
BERKELEY LAB
BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY
U.S. DEPARTMENT OF
ENERGY
Modern Financial Markets and
Data Intensive Science:
Leveraging 35 Years of Federal Research
Wes Bethel and David Leinweber
Presentation to CFTC Technical Advisory Committee
Washington, D.C.
December 13, 2011

Lawrence Berkeley National Laboratory

One of world's premier research institutions

Mission: Solve the most pressing and profound scientific problems facing humankind

- *Basic science for a secure energy future*
- *Understand living systems to improve the environment and energy supply*
- *Understand matter and energy in the universe*

16 Nobel Prizes,
2 Elements (Lawrencium & Berkelium)

Pioneer and Center of Excellence in **Data Intensive Science**

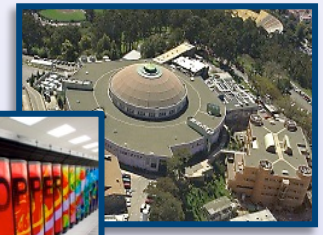
People

- *3,863 FTE*
- *3,040 Employees*
- *267 Joint faculty*
- *491 Postdoctoral researchers*
- *328 Graduate students*
- *194 Undergraduates*
- *8,025 Facility users*
- *1,612 Visiting scientists and engineers*

FY10 Total Operating Costs: $680.6M

LBNL at-a-glance

Advanced Light Source

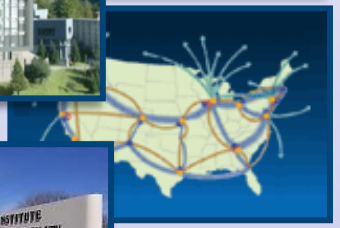


National Energy Research Scientific Computing Center (NERSC)

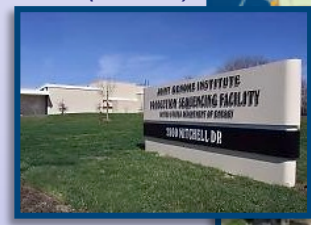
88-Inch Cyclotron

Molecular Foundry

Energy Sciences Network (ESnet)



Joint Genome Institute



National Center for Electron Microscopy

Berkeley Lab's largest **research facilities** see more than 25,000 users per year˝

