SAND2009-3764P

Sandia Initiatives in HPC Architectures and Applications

Overview for

Intel Labs GmbH

Nikolaus Lange, Jan Uerpmann and Team
Braunschweig, Germany

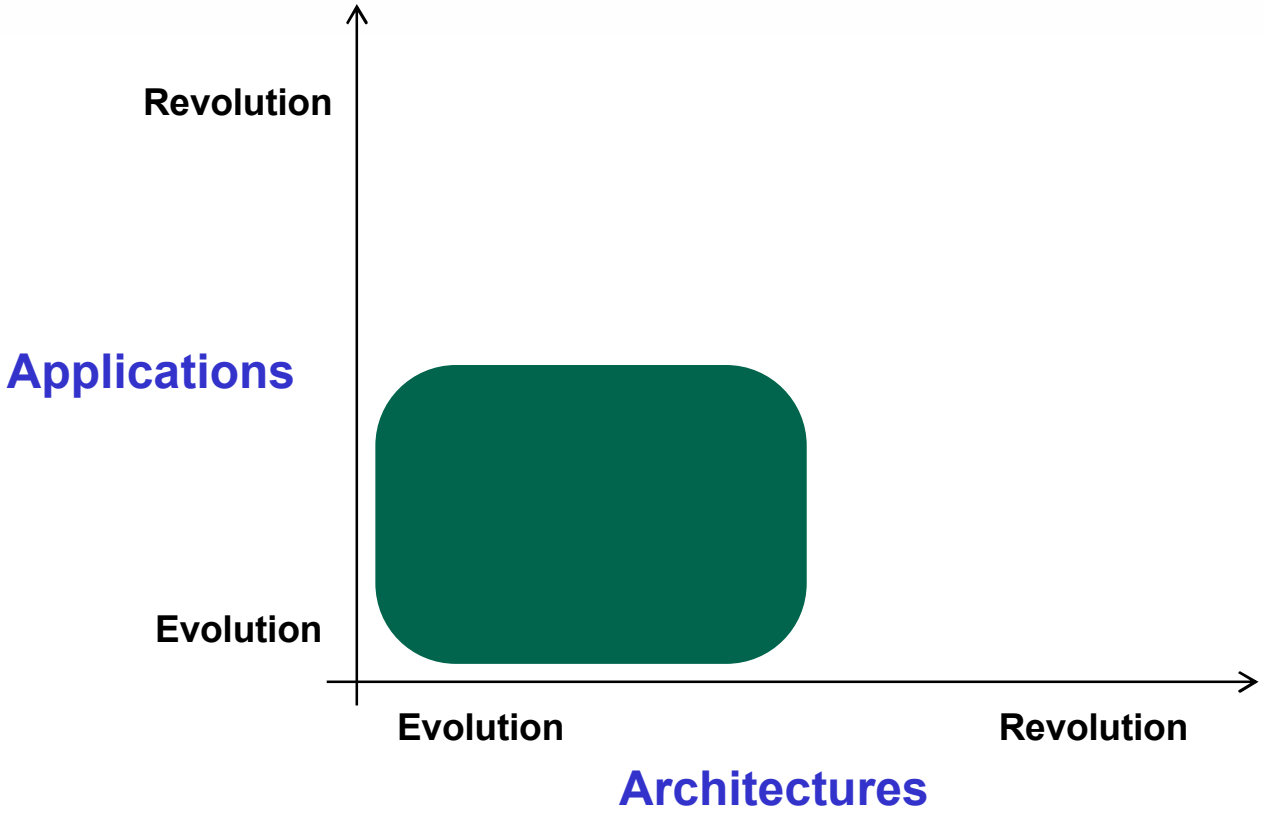
June 22, 2009

James A. Ang, Ph.D. & Manager
Scalable Computer Architectures
Computation, Computers, and Mathematics Center

Sandia is a Multiprogram Laboratory Operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy Under Contract DE-ACO4-94AL85000.

Sandia National Laboratories

Conceptual View of the Application/Architecture Domain Space



- For the last ~15-20 years HPC applications and architectures have benefitted from a period of remarkable stability and evolution

Sandia National Laboratories

Red Storm (2006)

True MPP, designed to be a single system
- Full 3-D mesh interconnect
- 12,960 compute nodes (Dual core AMD Opterons @ 2.4 GHz)
- 39.2 Terabytes of memory
- 400 Terabytes of disk storage

Sandia contributions included
- Overall System Architecture Design
- Collaboration to design interconnect
- Development of compute node operating system based on Sandia's light weight kernel technology

Diverse set of Tri-lab problems are solved on Red Storm

Promising early performance
- 101.4 Teraflops on Linpack (approximately 80.0% of peak)

*Together with **Sandia National Laboratories, who partnered with Cray in designing the 'Red Storm' architecture**, we are very excited that PSC has selected 'Red Storm' for their very diverse and demanding scientific supercomputing workload.*

