - Do Nothing: an MTBI of 10's of minutes
- **Programming Environment**
  - Data movement and heterogeneous architectures will drive new paradigms
- **Data Movement**
  - **Local**
    - node architectures
    - memory
  - **Remote**
    - Interconnect
    - Link BW
    - Messaging Rate
  - **File I/O**
    - Network Architectures
    - Parallel File Systems
    - Latency and Bandwidth

Power Challenge (20MW for 1 EF) is
Driving Near-term GPU Architectures and longer term Heterogeneous Many-core nodes

- As reference - Jaguar averages 5MW but can peak over 6MW
- Cost of electricity ($1M/ MW-year) over system's lifetime should not exceed the cost of the actual computer.
- This leads to 2018 Exascale system averaging 20MW
- GPUs are presently give the most flops per Watt.
- In 2015 and beyond expect chip manufacturers to start pulling the GPU right onto the CPU chip.

Four Nvidia Fermi in a 1U package

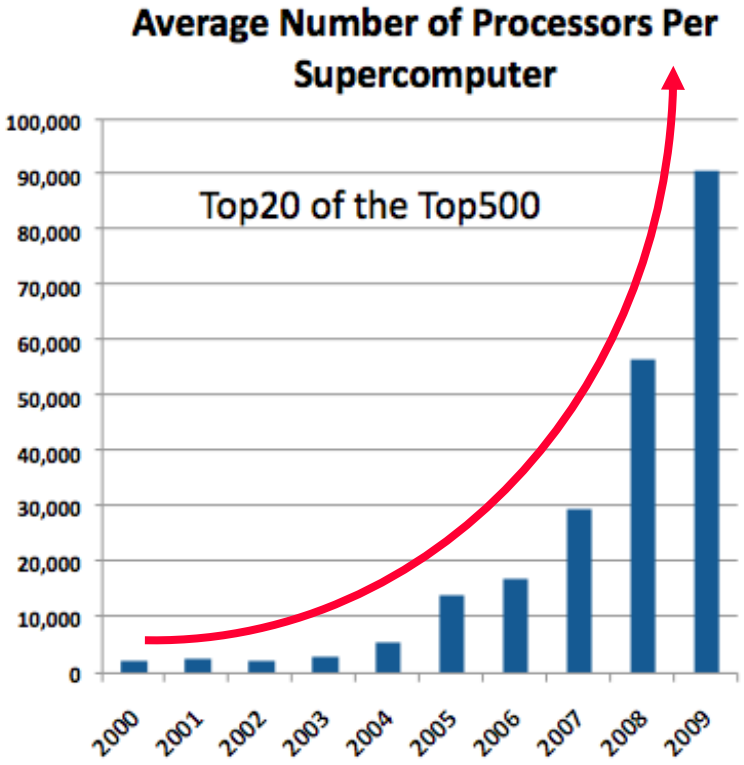


Intel many-core prototype

Challenge: Exponential Growth in Parallelism

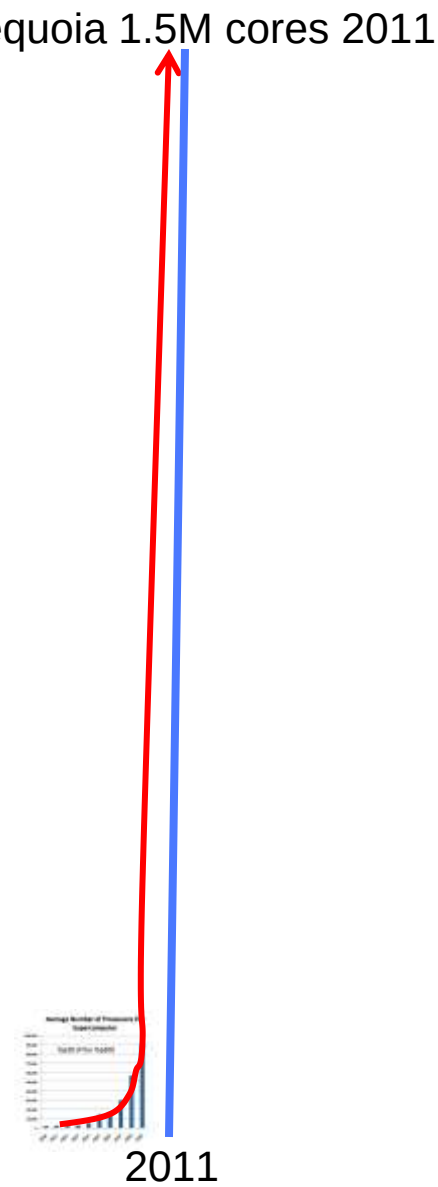
Sequoia 1.5M cores 2011

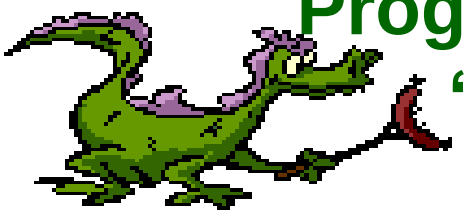
- Power requirements leads to more processors rather than faster ones.
- Fundamental assumptions of today's system software and applications did not anticipate **exponential growth** in parallelism
- Biology applications need to be sure they can scale (and scale rapidly) in the amount of available parallelism.
- Number of components and MTBF leading to a paradigm shift – **Faults will be the norm rather than rare events. SW will have to adapt to frequent failures on the fly.** Checkpoint/restart won't work



