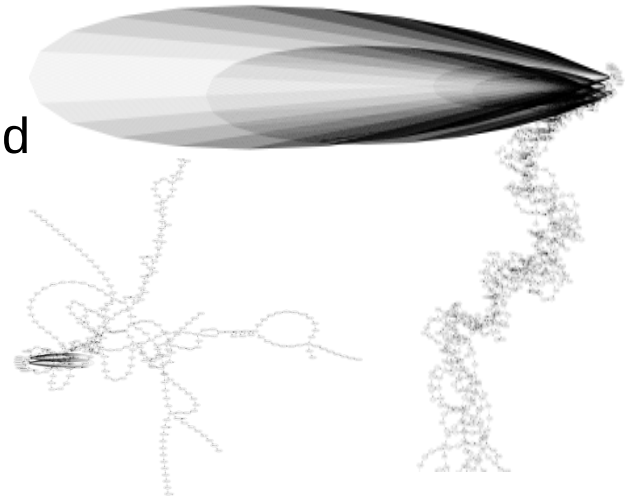
Data Intensive Supercomputing and Extreme Scale Information Analytics – Applications in Cyber Security

John R. Johnson

John.Johnson@pnl.gov

Data Intensive Supercomputing Focus Areas

- High performance computing for data-intensive problems
- Irregular applications that require novel or hybrid computing architectures
- Example application areas
  - Streaming data analysis
  - Complex network modeling and analysis
  - Large scale graph analytics
  - Multi-media analysis
  - Large scale metadata management and storage

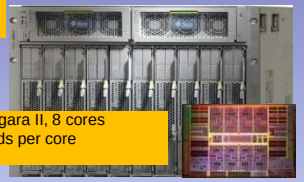


Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

PNNL Data Intensive Computing Systems

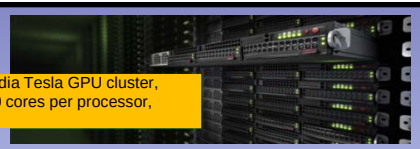
Excellence in Compute Intensive Applications

Cray XMT ,128 Nodes
128 Threads per node,
1 TB memory



Sun Niagara II, 8 cores
8 Threads per core

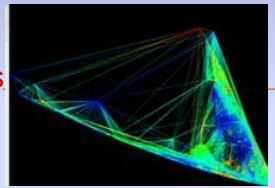
IBM Cell , 9 cores
Over 200 Gflops/processor



nVidia Tesla GPU cluster,
240 cores per processor,

Multi-threading

Irregular Applications



Complex Networks

Multi-core

NetezzaTwinFin

Storage

FusionIO NAND SSD

LSI Engenio 7900

Excellence in Storage Intensive Applications

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

PNNL HPC Technology Transition

Open Research Testbed | **Data Demonstration Testbed** | **Mission Support**

Number of systems decreases →

Size of systems increases →

Basic Research / University Collaboration

POC / Fundamental Capability Demonstration → **Community Testbed Real Problems, Real Data** → **Targeted Program Execution**