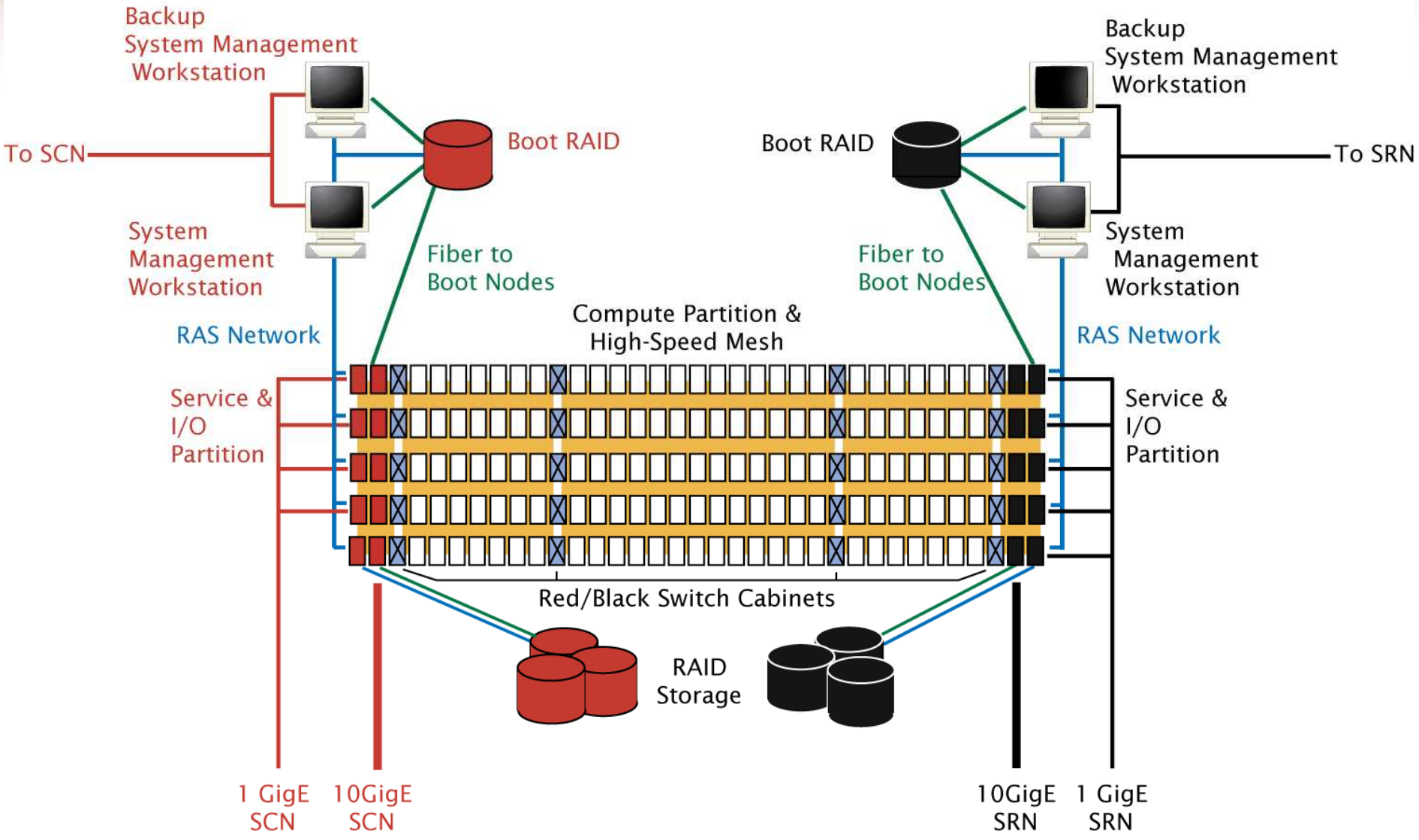
• **Peter Ungaro, Cray President and CEO**
[http://www.hpcwire.com/hpcwire/hpcwireWWW/04/0507/107608.html]

*Many HPC systems today are designed to excel on peak performance and Linpack numbers that are poor predictors of actual problem-solving performance on many end user applications. Like the Cray T3E before it, the new **Cray XT3 is designed for high performance on large-scale customer HPC applications and workloads**.*

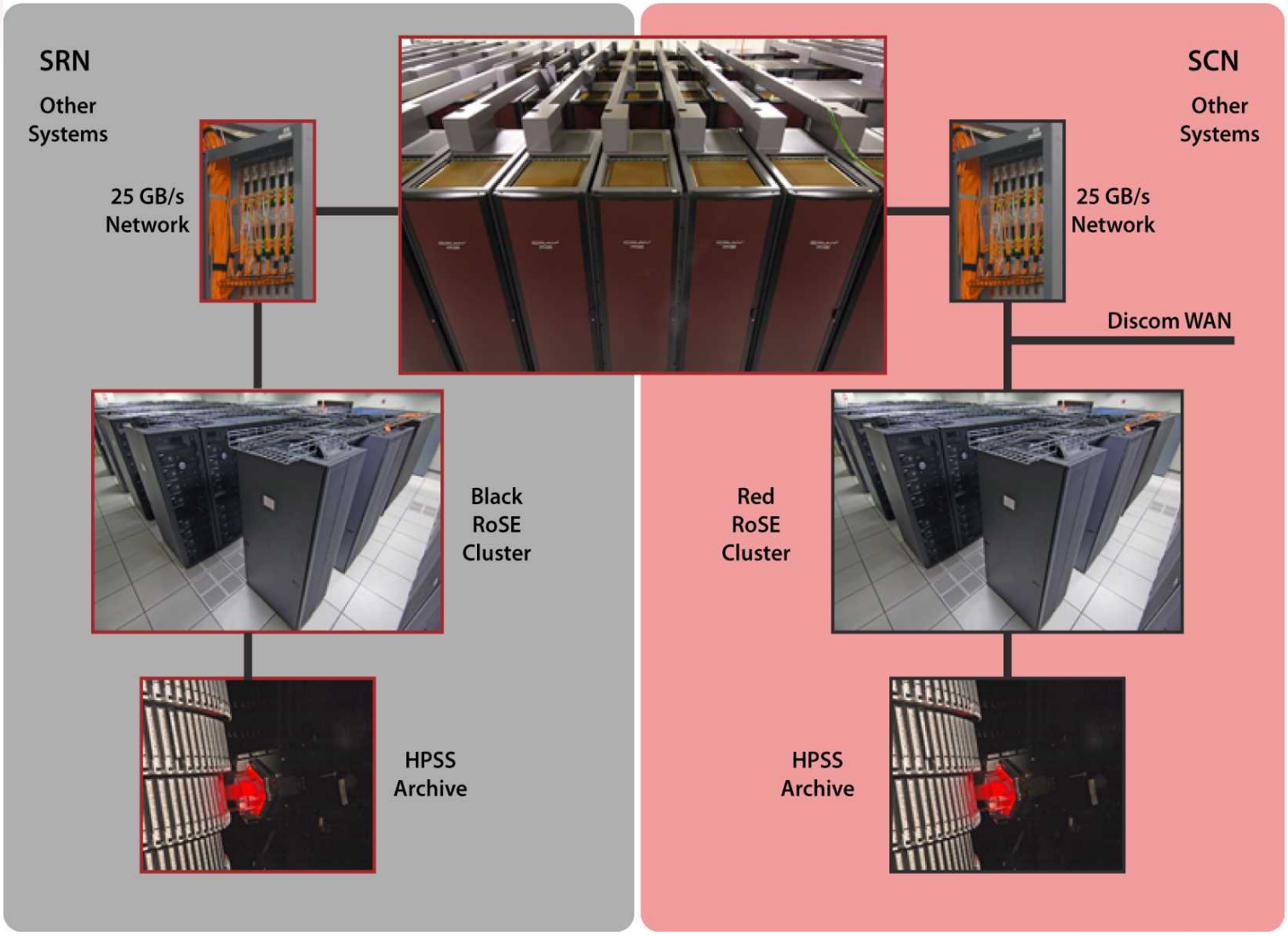
• **Earl Joseph, IDC Program Vice President**
[http://www.cray.com/products/xt3/index.html]