Challenge: Exponential Growth in Parallelism

- Power requirements leads to more processors rather than faster ones.
- Fundamental assumptions of today's system software and applications did not anticipate **exponential growth** in parallelism
- Biology applications need to be sure they can scale (and scale rapidly) in the amount of available parallelism.
- Number of components and MTBF leading to a paradigm shift – **Faults will be the norm rather than rare events. SW will have to adapt to frequent failures on the fly.** Checkpoint/restart won't work





Programming Paradigm Shift
"Failure is the Norm"

Factors Driving up the Fault Rate

Number of components both memory and processors will increase by an order of magnitude which will increase hard and soft errors

Smaller circuit sizes, running at **lower voltages** to reduce power consumption, increases the probability of switches flipping spontaneously due to thermal and voltage variations as well as radiation, increasing soft errors

Power management cycling significantly decreases components lifetimes due to The thermal and mechanical stresses

Heterogeneous systems make error detection and recovery even harder, for example, detecting and recovering from an error in a GPU (hundreds of cycles to drain the pipes)

It Already Is

Eg. in 3 days

error msgs

614	inactive link
861	machine check exception
24,215	deadlock timeouts

Yet system stays up for a week at a time

Data Movement on Hybrid CPU/GPU Nodes

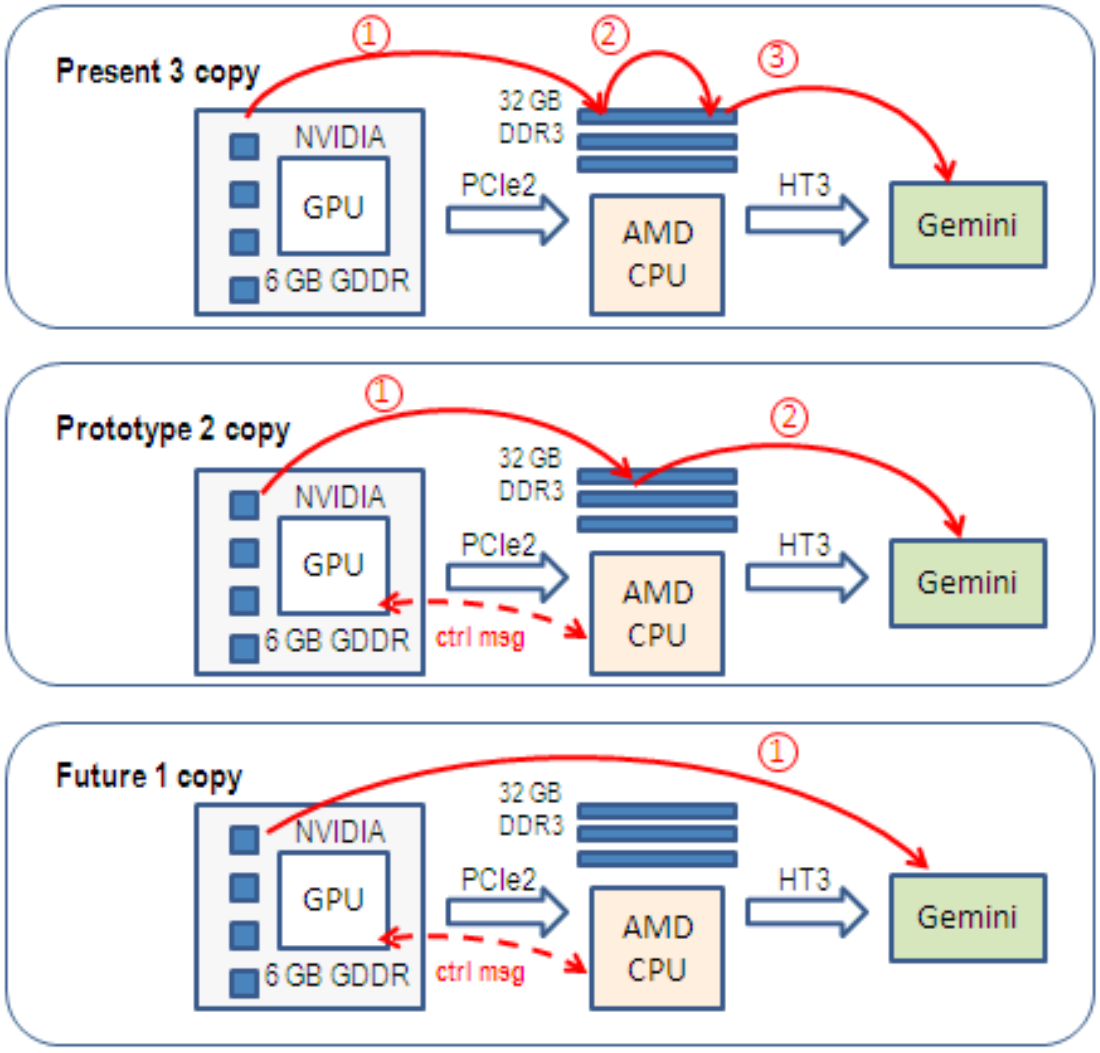
Awareness of Multiple types and banks of memory on the node (GDDR, DDR3)