Targeted, optimized, domain-specific solutions →

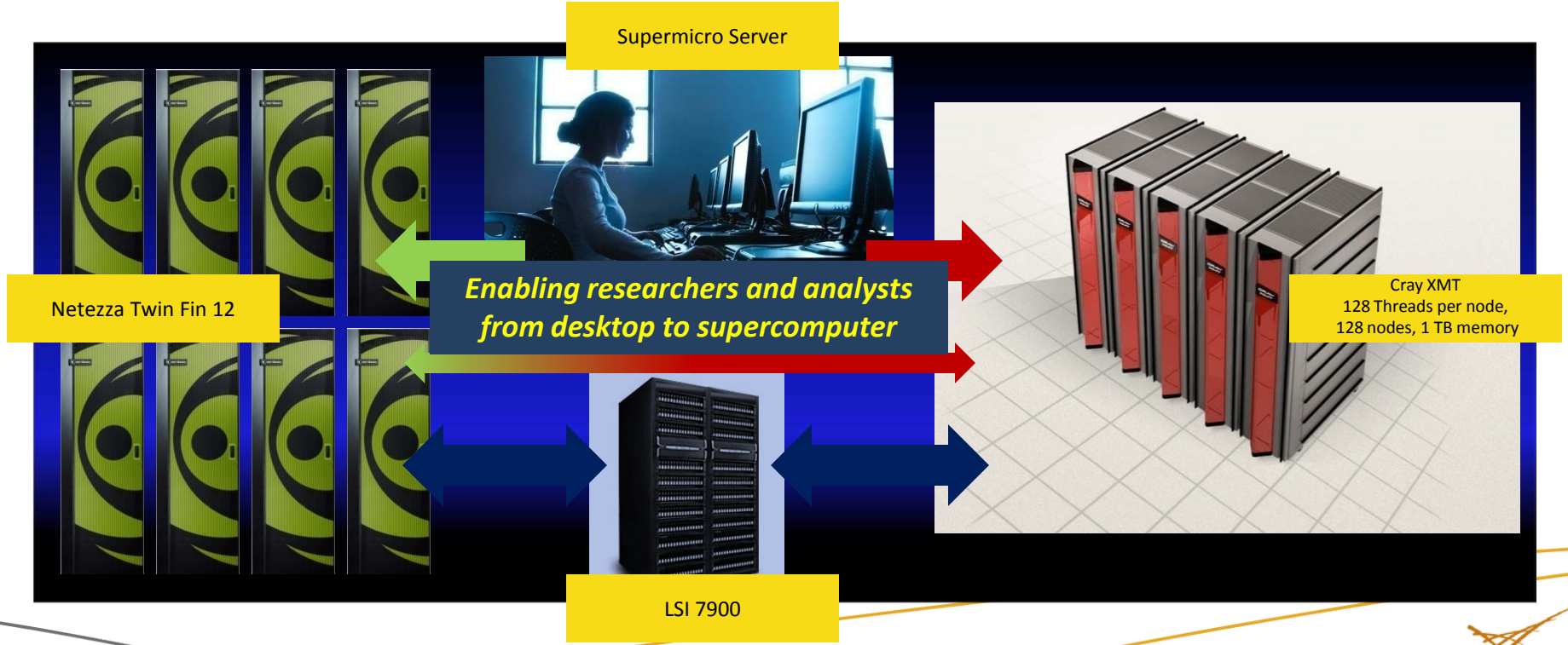
Number of algorithms / kernels decrease →

Size of applications increases →

Mission Driven Operational Programs

Proudly Operated by Battelle *Since 1965*

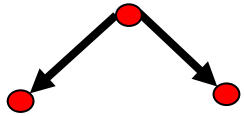
Data intensive supercomputing architecture