Sandia National Laboratories

Red Storm Layout



Architected Red Storm Environment



Sandia National Laboratories

Red Storm Configurations

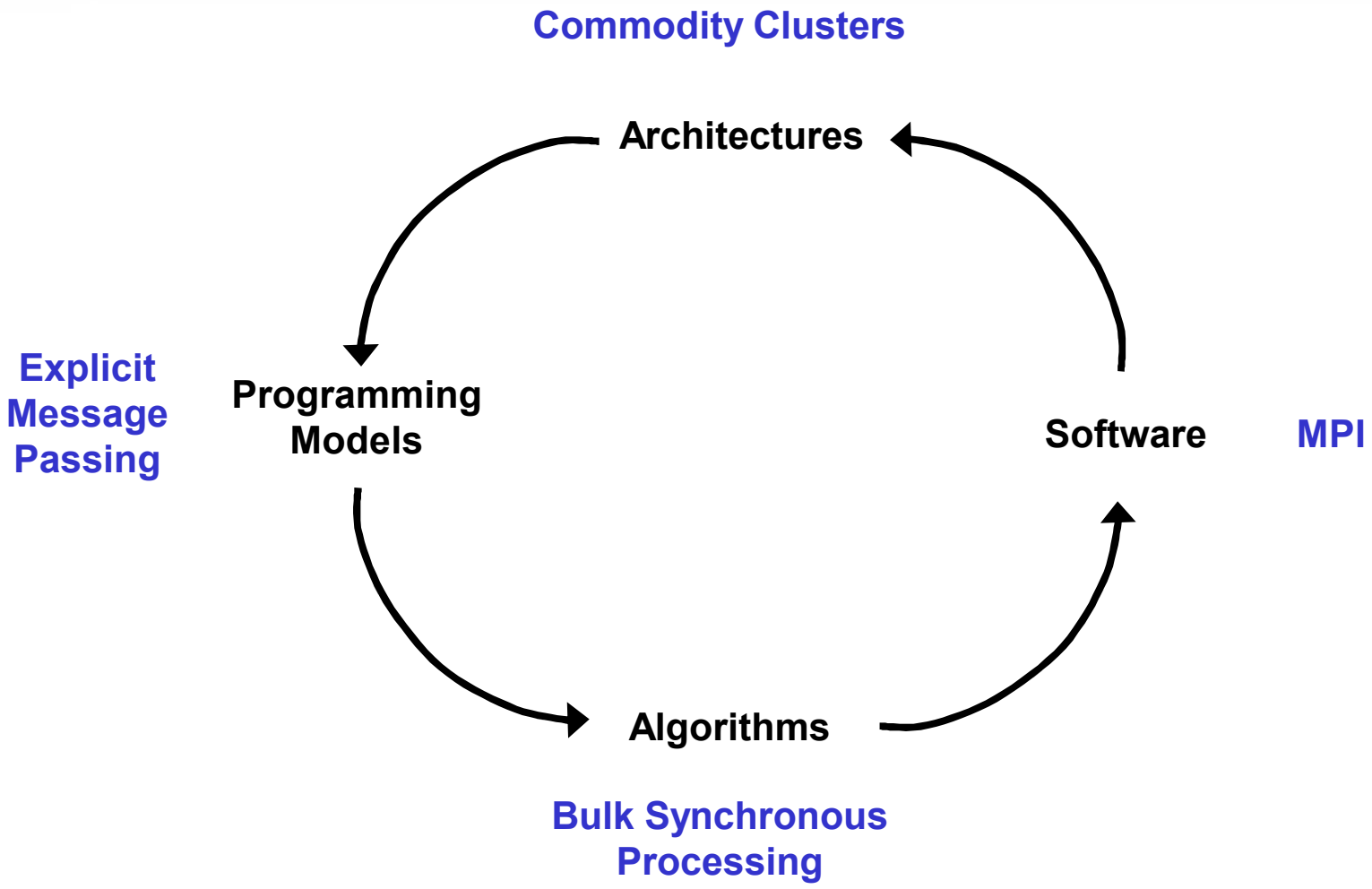
Operational Year	2005	2006	2008
Theoretical Peak (TF)	41.47	124.42	284.16
HPL Performance (TF)	36.19 on 10,880 processors	101.4 on 26,544 processors	204.2 on 38,208 processors
# compute cores	10,368	25,920 (12,960 nodes)	38,400 (12,960 nodes)
Processor	AMD Opteron™ @ 2.0 GHz	AMD dual core Opteron™ @ 2.4 GHz	6720 AMD dual-core Opteron™ @ 2.4 GHz 6240 AMD quad-core Opteron™ @ 2.2 GHz
Total Memory	33.38 TB	39.19 TB	78.75 TB
System Memory B/W	57.97 TB/s	78.12 TB/s	122 TB/s
User Disk Storage	340 TB	340 TB	1753 TB
Min Bi-section B/W	3.69 TB/s	4.61 TB/s	4.61 TB/s
System Foot Print	~3000 sq ft	~3500 sq ft	~3500 sq ft
Power Requirement	1.7 MW	< 2.5 MW	2.5 MW

Enablers for Current Successes

- Clusters
  - "Killer micros" enable commodity-based parallel computing
  - Attractive price and price/performance
  - Stable model for algorithms & software
- MPI
  - Portable and stable programming model and language
  - Allowed for huge investment in software
- Bulk-Synchronous Processing (BSP)
  - Basic approach to almost all successful MPI programs
  - Compute locally; communicate; repeat
  - Excellent match for clusters + MPI
  - Good fit for many scientific applications
- Algorithms
  - Stability of the above allows for sustained algorithmic research
  - Key advances include domain decomposition and partitioning algorithms

