# Query-Driven Visualization Accelerates Scientific Insight

**13 November 2006**
**SC06 – Tampa FL, USA**

**E. Wes Bethel**
***Lawrence Berkeley National Laboratory***




## The Take-Home Message

- Gaining insight from large and complex datasets is a major bottleneck.
- Thesis: effective software tools for gaining insight from such data must use a blend of technologies from:
  - Scientific Data Management
  - Visualization, Analytics
  - Computer Science
- Query-Driven Visualization is one such approach.



## Motivation

- Simulations and experiments are generating data faster than it can be analyzed and understood.
- Science bottleneck: information analysis and understanding.



## High Performance Visualization

- Observation: 80% of the code in any visualization application does "data management."
- Theses:
  - "Scalable visualization" solutions do not necessarily increase the likelihood of scientific insight.
  - More visualization output can cause more problems than it solves.
    - Increased depth complexity.
    - Increased cognitive workload.
  - Our approach to high performance visualization is to focus visualization processing on that subset of data deemed to be scientifically interesting.
  - Any tractable solution to "large data visualization" must address scientific data management issues.
  - Use SDM technology for data management.

# Query-Driven Visualization

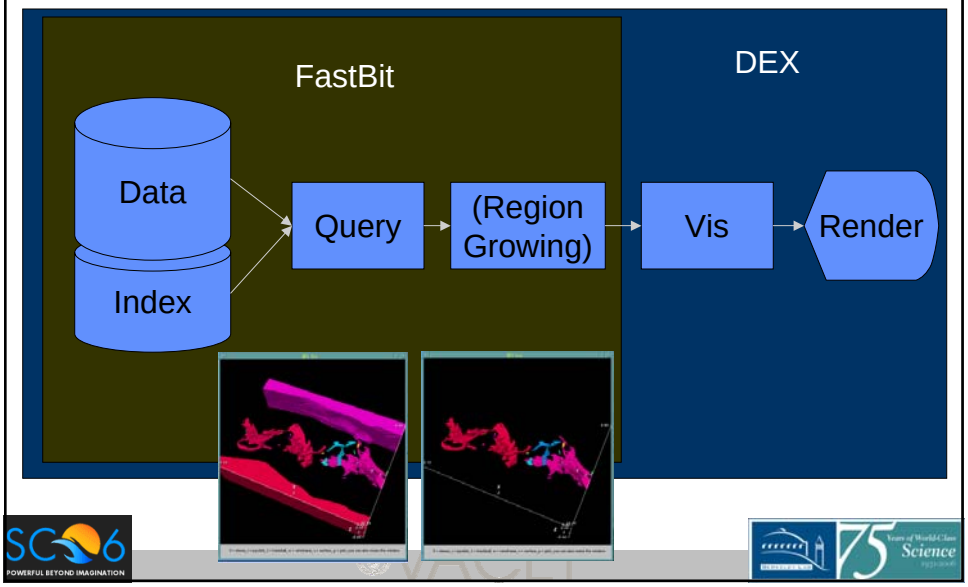
- What is Query-Driven Visualization?
  - Find “interesting data” and limit visualization, analysis, machine and cognitive processing to that subset.
- One way to define “interesting” is with compound boolean range queries.
  - E.g., ($CH_4 > 0.1$) AND ($T_1 < temp < T_2$)
- Quickly locate those data that are “interesting.”
- Pass results along to visualization and analysis pipeline.