High Frequency Trading:
Markets Become Data Intensive

Pervasive in stocks

- Growing in scope & across markets

Race to the bottom in latency

- Race to the top in message traffic
- Market "traffic jams" can happen

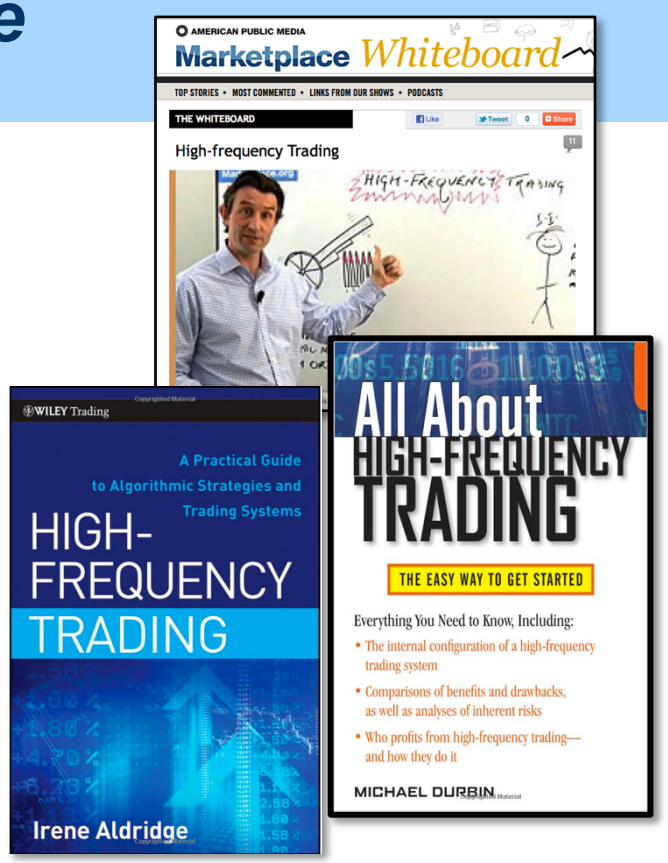
Major flavors

- Market making – liquidity provision
- Arbitrage & Short-term prediction

Contrast to slower buy-side algos

- Source of increasingly nervous buy-side & individual investors

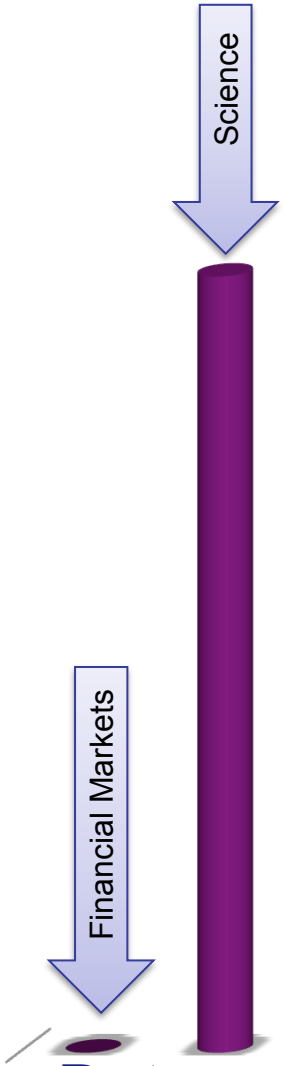
Creates a "Big Data" Problem



Marketplace: http://bit.ly/spxLPM
Aldridge: http://amzn.to/rRCz1P
Durbin: http://amzn.to/rBHWlC

BERKELEY LAB

“Big Data” and Federal Regulation of Financial Markets



IT woes are identified as a critical problem by CFTC, SEC

- Overwhelmed by “100 Gigabytes a day”
- Growing fast

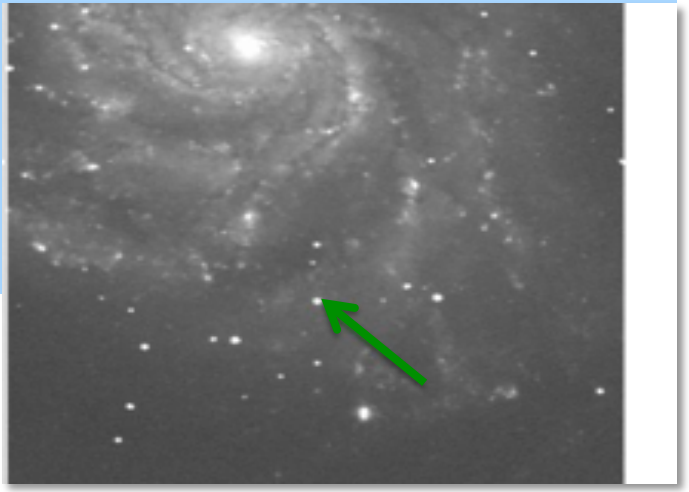
Supercomputing & Data Intensive Science routinely deal with flows and volumes ***hundreds of thousands times larger***

Leverage for successful modern market surveillance at reduced cost and risk

Early HFT Research

BERKELEY LAB

Flash in the market and a Flash in the sky



Imagine, for the sake of comparison:

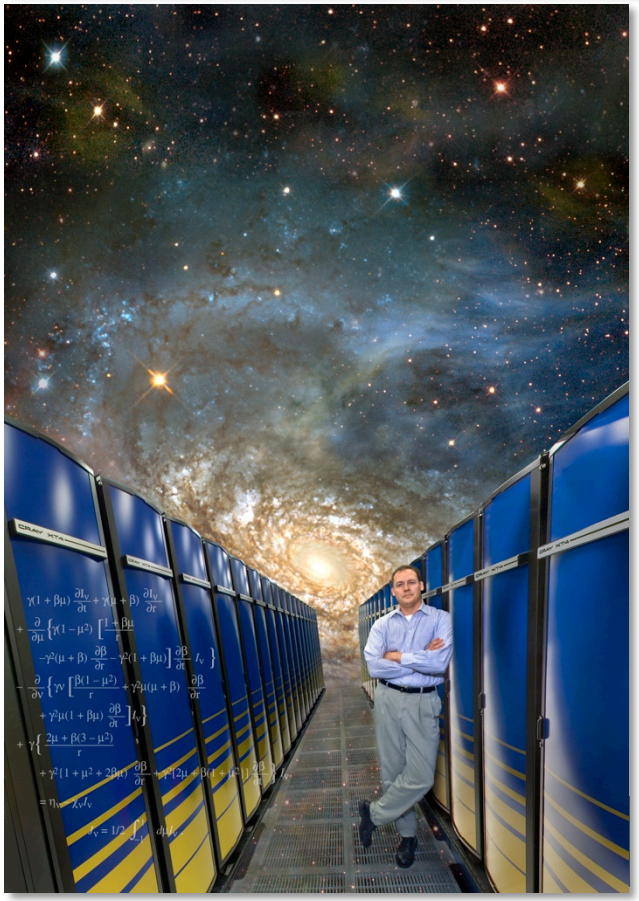
- Flash in the sky, (not the stock market), that you wanted to understand, and look at in other ways using other telescopes, satellites, and histories (not futures or options or multi-venues).