Sandia National Laboratories

A Virtuous Circle…



…but also a suffocating embrace

Sandia National Laboratories

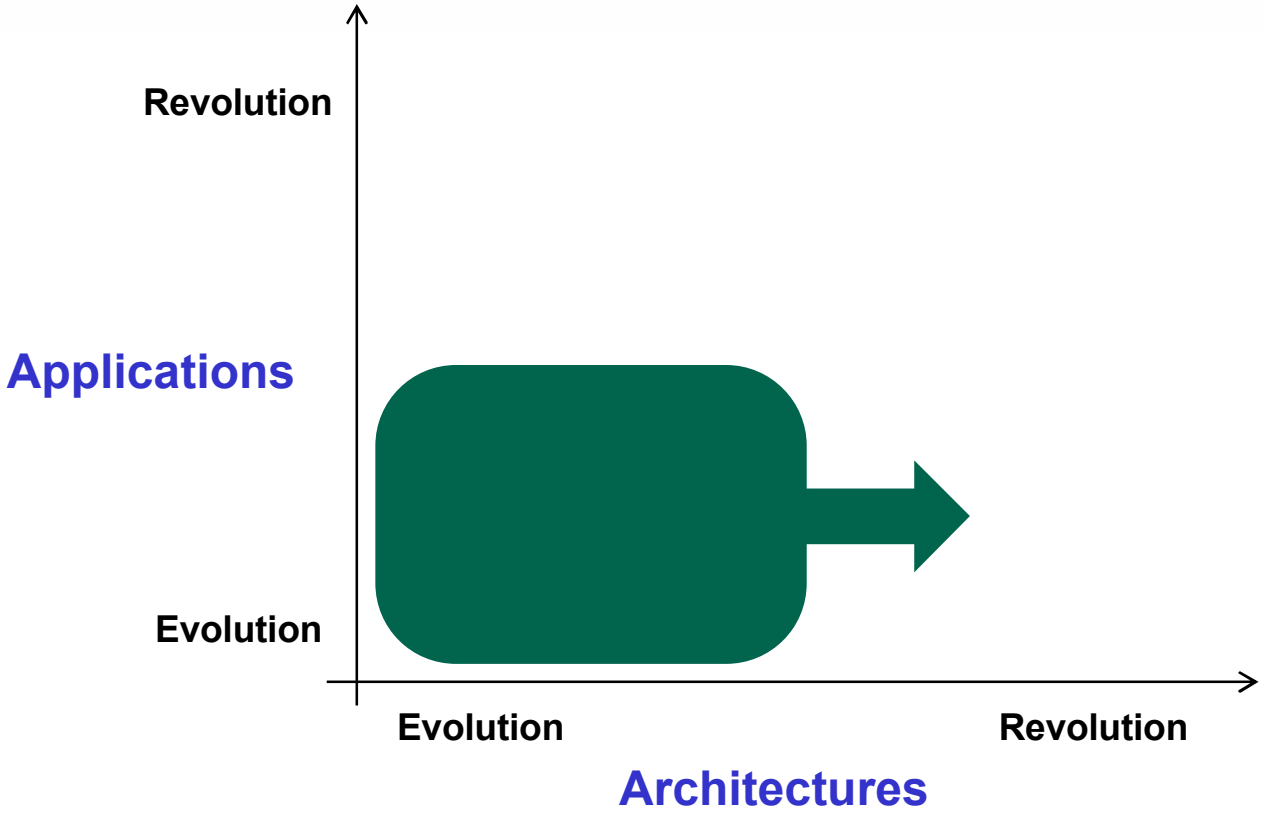
Applications Are Evolving

- **Leading edge scientific applications increasingly include:**
  - **Adaptive, unstructured data structures**
  - **Complex, multiphysics simulations**
  - **Multiscale computations in space and time**
  - **Complex synchronizations (e.g. discrete events)**

- **These raise significant parallelization challenges**
  - **Limited by memory, not processor performance**
  - **Unsolved micro-load balancing problems**
  - **Finite degree of coarse-grained parallelism**
  - **Bulk synchronous processing not always appropriate**

- **These evolutionary changes will stress existing approaches to parallelism**

Conceptual View of the Application/Architecture Domain Space



- Architectures are undergoing a revolutionary change

Sandia National Laboratories

Industry Trends

Existing industry trends not going to meet HPC application needs

Semi-conductor industry trends

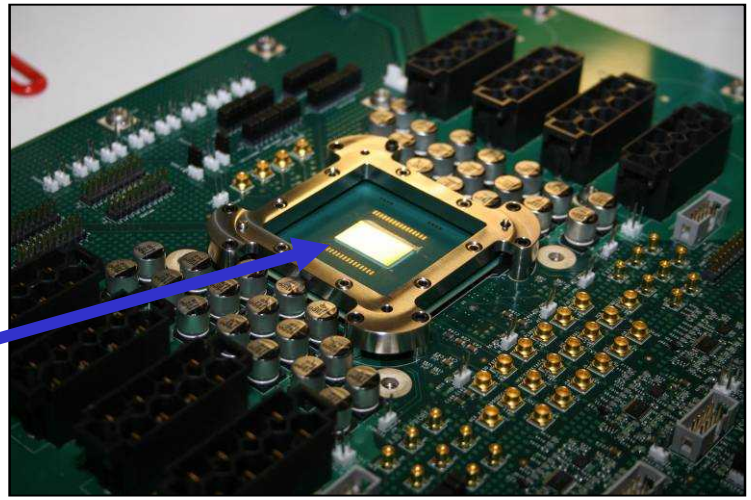
- Moore's Law still holds, but clock speed now constrained by power and cooling limits
- Processors are shifting to multi/many core with attendant parallelism
- Compute nodes with added hardware accelerators are introducing additional complexity of heterogeneous architectures
- Processor cost is increasingly driven by pins and packaging, which means the memory wall is growing in proportion to the number of cores on a processor socket

Development of large-scale Leadership-class supercomputers from commodity computer components requires collaboration

- Supercomputer architectures must be designed with an understanding of the applications they are intended to run
- Harder to integrate commodity components into a large scale massively parallel supercomputer architecture that performs well on full scale real applications
- Leadership-class supercomputers cannot be built from only commodity components

