Data Analytics & High Performance Computing: When Worlds Collide

Bruce Hendrickson

Senior Manager for Math & Computer Science
Sandia National Laboratories, Albuquerque, NM

University of New Mexico, Computer Science Dept.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

What’s Left to Say!?

Worlds Apart

	High Performance Computing	Data Analytics
Programming Model	MPI	SQL / MapReduce
Performance Metric	Single Application Runtime	Throughput
Performance Limiter	Processor	Memory System
Execution Model	Batch	Interactive
Architecture Driver	Performance	Resilience
Data Volumes	Small in, Large out	Large in, Small out
…		

Outline

- **Today's HPC landscape**

- **HPC Applications are changing**
  - Evolution
  - Revolution

- **Architectures are changing**
  - Evolution
  - Revolution

- **Conclusions:**
  - Organic forces will make HPC more data friendly
  - External forces will make HPC more data-centric

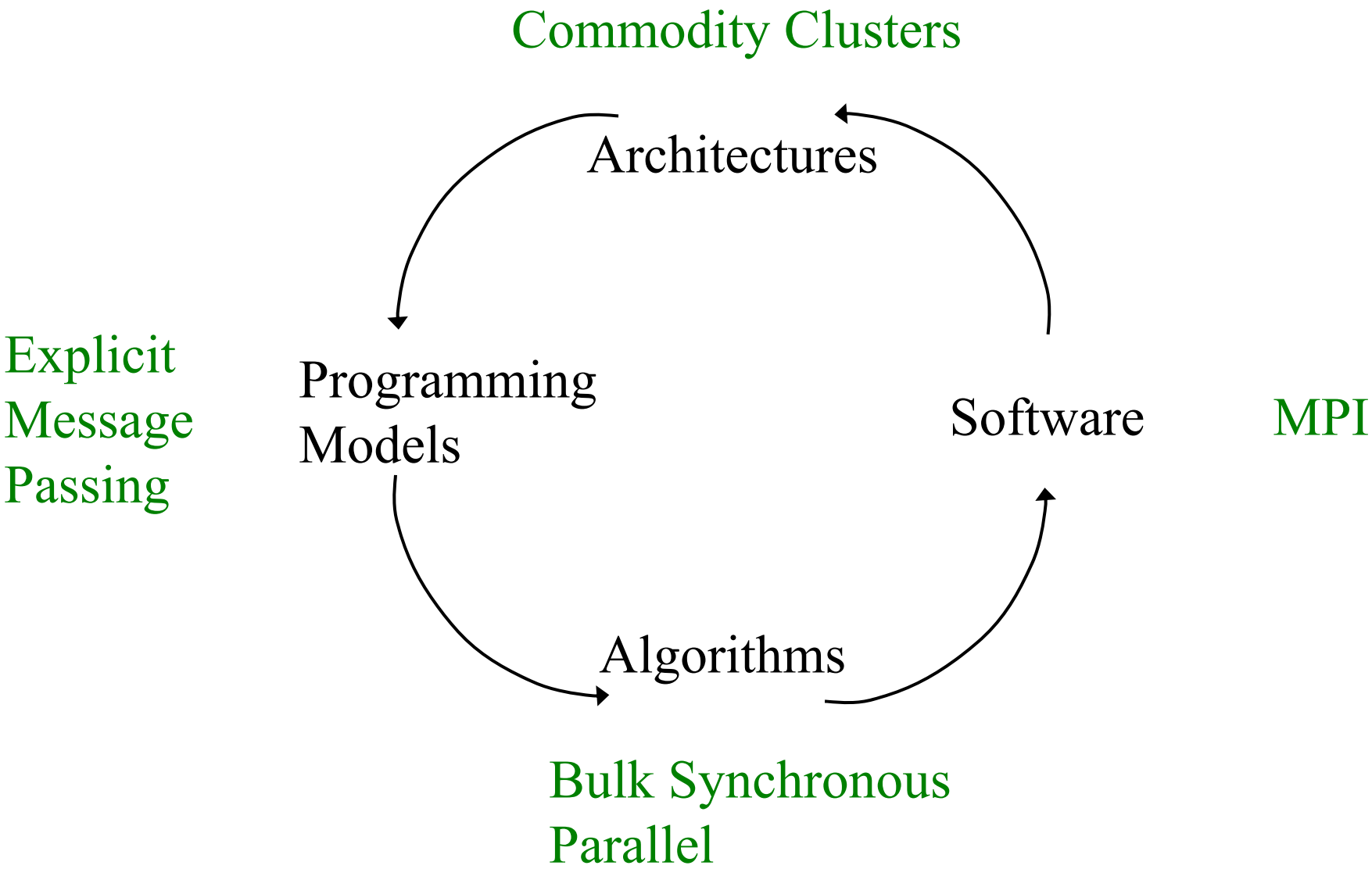
Sandia National Laboratories

Enablers for Mainstream HPC

- **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 Parallel Programming (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

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 parallel not always appropriate

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

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

- **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

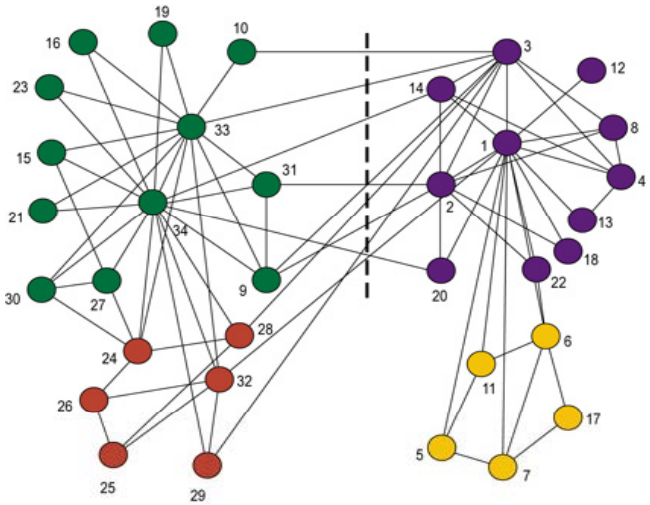
- **What abstractions, paradigms and algorithms are needed?**

Example: Network Science

- **Graphs are ideal for representing entities and relationships**

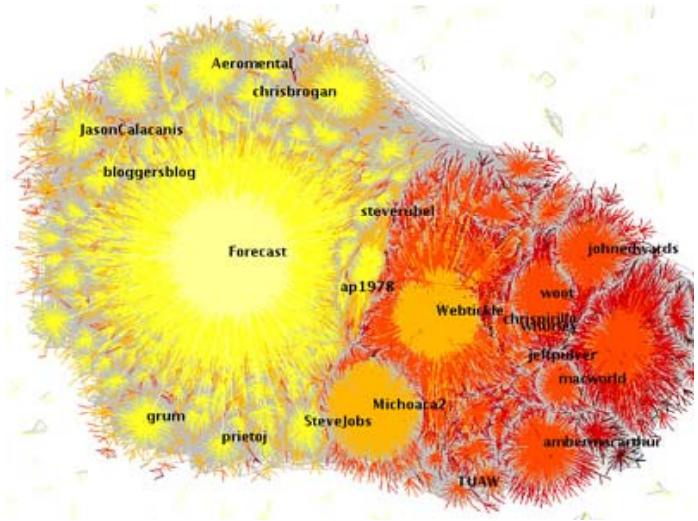
- **Rapidly growing use in biological, social, environmental, and other sciences**

The way it was …



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

The way it is now …



Twitter social network (|V|≈200M)