Once, doing this was the same messy tangle of incompatible scattered data we see in markets.

Now, a giant unified “Digital Eye in the Sky”

Nice coincidence:

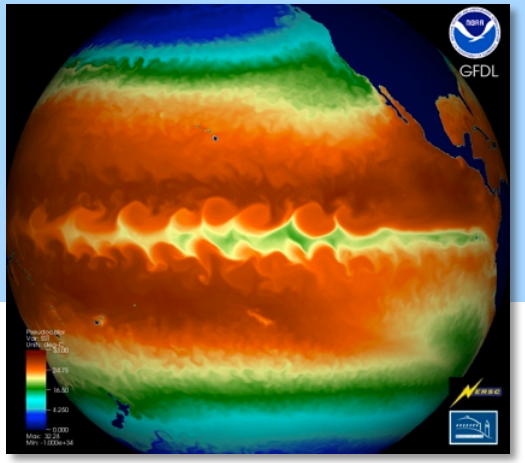
- Used as Flash Crash Metaphor: April 2011
- Supercomputer discovers actual supernova in “real-time” : August 2011

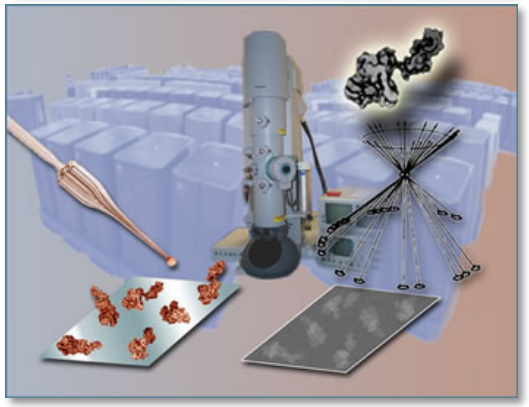


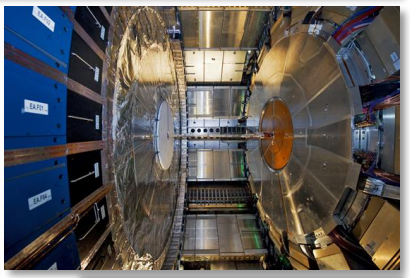
http://newscenter.lbl.gov/feature-stories/2011/08/25/supernova/

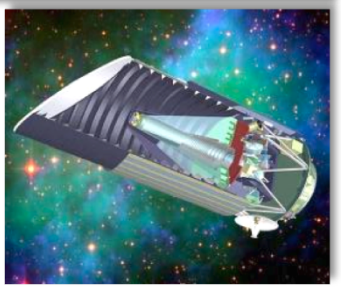
Data Intensive Science Success Stories

- Supernovas in real time
- Revolution in Earth Observation
- Two Nobel Prizes in Five Years based on data intensive research
- *Hardware, Software & Expertise*









http://www.nersc.gov/news-publications/publications-reports/science-highlights-presentations/