Sometimes data movement across the node is more complex than expected

Congestion and bottlenecks are a real concern in the links between GPU, CPU, and network

Managing (and minimizing) data movement on a GPU node is critical to performance

Co-design improves awareness



Co-design: What does it mean? How can it help?

Means the **integrated co-design of tools, algorithms, and architectures** to enable more efficient and timely solutions to mission critical problems

Integrated team of math, CS, and bio-molecular experts has expertise and cross fertilization of ideas that increases the chances of working around the challenges.

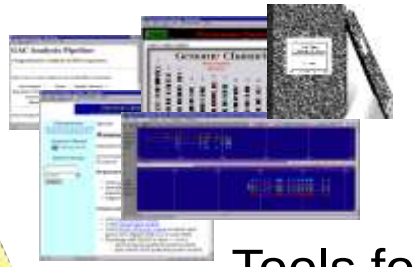
How Co-design helps:

Architecture-aware algorithms and associated runtime enable science applications to better exploit the architectural features of DOE's Leadership systems, **such as GPUs**.

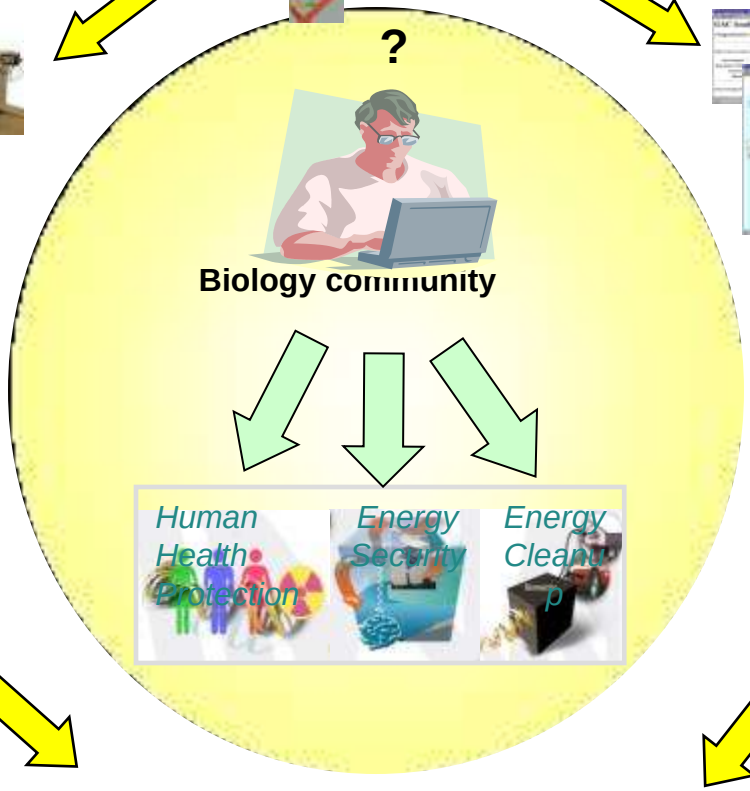
Applications team members immediately incorporate new algorithms providing **near-term high impact on science**