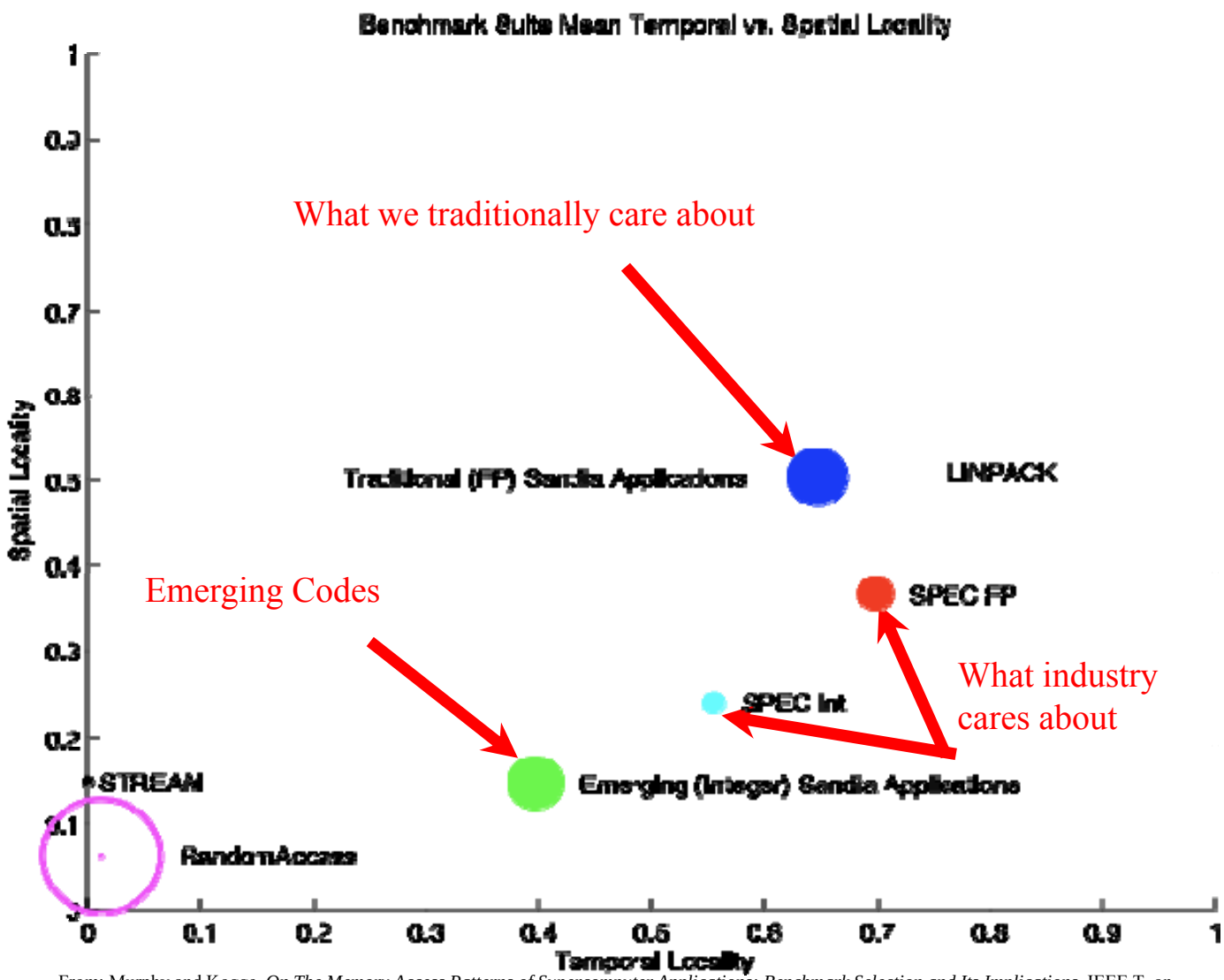
Sandia National Laboratories

Computational Challenges for Network Science

- **Unlike meshes, complex networks aren't partitionable**
- **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

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

Outline

- **Today's HPC landscape**

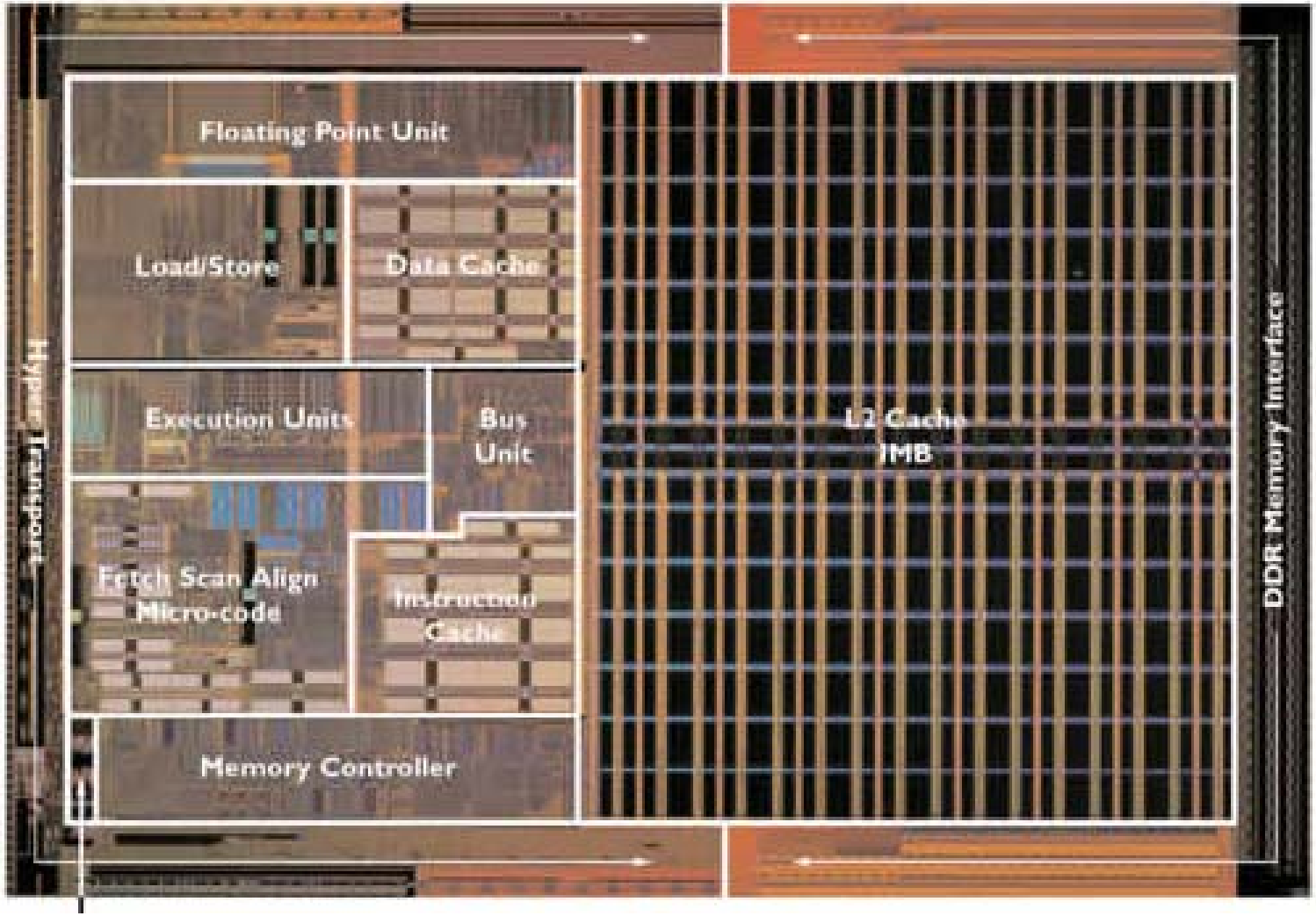
- **HPC Applications are changing**
  - Evolution
  - Revolution