Major US DOE Lab Resources For Data Intensive Methods in Markets

Investment of >$30B since the 70s

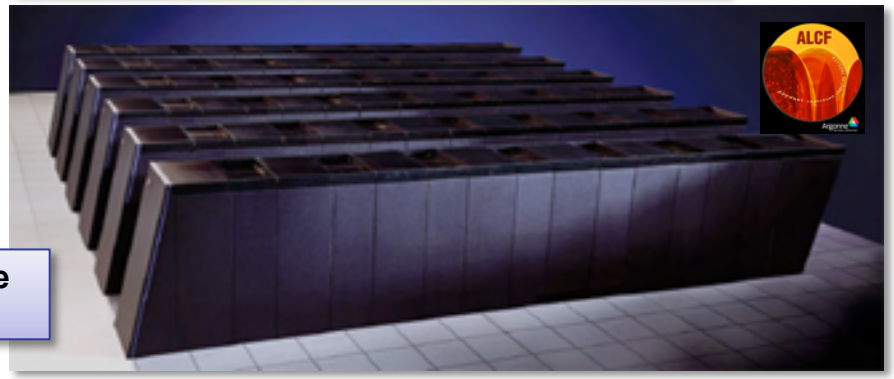
Components:

- Supercomputers
  - Now in all sizes
- Large scale data tools:
  - Analysis & Visualization
  - Simulation & Experiment
- ***Expertise and Experience***

Hopper (NERSC-6): Cray XE6 at Berkeley Lab

- 153,216 cores
- 212 Terabytes main memory
- 2000 Terabytes disk
- 2nd Fastest computer in US
- ~$5 Million/year electric bill!

ORNL Jaguar



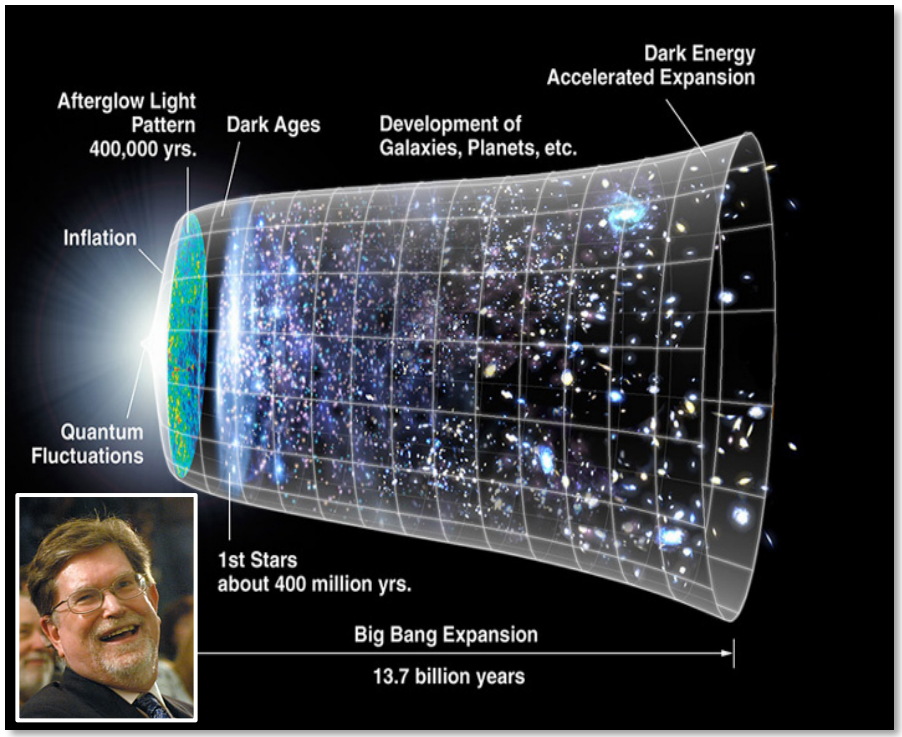
Argonne ALCF

U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

Berkeley Lab NERSC Data Intensive Science: Two Physics Nobel Prizes in 5 Years

George Smoot, 2006



Cosmic Background Radiation

Saul Perlmutter, 2011

Accelerating Expansion of the Universe

BERKELEY LAB

20 Key Questions for Markets
A "Jim Gray" starter list

Systemic Structural Risk

- Are **complex interactions between market centers** a source of risk due to unanticipated interactions when they are **operating as designed**?