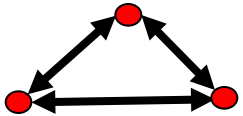
Triadic analysis of network data

Technical POC: George Chin
George.Chin@pnl.gov

Transmission,
Transactions,
Chain of Command



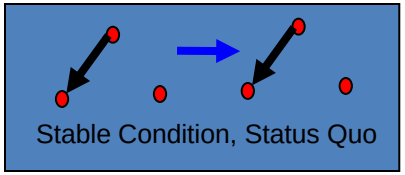
Hierarchies

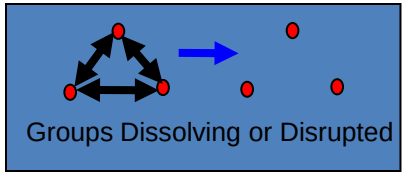


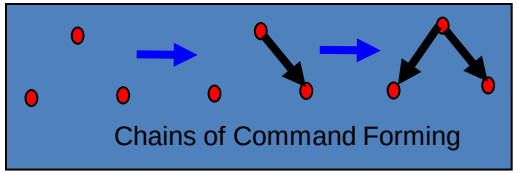
Close-Knit Groups,
Strong Dependencies

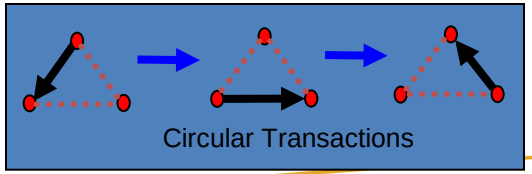
Bridges,
Points of Disruption

Triad patterns can have meaning that changes over time

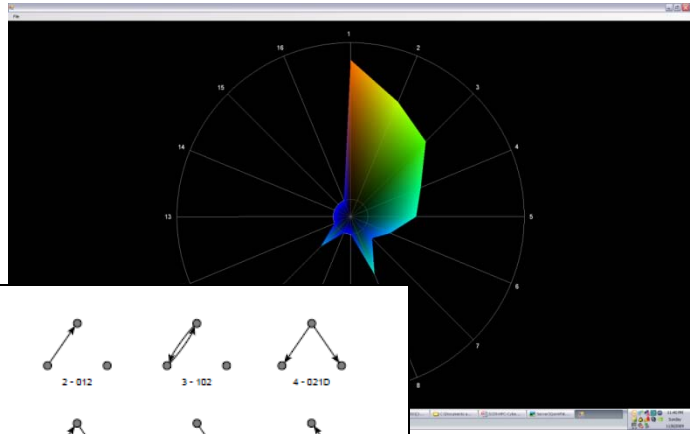


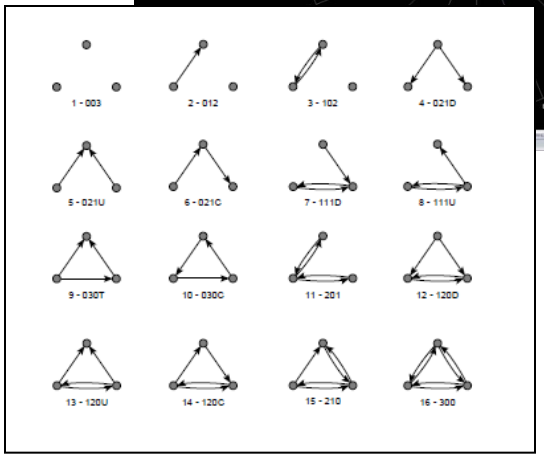




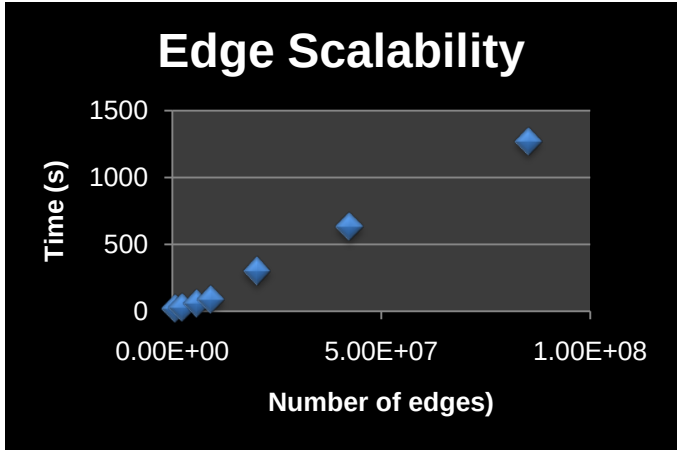


Triad Census



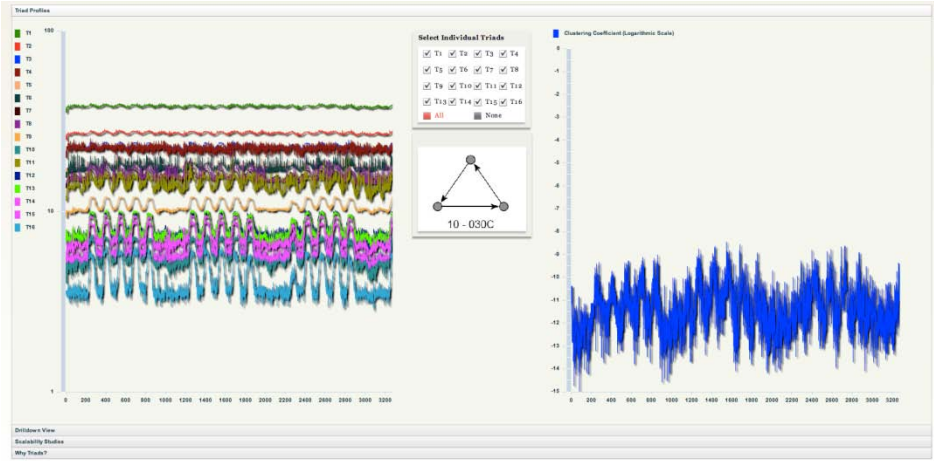


16 possible patterns



- Triad census computes the number of instances of each pattern