- **Architectures are changing**
  - Evolution
  - Revolution

- **Conclusions:**
  - Organic forces will make HPC more data friendly
  - External forces will make HPC more data-centric

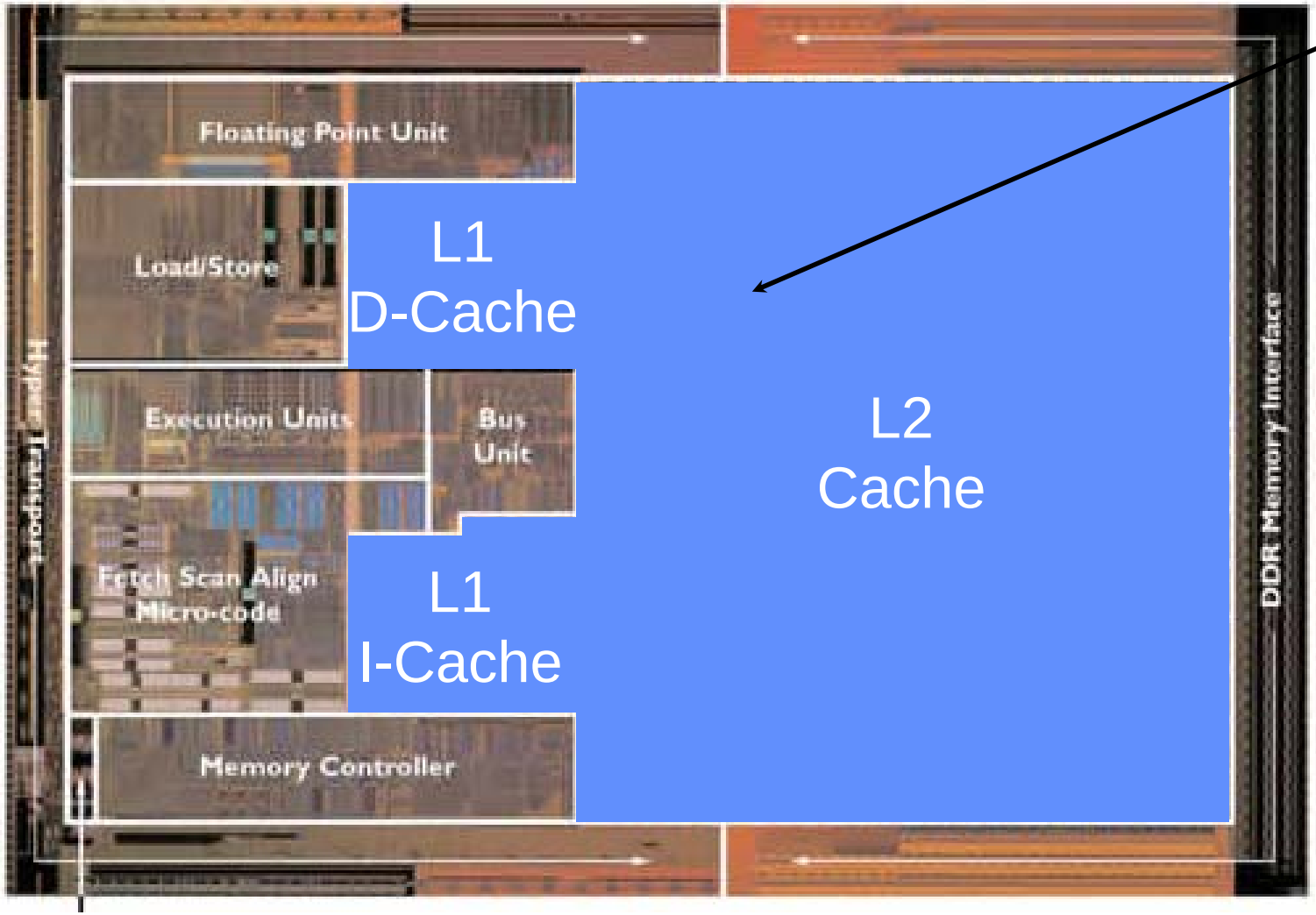
Sandia National Laboratories

Example: AMD Opteron



Sandia National Laboratories

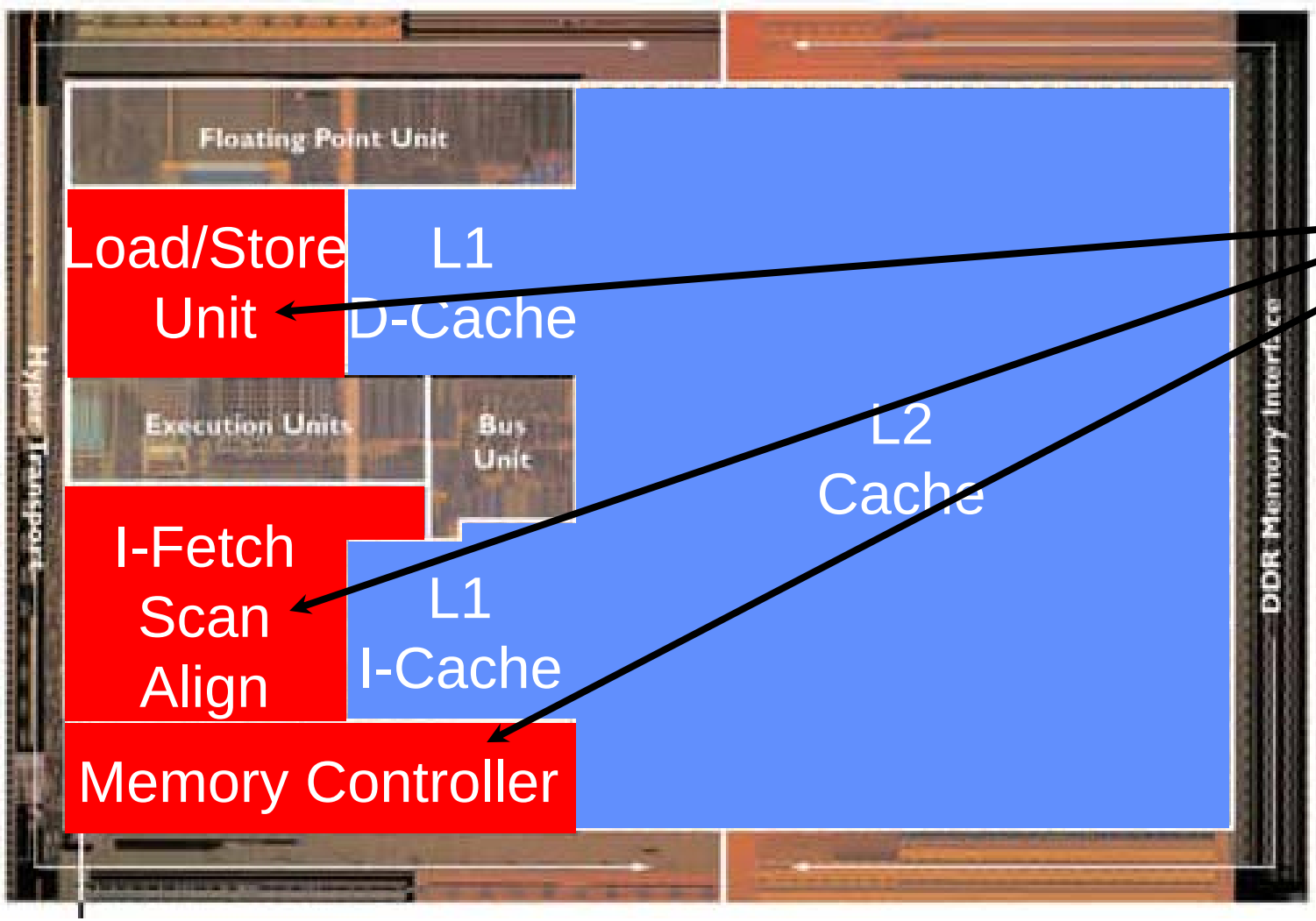
Example: AMD Opteron



Memory (Latency Avoidance)