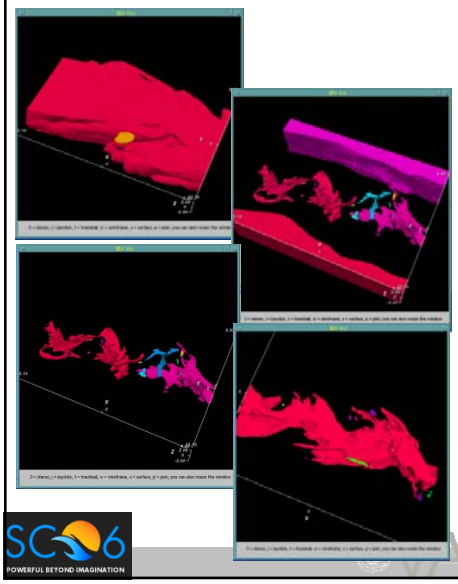
# Query-Driven Visualization


The Canonical Visualization Pipeline

# Query-Driven Visualization




# Query-Driven Visualization



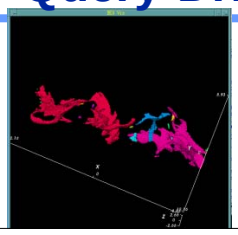
- $CH_4 > 0.3$
- $Temp < T_1$
- $CH_4 > 0.3$ AND $temp < T_1$
- $CH_4 > 0.3$ AND $temp < T_2$
  - $T_1 < T_2$

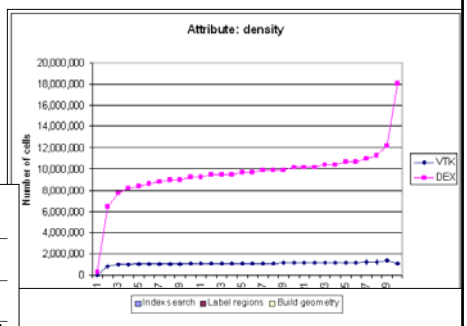
# Query-Driven Visualization

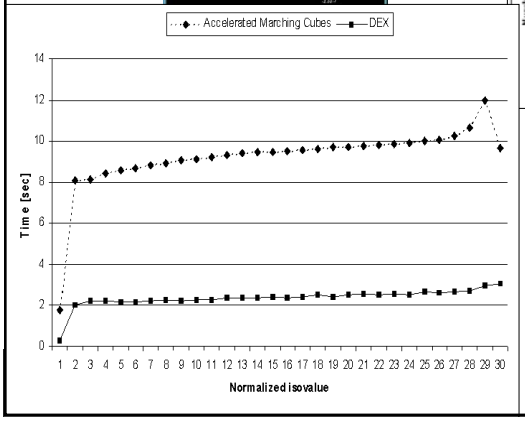
- How fast is it?
  - Comparison: Isosurface algorithms:
    - Nice summary in: Sutton et. al., "A Case Study of Isosurface Extraction Algorithm Performance," *2nd Joint Eurographics-IEEE TCCG Symposium on Visualization*, May 2000
  - For *n* data values and *k* cells intersecting the surface:
    - Marching Cubes: O(*n*)
    - Octtree methods: O(*k* + *k* log (*n*/*k*))
      - Acceleration: pruning; sensitive to noisy data
    - Span-space methods:
      - NOISE: O(sqrt*(n)* + *k)*
      - ISSUE: O(log *(n/L)* + sqrt*(n)/L* + *k)*
        - » $L$ is a tunable parameter
      - Interval Tree: O(log *n* + *k*)
- FastBit: O(k) – the theoretical optimum.
  - Profound performance gain for Petascale visualization!

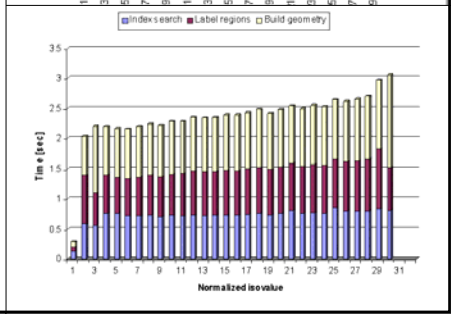


# Query-Driven Visualization





## Query-Driven Visualization

- What do these timing results mean?
  - In a one-sided matchup (DEX doing a lot more work), our performance results are markedly better for a given task than an industry-standard isocontouring implementation.
- These are single-valued queries.
  - DEX capable of *n-dimensional* queries.
  - Tree-based indexing methods not capable of *n-dimensional* queries.
- Why compare against isosurfacing?
  - Familiar to the visualization community.



## Query-Driven Visualization

- The previous work explored the feasibility of the approach by performing a speed comparison with the fastest industry-standard algorithm for finding “interesting” scalar data.
- The next sequence of slides discusses application of the work to a cybersecurity application – proof that the idea is generally applicable to large data visualization.