Systemic Implementation Risk

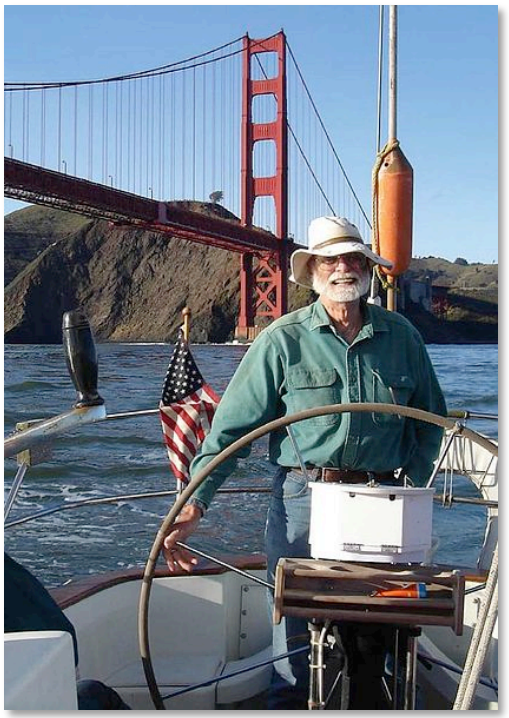
- **Same question as above, but** recognizing that markets are built on real computers, with delays, crashes, races, queues, slow-downs…

Enforcement

- Can you spot a **market manipulator who works in microseconds**?

Financial Cyber-security

- **The worst calls the heads of the** SEC/OFR/CFTC could get is **"Are our markets under attack?"**
- If that happened, test probes would certainly precede it. Would we know?

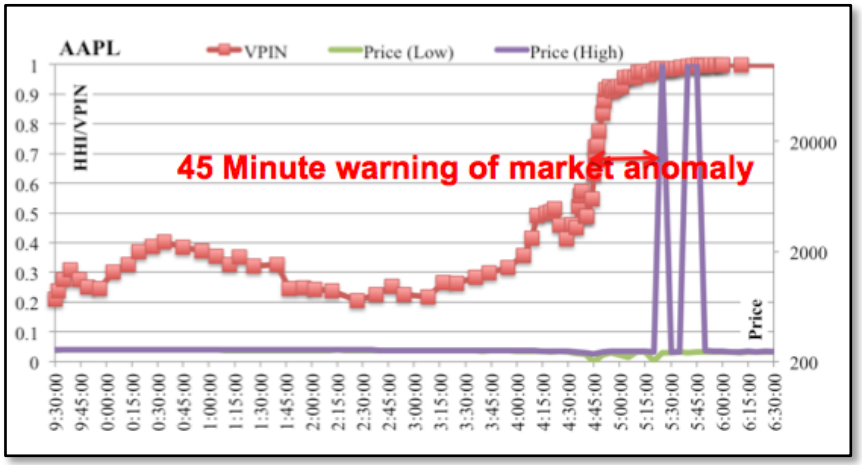


Policy Analysis

- Can we **simulate, analyze, model and visualize** what would happen if we make changes in the rules? Avoid unintended consequences..

Data Intensive Science Financial Prototypes at Berkeley Lab

- Replicate portions of Flash Crash Analysis
- Extend to test improved early warning "Soft Circuit Breaker Methods"
- Cooperation with agencies and market participants