Sandia National Laboratories

Example: AMD Opteron

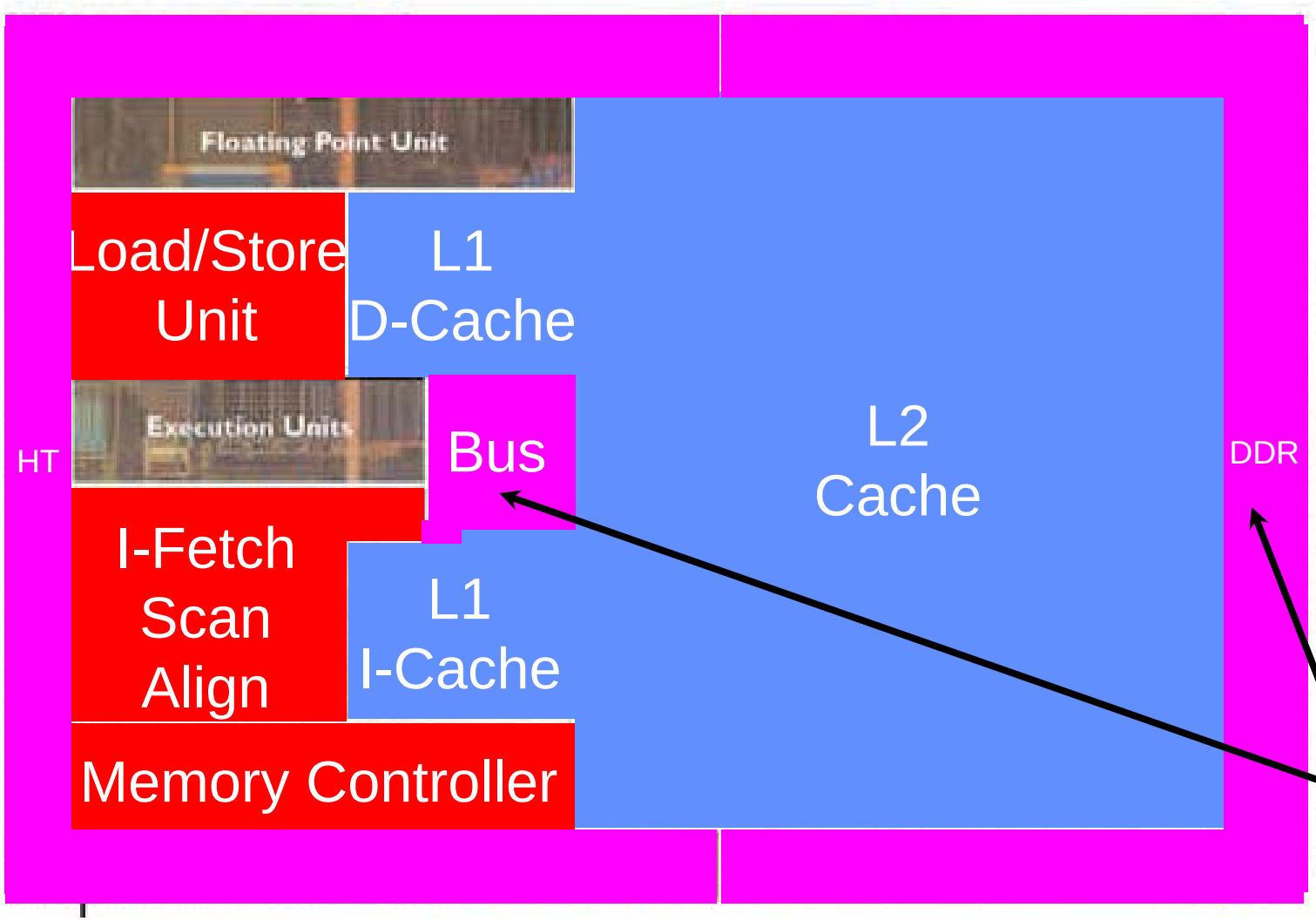


Memory
(Lat. Avoidance)

Out-of-Order
Exec
Load/Store
Mem/Coherency
(Latency
Tolerance)

Sandia National Laboratories

Example: AMD Opteron



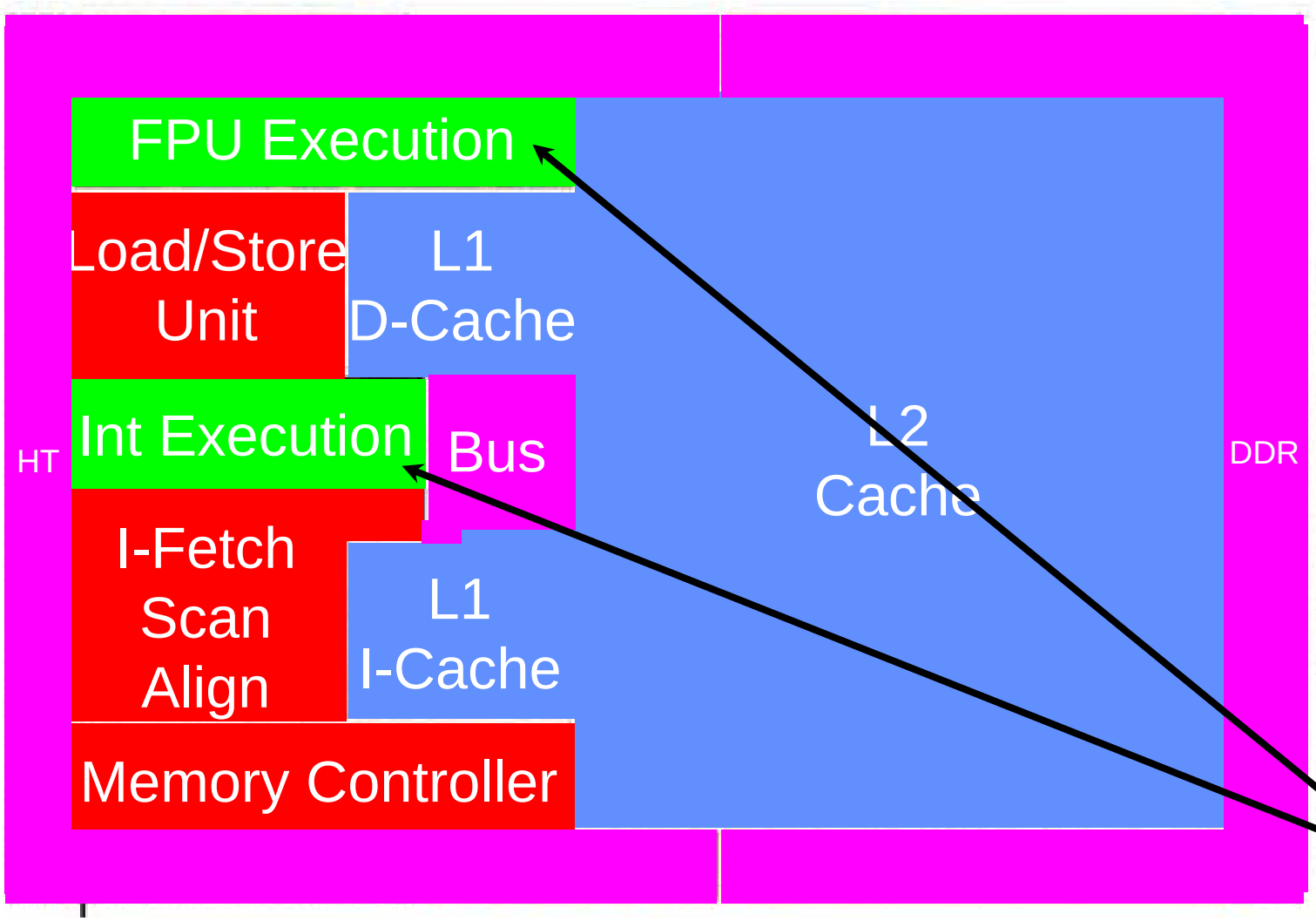
Memory (Latency Avoidance)