## QDV – Detecting Distributed Scans

- The problem:
  - One day's worth of traffic consists of tens of millions of individual connections.
  - Traffic increasing by an order of magnitude every 48 months.
    - ESnet monthly traffic levels now exceed 1 PB.
  - The Internet is a hostile environment, and it will get worse.
  - Objective: enable rapid forensic data analysis (network flow records).



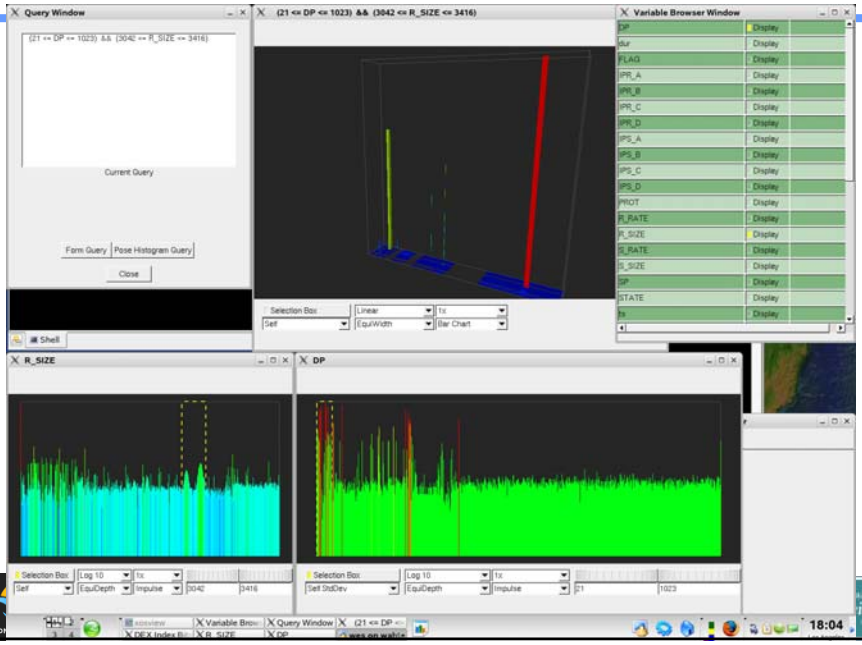
## QDV – Detecting Distributed Scans

- The data:
  - 42 weeks' of connection records from Bro (NERSC).
  - 281GB for raw data, 78GB for compressed bitmap indices.

- "Hero-sized problem"
  - No previous network analysis work has ever attempted to perform interactive visual analytics on data of this scale.
  - Result: what once took days (if at all possible) now takes seconds.

SC06 POWERFUL BEYOND IMAGINATION VACET 75 Years of World-Class Science

# QDV – Detecting Distributed Scans

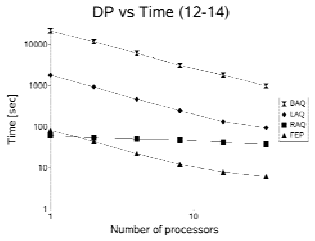



# QDV – Detecting Distributed Scans

- How fast is it?
  - 3 to 6 orders of magnitude faster than shell-based tools.
  - 2 to 3 orders of magnitude faster than ROOT, the “gold standard” in the HEP community.
  - Shows favorable scaling characteristics up to 32P.

1D Histogram, per-bin queries

| PEs | Shell-based | ROOT/Projection Index | ROOT/FastBit |
|---|---|---|---|
| 1 | 156381.14 | 1357.07 | 5.36 |
| 2 | 71835.32 | 600.05 | 3.72 |
| 6 | 21952.12 | 214.14 | 2.66 |
| 13 | 9389.96 | 113.88 | 2.58 |
| 21 | 2237.53 | 98.95 | 2.05 |