Moore's Law + Multicore → Rapid Growth in Computing Power

2007 - Intel Polaris R&D Processor
1 TeraFLOPs on a chip
• 275 mm^2 & 62 W



1997 - Intel ASCI Red
1 TeraFLOPs in a house
• 2,500 ft^2 & 500,000 W

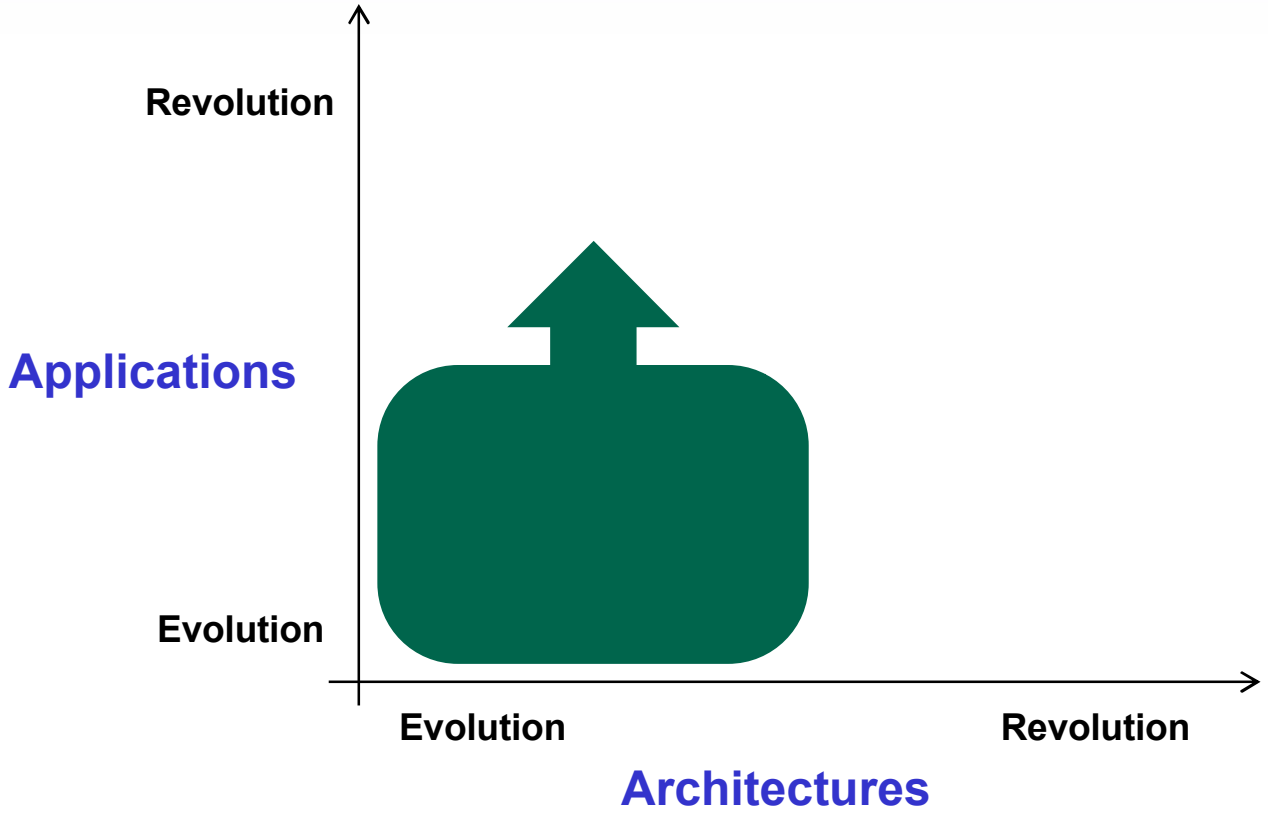
Sandia National Laboratories

A Renaissance in Architecture Research

- Good news
  - Moore's Law marches on
  - Real estate on a chip is essentially free
    - Major paradigm change – huge opportunity for innovation
- Bad news
  - Power considerations limit the improvement in clock speed
  - Parallelism is only viable route to improve performance
- Current response, multicore processors
  - Computation/Communication ratio will get worse
    - Makes life harder for applications
- Long-term consequences unclear

Sandia National Laboratories

Conceptual View of the Application/Architecture Domain Space



- Our Goal: computational science applications expand into new domains

Sandia National Laboratories

Revolutionary Applications

- **What is "Computational Science"?**
- **We often equate it with modeling and simulation.**
  - **But this is unnecessarily limited.**
- **From Dictionary.com:**
  - **sci·ence – *(noun)* A branch of knowledge or study dealing with a body of facts or truths systematically arranged and showing the operation of general laws.**
  - **com·pu·ta·tion·al (*adjective)* Of or involving computation or computers.**

Sandia National Laboratories

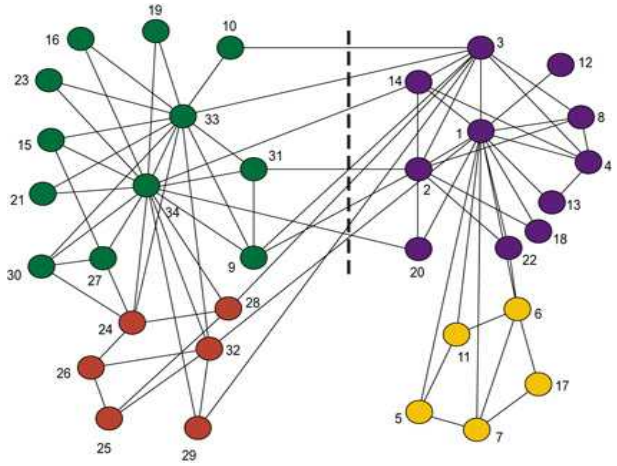
Emerging Uses of Computing in Science

- Science is increasingly data-centric
  - Biology, astrophysics, particle physics, earth science
  - Social sciences
  - Experimental, computational and literature data
  - Validation of Science and Engineering Modeling & Simulation results with quantitative comparisons with experimental measurements
- Sophisticated computing often required to extract knowledge from this data
- Computing challenges are different from mod/sim
  - Data sets can be huge (I/O is a priority)
  - Response time may be short (throughput is key metric)
  - Computational kernels have different character
- What abstractions, paradigms and algorithms are needed?