Out-of-Order Exec Load/Store Mem/Coherency (Lat. Toleration)

Memory and I/O Interfaces

Sandia National Laboratories

Example: AMD Opteron



Memory (Latency Avoidance)

Out-of-Order Exec
Load/Store
Mem/Coherency
(Lat. Tolerance)

Memory and I/O Interfaces

COMPUTER

Thanks to Thomas Sterling

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**

Architectural Wish List for Graphs

- **Low latency / high bandwidth**
  - For small messages!
- **Latency tolerant**
- **Light-weight synchronization mechanisms for fine-grained parallelism**
- **Global address space**
  - No graph partitioning required
  - Avoid memory-consuming profusion of ghost-nodes
  - No local/global numbering conversions

- **One machine with these properties is the Cray XMT**
  - Descendent of the Tera MTA

How Does the XMT Work?

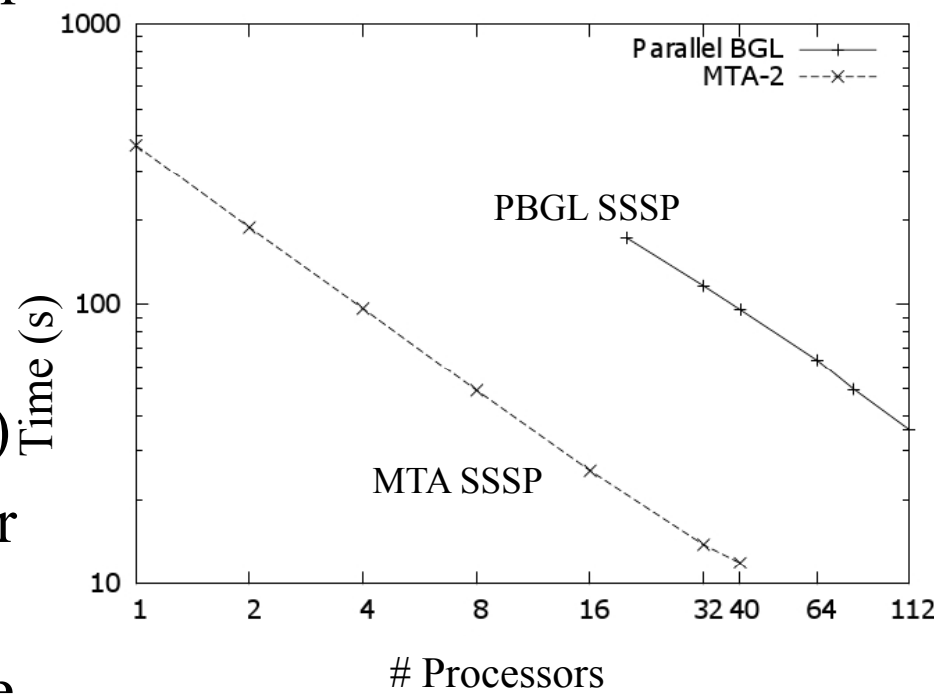
- **Latency tolerance via massive multi-threading**
  - Context switch in a single tick
  - Global address space, hashed to reduce hot-spots
  - No cache or local memory.
  - Multiple outstanding loads
- **Remote memory request doesn't stall processor**
  - Other streams work while your request gets fulfilled
- **Light-weight, word-level synchronization**
  - Minimizes conflicts, enables parallelism
- **Flexible dynamic load balancing**

- **Slow clock, 400 MHz**

Case Study: Single Source Shortest Path

- **Parallel Boost Graph Library (PBGL)**
  - Lumsdaine, et al., on Opteron cluster
  - Some graph algorithms can scale on some inputs

- **PBGL – MTA2 Comparison on SSSP**
  - Erdös-Renyi random graph ($|V|=2^{28}$)
  - PBGL SSSP can scale on non-power law graphs
  - Order of magnitude speed difference
  - 2 orders of magnitude efficiency difference
    - Big difference in power consumption
  - [Lumsdaine, Gregor, H., Berry, 2007]



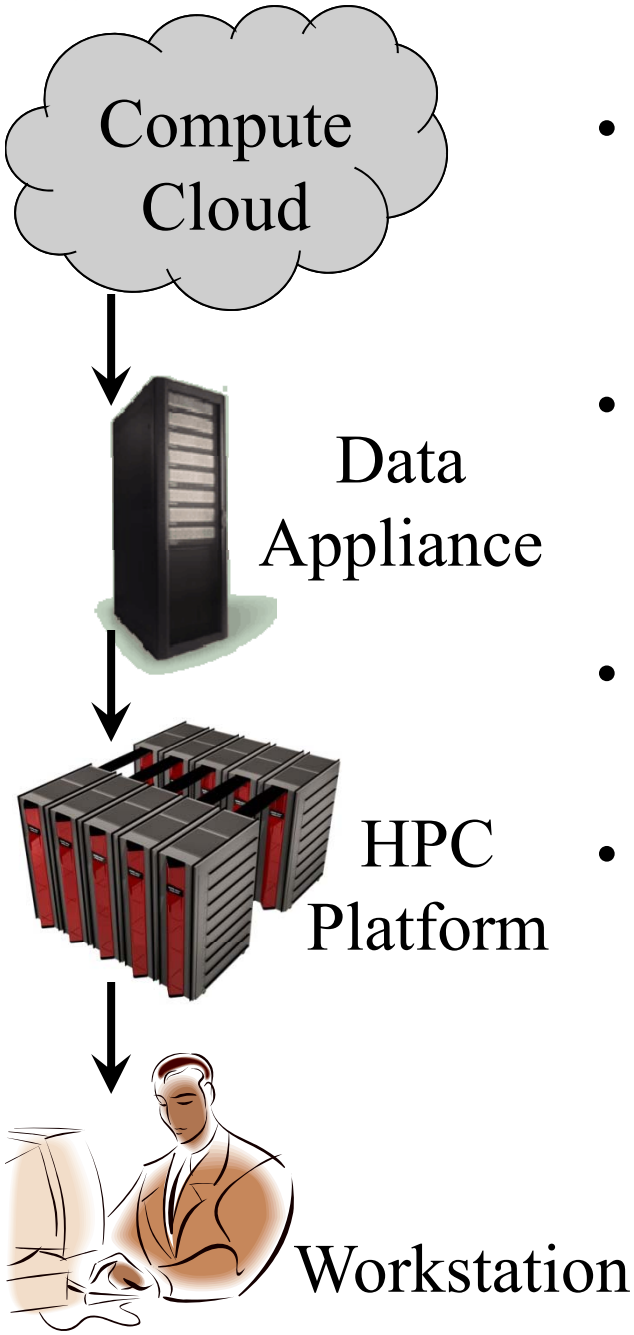
Sandia National Laboratories

Longer Term Architectural Opportunities

- **Near future trends**
  - Multithreading for latency tolerance on commodity processors
  - Growing heterogeneity in our compute nodes
  - Specialized machines targeting market segments
- **Further out**
  - Application-specific circuitry?
    - E.g. common scientific kernels
  - Reconfigurable hardware?
    - Adapt circuits to the application at run time
- **Disruptive changes to our virtuous circle!**