REUTERS

Post 'flash crash' monitoring emerges at Berkeley

http://reut.rs/tsIRwu

"Top Ten" SSRN Paper: http://bit.ly/w4eHxY

BERKELEY LAB

Why Real-time Makes Sense:
NTSB Approach: Ex Post Analysis

BERKELEY LAB

Why Real-time Makes Sense:
FAA - Real-time safety and stability

Why Real-time Makes Sense

NTSB Data:
Ex Post Analysis

FAA Monitoring:
Real-time safety and stability

Two challenging yet soluble problems in HPC and Data-Intensive Science

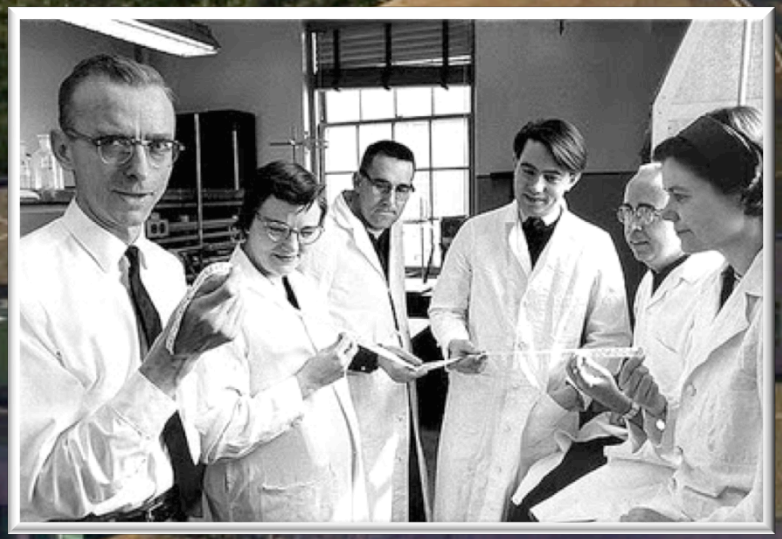
BERKELEY LAB

BERKELEY LAB

LAWRENCE BERKELEY NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Specific Technologies from Data Intensive Science and their Application to Financial Market Analysis/Understanding

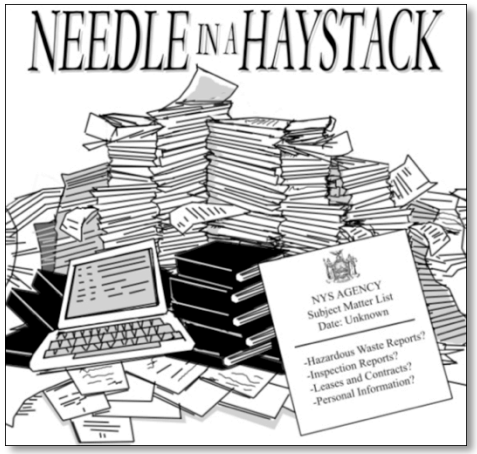


Studying HFT: A Big Data Problem

Regardless of "final definition" of HFT, studying it and its impact on markets is a "big data problem"

Two potential study scenarios:

- Historical: look at past data to find, study patterns, pose and test hypotheses
- Real-time: look at market data as it happens – want to find events, then make/execute policy decisions based upon events



BERKELEY LAB

Science Familiar with Big Data Problems: *Data Sources and Data Movement*

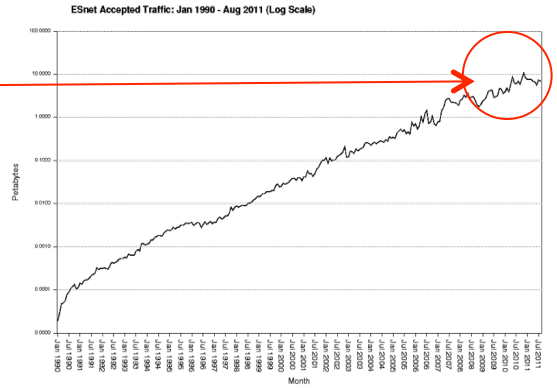
- **Supercomputer centers:** for computational modeling, "virtual experiments."

  Multi-petabyte data challenges

- **Experimental facilities:** particle accelerators, particle detectors, x-ray/laser light sources, etc.

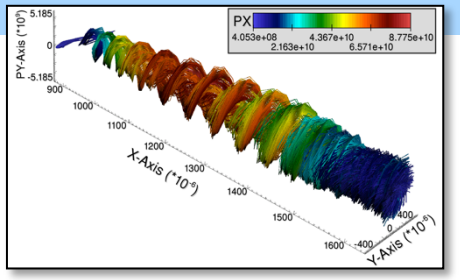
  Multi-petabyte data challenges

- **Networks:** high-speed, low-latency production networks that connect scientists worldwide to these facilities
  - DOE's Energy Sciences Network (ESnet):
    - Petabyte per month load now
    - 100 petabytes/month by 2015



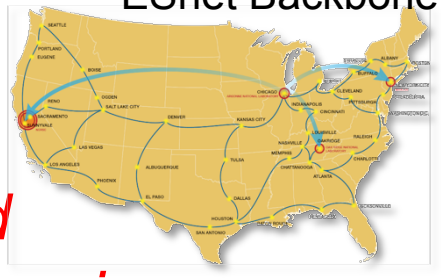
Science Familiar with Big Data Problems: *Analysis, Knowledge Discovery*

High Energy Physics



- Significant R&D investment on **tools, techniques for analyzing and understanding** ever-growing scientific data
- R&D on **monitoring, analysis** of world's largest production scientific network