Example: Network Science

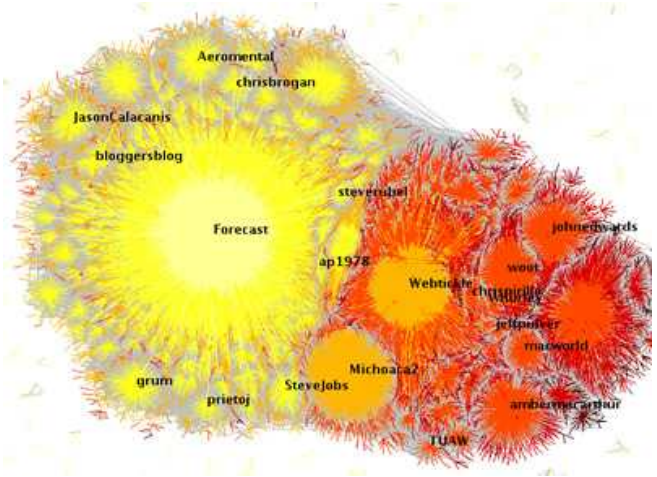
- **Graphs are ideal for representing entities and relationships**
- **Rapidly growing use in biological, social, environmental, and other sciences**

The way it was …



Zachary's karate club (|V|=34)

The way it is now …



Twitter social network (|V|≈200M)

Sandia National Laboratories

Emerging New Scientific Questions

- **New algorithms**
  - Community detection, centrality, graph generation, etc.
  - Right set of questions and concepts still unknown
- **New issues**
  - Noisy, error-filled data. What can we conclude robustly?
  - *Semantic* graphs with edges and vertices of different types
  - Temporal evolution of networks
- **New science**
  - Social dynamics and ties to technology & media
  - Large economic, social, political consequences

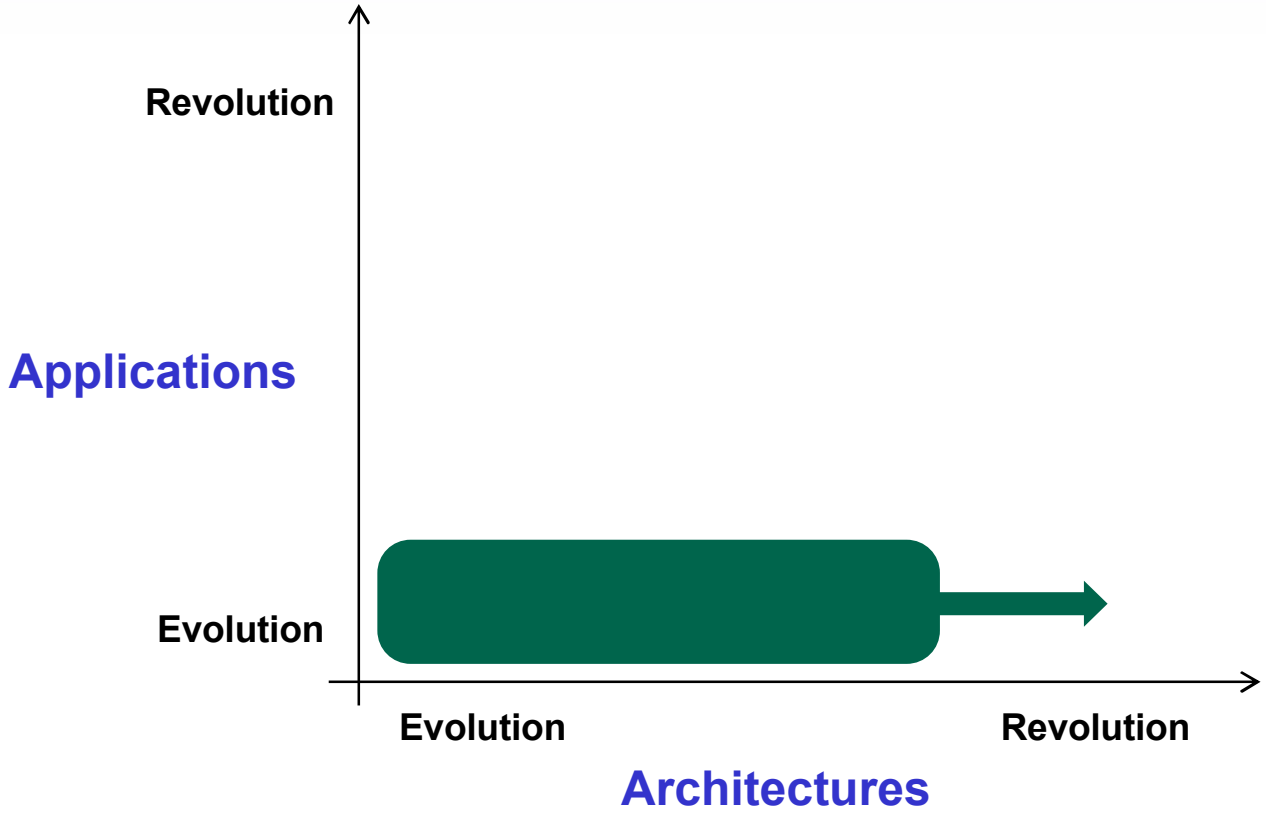
Sandia National Laboratories

Computational Challenges for Network Science

- Minimal computation to hide access time
- Runtime is dominated by latency
  - Random accesses to global address space
  - Parallelism is very fine grained and dynamic
- Access pattern is data dependent
  - Prefetching unlikely to help
  - Usually only want small part of cache line
- Potentially abysmal locality at all levels of memory hierarchy
- Many algorithms are not bulk synchronous
- Approaches based on virtuous circle don't work!