Co-design Example #1: Integrated Tools, Algorithms, and Architectures

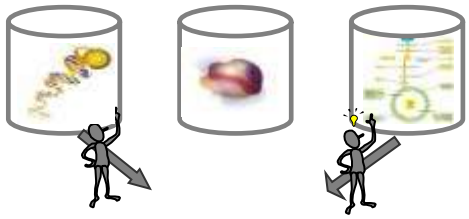
Experiments



Tools for Ease of Use



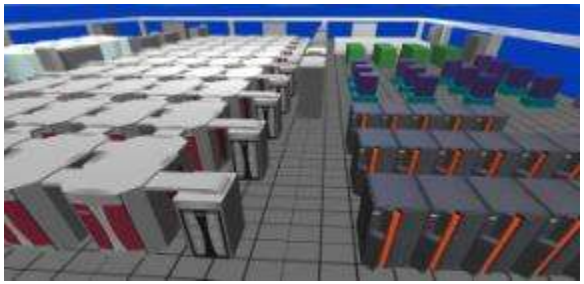
Distributed Databases



Knowledge Discovery

Advanced Computational Algorithms

Simulation & Modeling



Supercomputer Resources

Co-Design Example #2: Integrated Algorithms, Runtime, and Simulation

Architecture Aware Algorithms
Minimize data movement, Exploit GPU

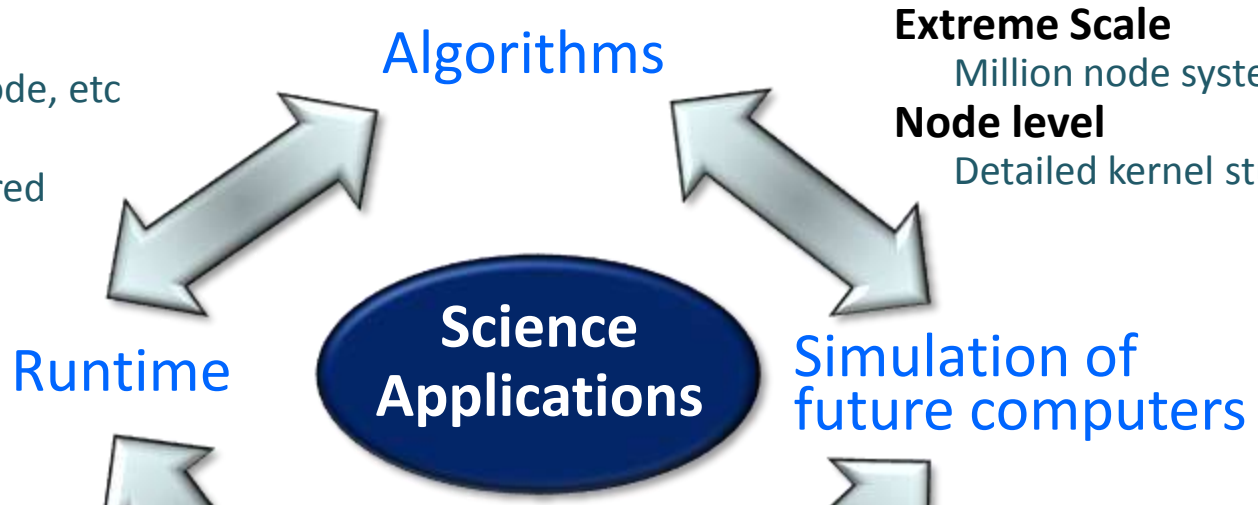
Multi-precision Algorithms
Higher SP performance

Hierarchical MPI
MPI_Comm_Node, etc

Shared memory
MPI_Alloc_Shared

Extreme Scale
Million node systems

Node level
Detailed kernel studies



Multi-core
Processor affinity
Memory affinity
Scheduling
Threading

Memory hierarchy
Future designs

Interconnect
Latency/BW effects

Influence design

Thank You