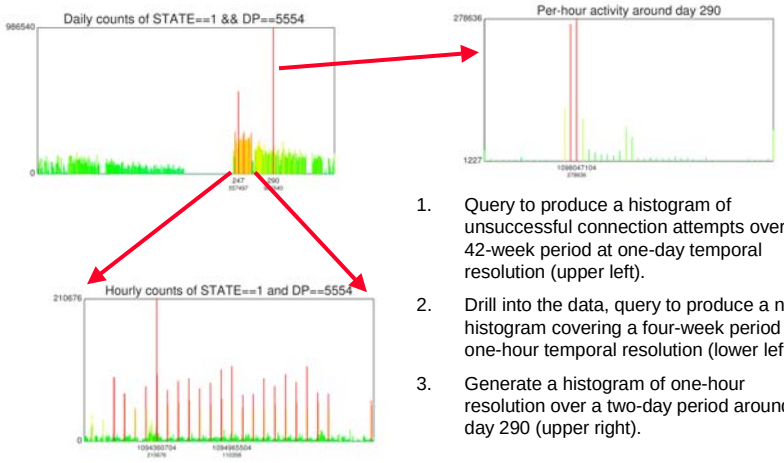



Conditional 2D histogram processing time

Query: (1000 < DP < 11000) AND (50 <= tsyday <= 350) AND (state==1) AND (12 <= tshour <=14), 10K total bins.

# QDV – Detecting Distributed Scans

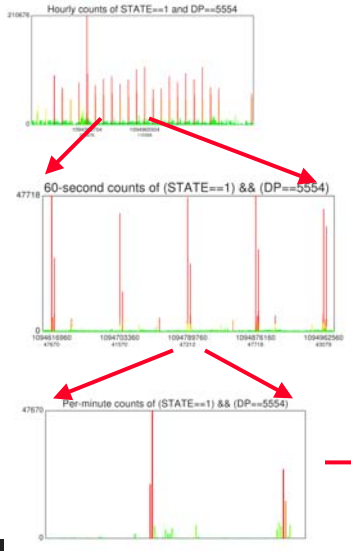
- The Case Study



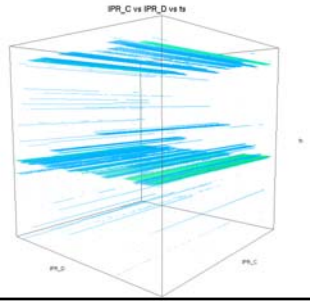

1. Query to produce a histogram of unsuccessful connection attempts over a 42-week period at one-day temporal resolution (upper left).
2. Drill into the data, query to produce a new histogram covering a four-week period at one-hour temporal resolution (lower left).
3. Generate a histogram of one-hour resolution over a two-day period around day 290 (upper right).



# QDV – Detecting Distributed Scans




5. Query to generate a histogram of unsuccessful connection attempts over a five-day period sampled at one-minute temporal resolution (middle, left). Regular attacks occur at 21:15L, followed by a second wave 50 minutes later.
6. Query to generate histogram over a two-hour period at one-minute temporal resolution (lower left).
7. Query to generate a 3D histogram showing the coverage of attacks in destination address space (lower right).

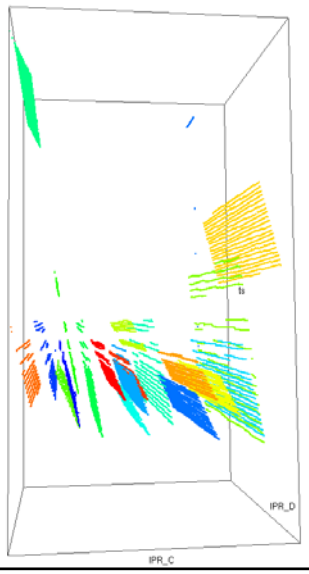


# QDV – Detecting Distributed Scans

After establishing that (1) a temporally regular activity is occurring, and (2) that it is in fact a systematic probe (scan) of entire blocks of network addresses, the next task is to determine the set of remote hosts participating in the attack.

Working backwards, we isolate the A, B, C and D address octets of the hosts participating in the attack.

This image shows a 3D histogram of the destination address space being attacked by each of 20 different hosts. The vertical axis is time – a seven-minute window at one-second temporal resolution.