Sandia National Laboratories

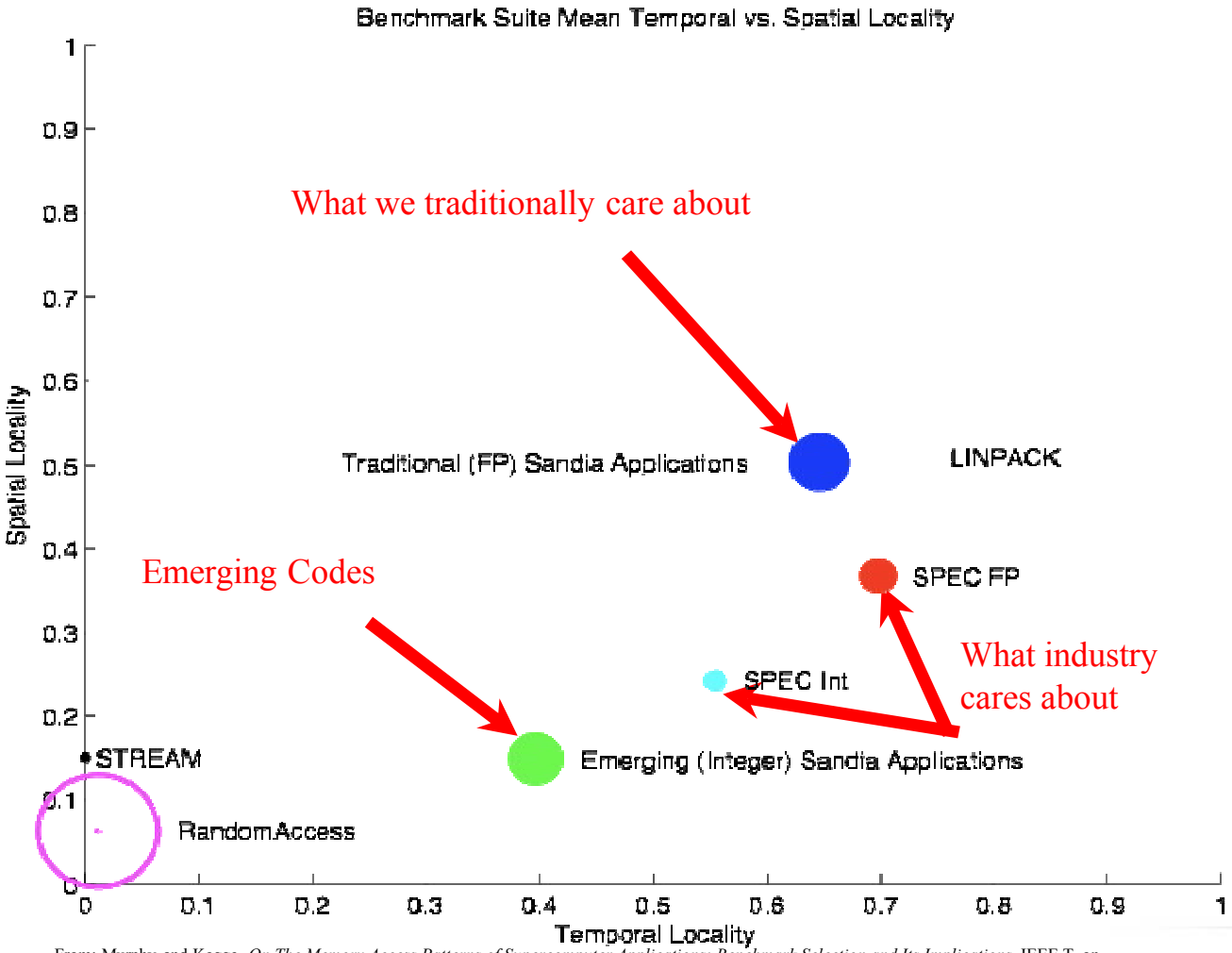
Conceptual View of the Application/Architecture Domain Space



- Our Fear: architectural revolution shrinks the domain of computational science applications

Sandia National Laboratories

Locality Challenges



From: Murphy and Kogge, *On The Memory Access Patterns of Supercomputer Applications: Benchmark Selection and Its Implications*, IEEE T. on Computers, July 2007

Sandia National Laboratories

Exascale Technical Challenges

- **Technical Challenges (from the DARPA ExaScale Computing Study)**
  - **Energy & Power**
  - **Memory & Storage**
  - **Concurrency and Locality**
  - **Resiliency**
- **Solutions will require:**
  - **Co-development and optimization of both hardware and software**
  - **Systems perspective and integration of inter-disciplinary skills**

ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number **FA8650-07-C-7724.** This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings

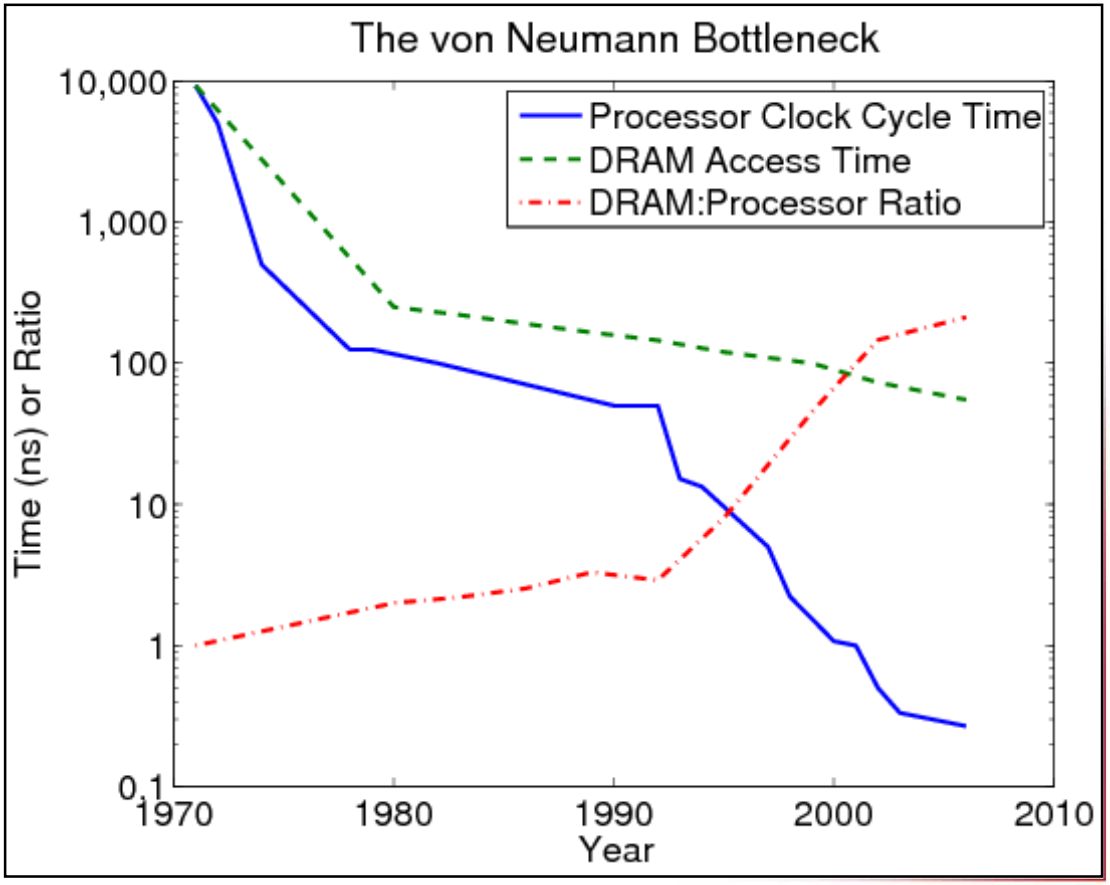
Sandia National Laboratories

The Memory Wall

"FLOPS are 'free'. In most cases we can now compute on the data as fast as we can move it." - Doug Miles, The Portland Group