ESnet Backbone



- *Many of these same tools, technologies, and expertise can be brought to bear on data-intensive challenges of financial markets*

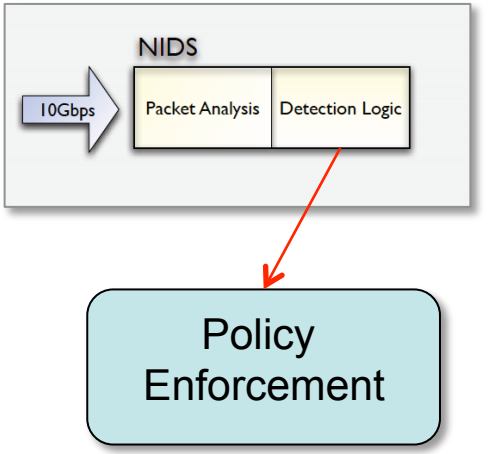
Two forensic cybersecurity examples follow

BERKELEY LAB

Berkeley Lab Competency: High Performance, Real-time Network Traffic Analysis and Policy Enforcement

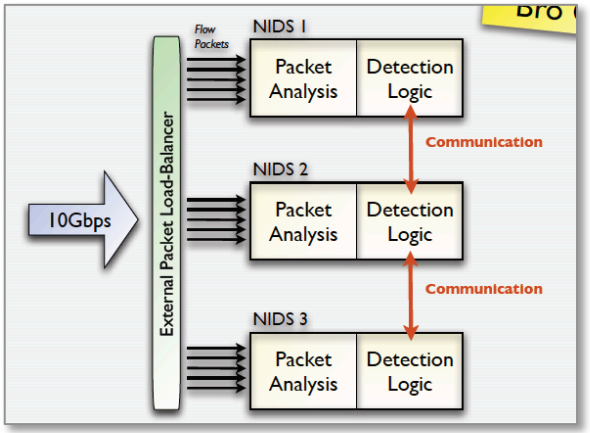
The **Bro Intrusion Detection System:**

- Detect events: packet-level analysis engine uses signature matching to detect events, generates a trigger
- Policy enforcement: given event triggers, make decisions about what to do



Faster networks require faster processing

- Scaling Bro-IDS: the "Bro Cluster" – required for processing traffic loads on 10Gbps links, in use now

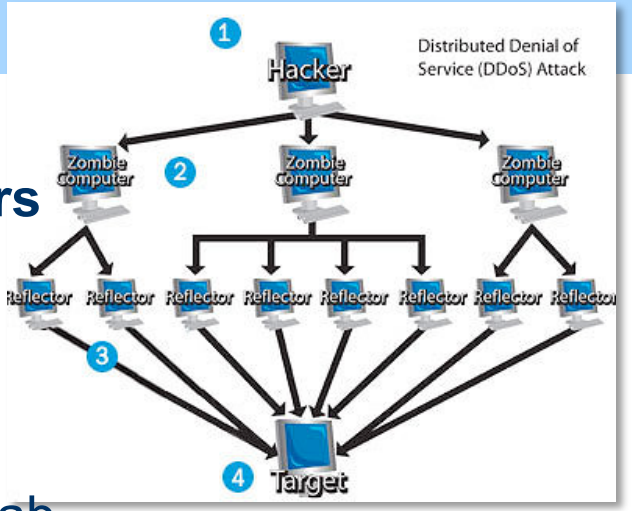


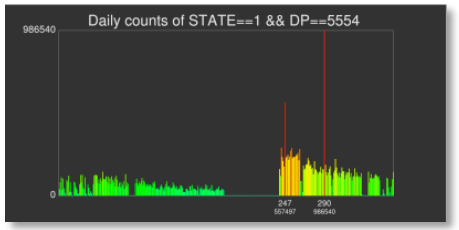
Many conceptual similarities to real-time financial market analysis

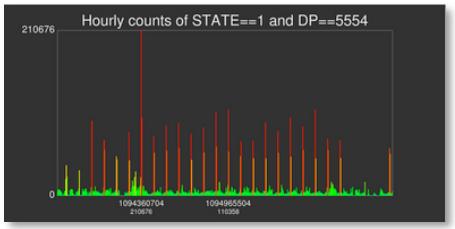
BERKELEY LAB

Berkeley Lab Competency: Accelerating Inspection, Analysis of Historical Network Data