# QDV – Detecting Distributed Scans

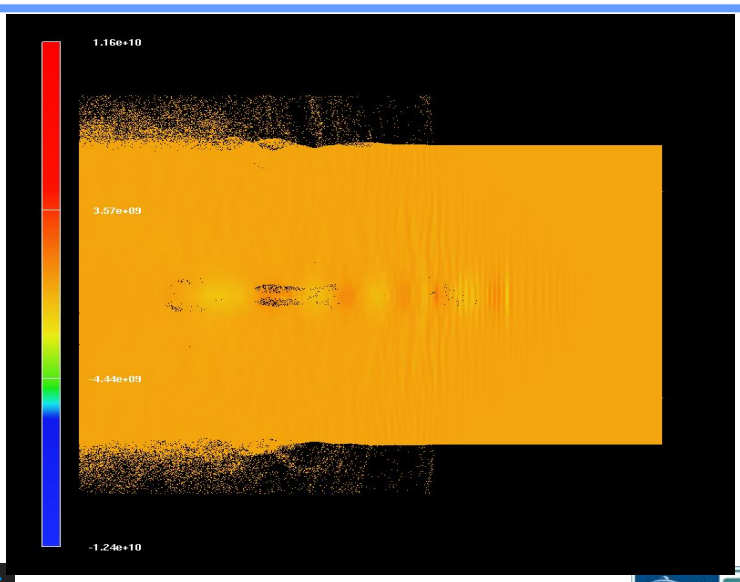
- Our analysis was performed in statistical space only.
  - We never accessed the raw data.
  - Our processing and visualization used only the index data.
- The same principles can be (and will be) applied to scientific data.
- Challenges:
  - How to define “interesting?”
  - Effective user interfaces for:
    - Support rapid interrogation, propagating query results from step to step in the analysis process.
    - Multivariate visualization
    - Drill-down (mining), linked/correlated views
  - Adapting, applying and deploying these principles to scientific data (e.g., AMR, distributed data/computing resources).

## Plasma-Wakefield Accelerators

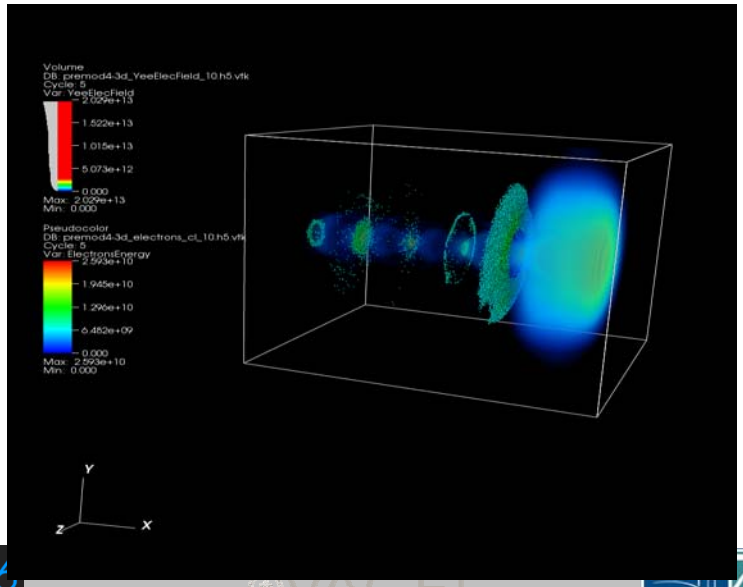
- 1000s of times more powerful than traditional RF-based accelerators.
- Short, intense laser pulses generate waves in plasma fields.
- Electrons are “bunched” into the wake of waves to a much higher degree than is possible with RF-based methods.
- This 2004 image shows 2D simulation results.
- Coincides with successful 2004 experiment at L’Oasis (LBNL)
- Present day – 3D simulations (DOE INCITE program).