Sandia National Laboratories

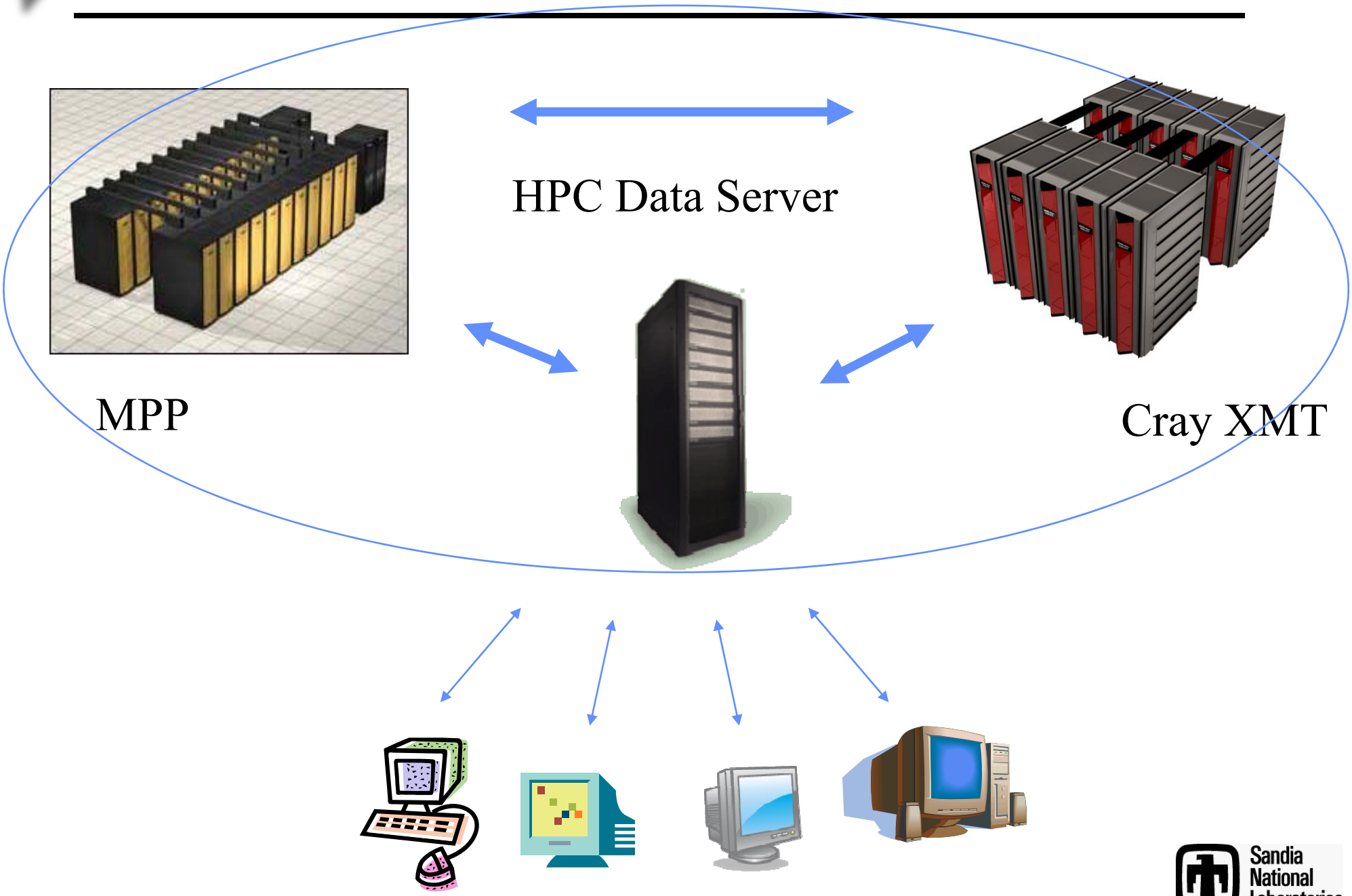
Role for HPC in Data Analysis Pipeline



- **Graph created from raw data, explored and studied under analyst direction.**
- **Analysis done at every stage in pipeline**
  - Data size decreasing downwards
  - Algorithmic complexity increasing downwards
- **Data flow not fully unidirectional**
  - Analysis results get updated in database
- **Human-centric HPC**
  - Interactive, not batch computing
  - Orchestration is immensely complicated
  - Usability and human factors are big challenges

Sandia National Laboratories

HPC Analysis System Architecture



Sandia National Laboratories

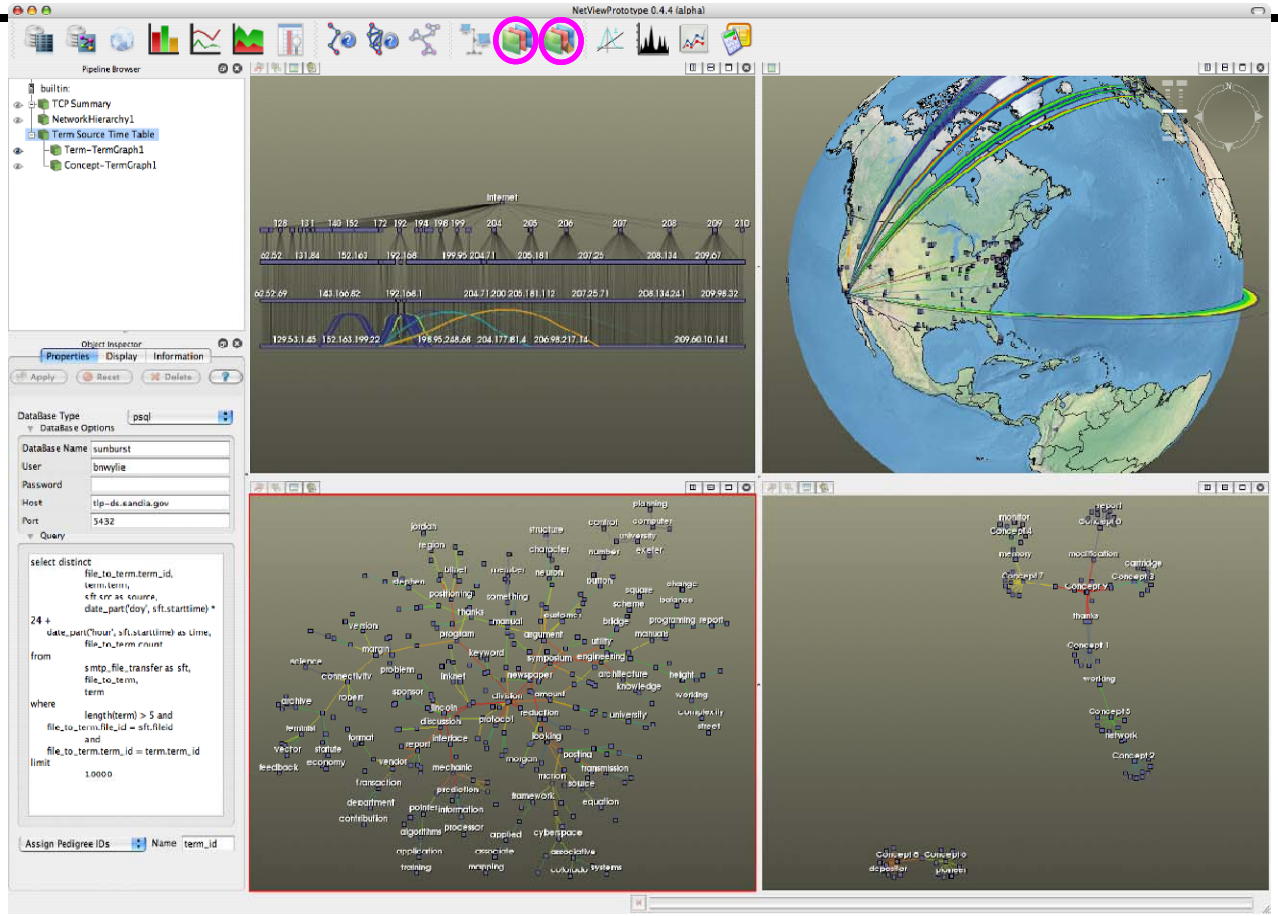
Exemplar: Sandia's Networks Grand Challenge