- Use supercomputers to discover **distributed "bot-net" attack on US government centers**
- Source data: 1 years' worth of 'connection records,' ¼ terabyte in size
- Processing time:
  - Minutes: with supercomputers and related lab software technologies
  - Months: using traditional, "widely accepted" practices
- How? Parallel computing, data models/formats, index/query for fast subset selection and analysis

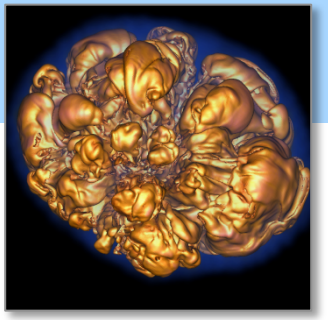


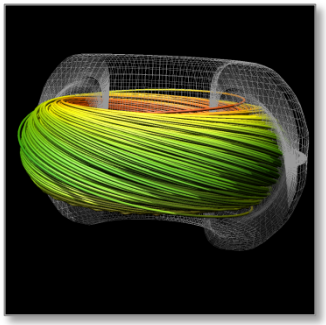


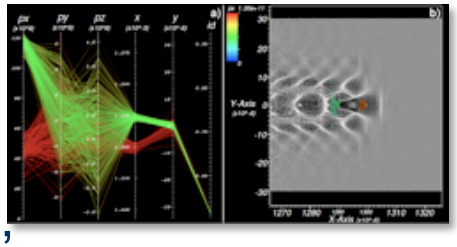


BERKELEY LAB

DOE Lab Computing Technologies Can Aid Market Analysis Efforts







Real-time analysis: scalable infrastructure for processing vast amounts of streaming data

- Look for "events" to enable execution/enforcement of policy
- Testbed for trying out new ideas, hypotheses

Historical: Scalable technologies for data indexing, searching

Related HPC technologies: I/O formats, data models, large scale computational modeling, advanced visual data analysis and exploration

Plenty of experience applying these technologies to diverse science problems: astronomy, astrophysics, climate modeling, combustion, fusion, high energy physics, …..

BERKELEY LAB

Federal Market Monitoring Lessons

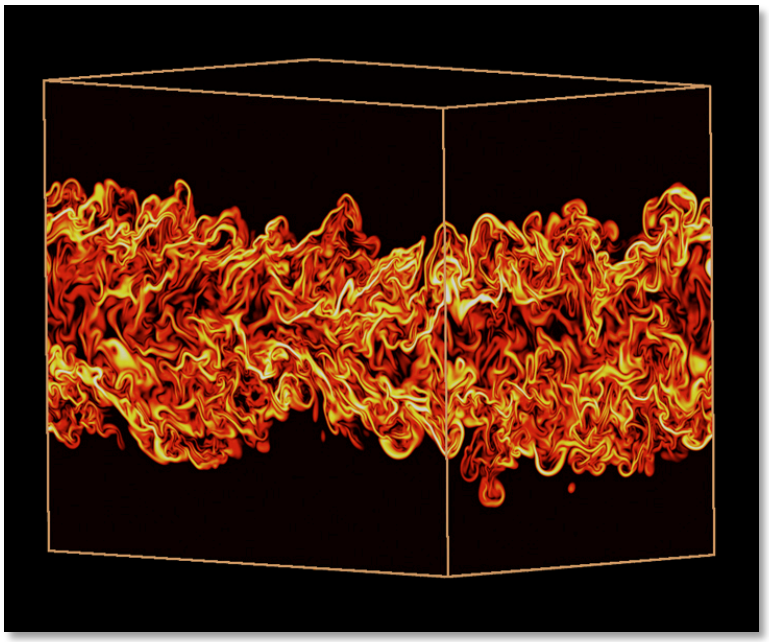
Financial markets are big data,
big *fast* data

- Bring the best technology to bear on the problem
- Nice that it already exists in Fed World

Real-time makes sense

- Don't apply rules of the road to aircraft

Stability, Safety, Security



For more information and slides with links

David Leinweber
LBL CIFT Cofounder
dleinweber@lbl.gov

E. Wes Bethel
Visualization Group Leader
ewbethel@lbl.Gov

David L. Brown
Computational Research
Division Director
dlb@lbl.gov

U.S. DEPARTMENT OF
ENERGY

UNIVERSITY OF
CALIFORNIA