## Plasma-Wakefield Accelerators

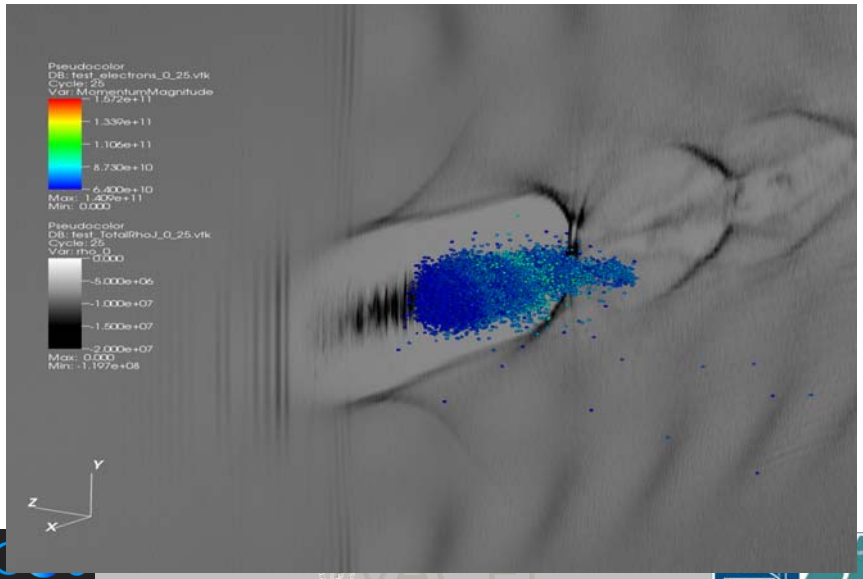


# Plasma-Wakefield Accelerators

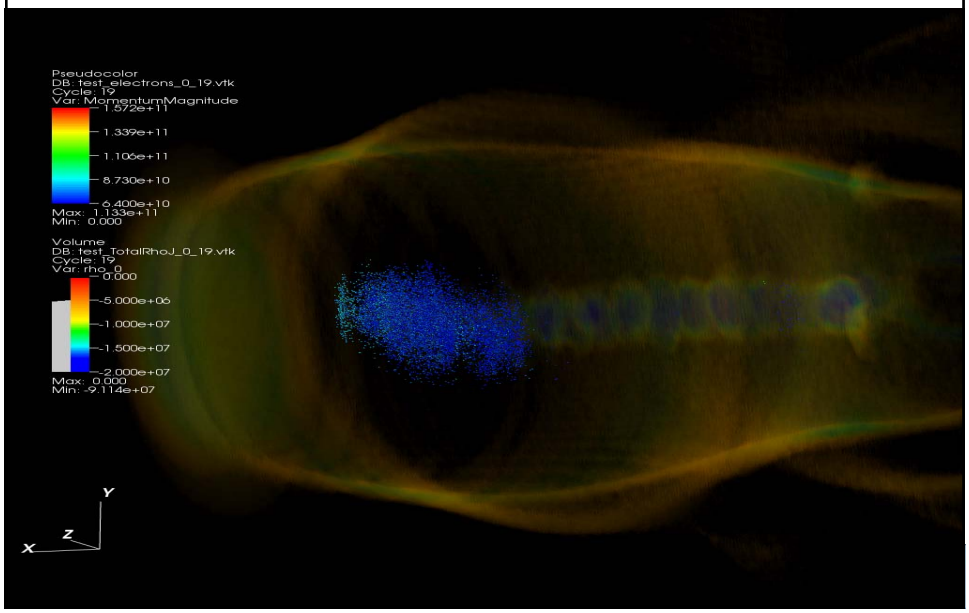





# Plasma-Wakefield Accelerators



# Plasma-Wakefield Accelerators




# Plasma-Wakefield Accelerator Lessons

- This work has lead to new scientific discoveries that are currently in press (at the request of the researchers, none of the preceding images show the new science).
- A visualization of “all the particles” is not scientifically helpful – only a small subset of particles “is interesting.”
  - Limiting analysis and vis to “interesting particles” is very interesting.
- An image of “all the data” isn’t typically scientifically useful.
  - However, those images show up on the cover of Nature.
  - Those images are very helpful in communicating with the public, funding agencies, etc.
  - Those images are very useful in creating excitement about science.

## QDV – Conclusion

- New capability: ability to focus visualization and analysis processing on interesting data.
  - Orders of magnitude faster than previous approaches.
  - Directly responsive to needs of scientific researchers.
  - Quantifiable reduction in data understanding duty cycle.
- Leverages state-of-the-art Scientific Data Management technology to accelerate searches.
- QDV concepts are general-purpose and scalable to 1000s of PEs.



## The End