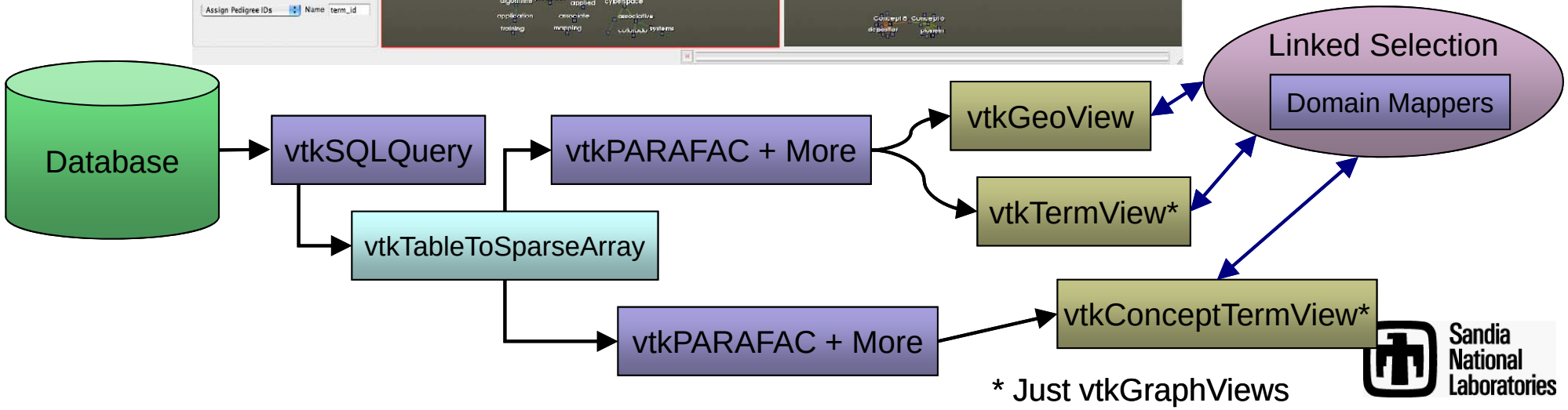
- **Large internally funded R&D project**

- **Goals:**
  - Enable exploratory analytics at scale
  - Support rich combinations of analytical methods
    - Graph analytics, algebraic methods, statistics, info-viz, …
  - Focus on usability and usefulness of capabilities
    - Significant human-factors investment
    - Close collaboration between researchers and target customers
  - Create open-source foundation for further R&D
    - Builds on ParaView

Use Case for First Prototype:

What 'payloads' were contained in the network transfers?