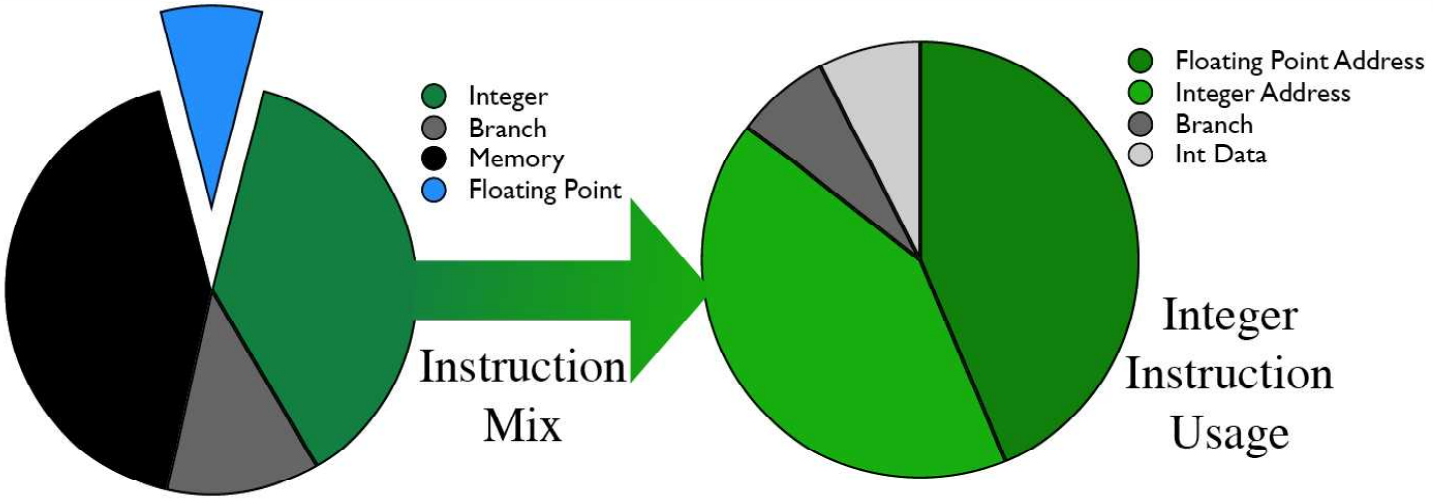
What we observe today:

- Logic transistors are free
- The von Neumann architecture is a bottleneck
- Exponential increases in performance will come from increased concurrency not increased clock rates if the cores are not starved for data or instructions



Sandia National Laboratories

The Memory Wall significantly impacts the performance of our applications

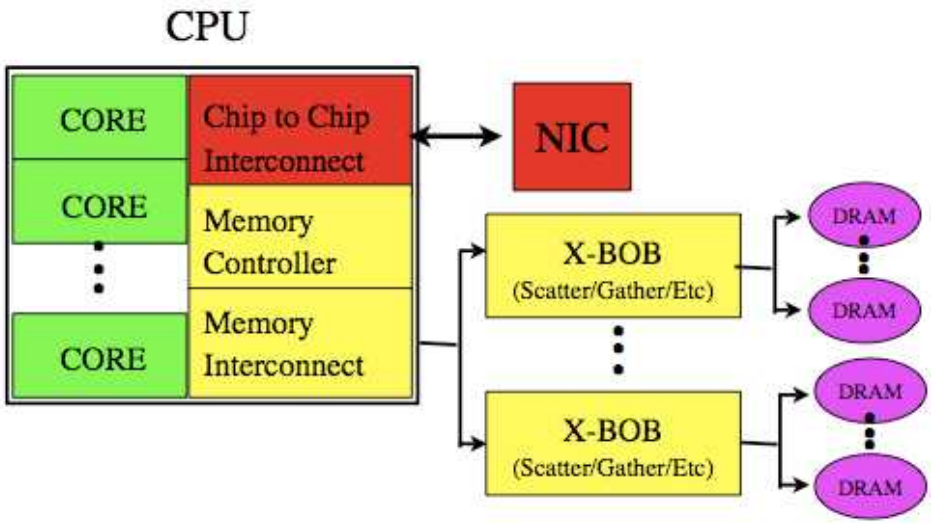


- Most of DOE's Applications (e.g., climate, fusion, shock physics, …) spend most of their instructions accessing memory or doing integer computations, not floating point
- Additionally, most integer computations are computing memory Addresses
- Advanced development efforts are focused on accelerating memory subsystem performance for both scientific and informatics applications

Memory Conceptual Design

Vision: Create a **commodity** memory part with support for HPC data movement operations.

Approach: new high-speed memory signaling technology inserts an ASIC (the Buffer-on-Board, or BOB) between the CPU and memory. Add data movement support in the ASIC.



Near Term Goals:

- Define in-memory operations (scatter/gather, atomic memory operations, etc.)
- Define CPU/X-BOB coherency

Long Term Goals:

- Create a commodity memory part that increases **effective** bandwidth utilization

Sandia National Laboratories

The Consolidated ExaScale Technical Challenge

- **Energy Metric: Simulation Results per Joule**

$$\propto \frac{[\text{Concurrency}]^{\gamma} \times [\text{Resiliency}]^{\rho}}{[\text{Memory Wall}]^{\mu}}$$

Acknowledgements:

- **Red Storm: James Tomkins, Suzanne Kelly, Robert Ballance**
- **Description of the *Virtuous Circle/Suffocating Embrace*, and Revolutionary Applications: Bruce Hendrickson**
- **Locality Challenges: Richard Murphy**
- **Application/Memory Analysis: Arun Rodrigues and Richard Murphy**

Sandia National Laboratories