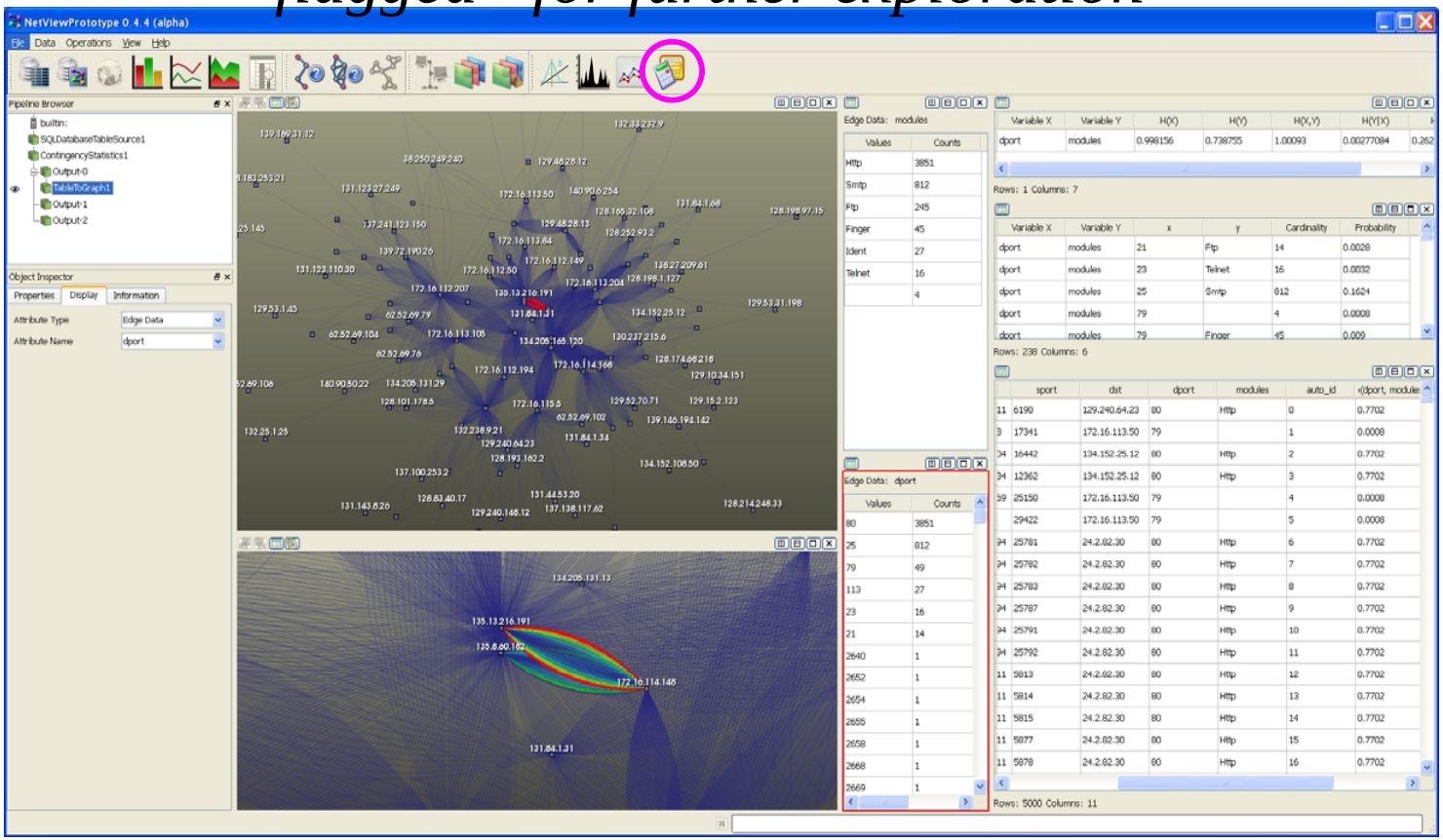


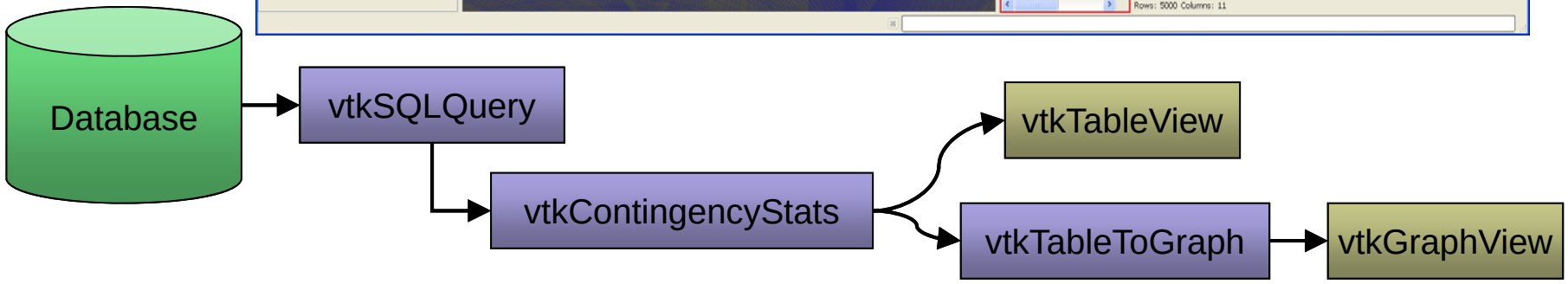


Sandia National Laboratories

Another Use Case:

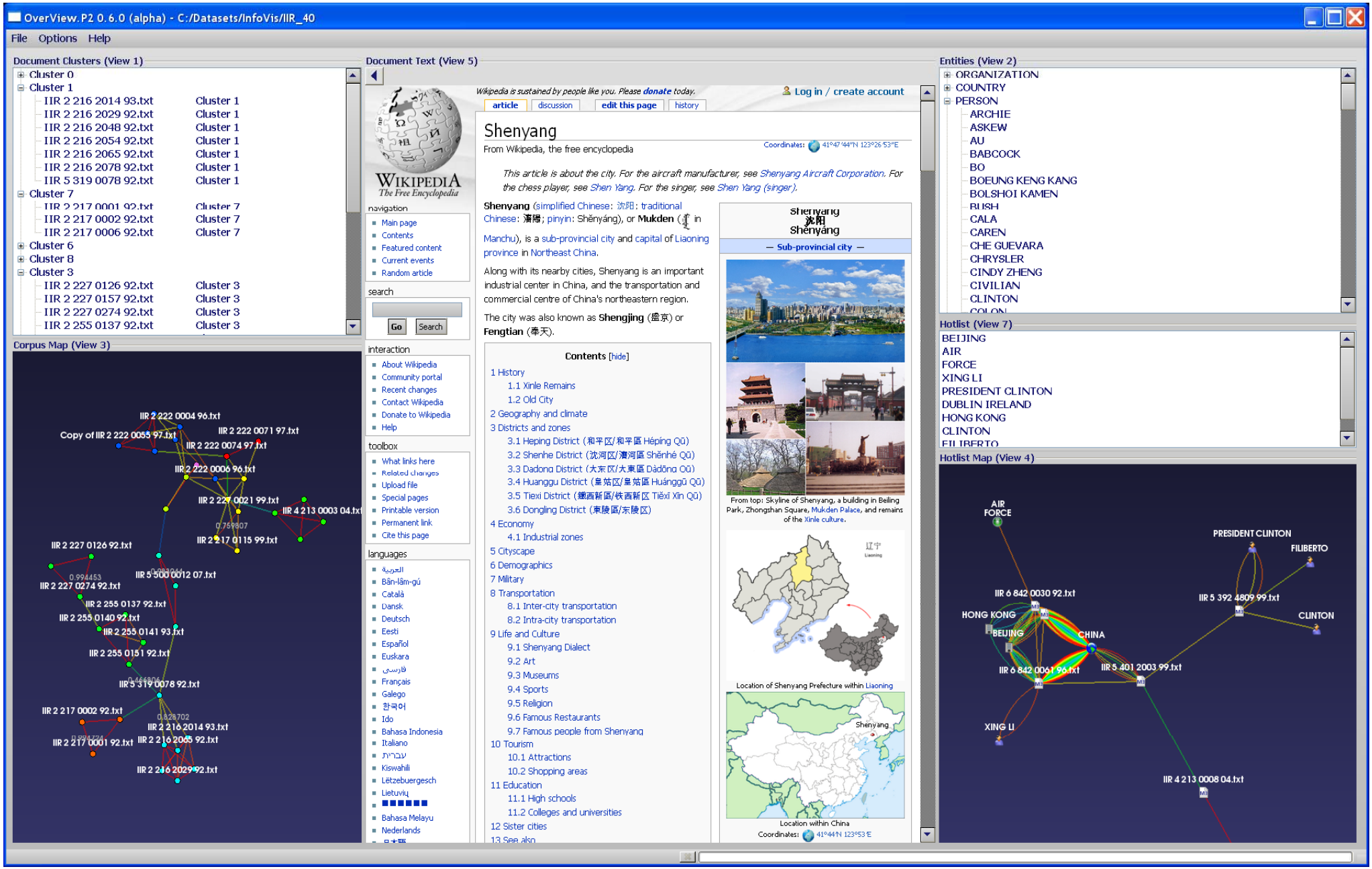
Analyst needs suspicious (out of the 'norm') behaviors "flagged" for further exploration





Sandia National Laboratories

Second Prototype



Conclusions

- **Organic forces will make HPC more data friendly**
  - BSP/MPI hegemony is breaking down
    - Multicore, more complex applications (traditional & emerging)
  - Memory performance is already the key to HPC
- **External forces will make HPC more data-centric**
  - Unprecedented opportunities for architectural innovation
  - Data rich applications of growing importance to science
  - Opportunities to impact applications currently not on the HPC radar
- **Enormous challenges ahead**
  - Some will need to be addressed by HPC community anyway
  - Cultural challenges may prove the most daunting

Sandia National Laboratories

Thanks

- **Cevdet Aykanat, Jon Berry, Rob Bisseling, Erik Boman, Bill Carlson, Ümit Çatalürek, Edmond Chow, Karen Devine, Iain Duff, Danny Dunlavy, Alan Edelman, Jean-Loup Faulon, John Feo, John Gilbert, Assefaw Gebremedhin, Mike Heath, Paul Hovland, Simon Kahan, Pat Knupp, Tammy Kolda, Gary Kumfert, Fredrik Manne, Michael Mahoney, Mike Merrill, Richard Murphy, Esmond Ng, Ali Pınar, Cindy Phillips, Steve Plimpton, Alex Pothen, Robert Preis, Padma Raghavan, Steve Reinhardt, Suzanne Rountree, Rob Schreiber, Viral Shah, Jonathan Shewchuk, Horst Simon, Dan Spielman, Shang-Hua Teng, Sivan Toledo, Keith Underwood, etc.**