- Linear scalability up to 88 Million edges on 64 node Cray XMT
- Over *1.7x10^19* triads calculated
- System utilization is 75% that of LINPACK - one of the highest levels obtained by an *irregular* application and much higher than commodity processor based systems

Drill down capabilities on triad census



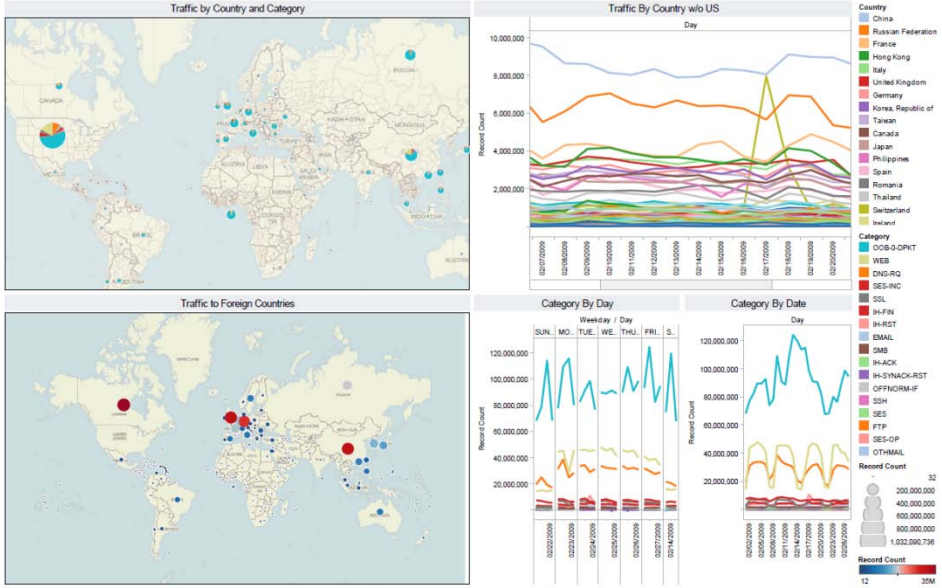
Summary Screen With Triads and Clustering Coefficient Information for entire Dataset

Viewing smaller time range while monitoring a Triad element at global scale

Generalized network intrusion technology (GNITE)

Technical POC:
Terence Critchlow
Terence.Critchlow@pnl.gov

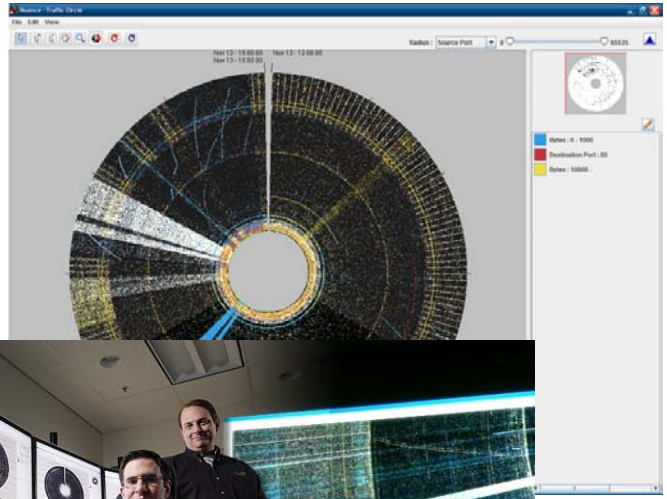
- Summary Information:
  - Aggregation over different views and subsets of the data
  - Flexible, graphical display of ad hoc projections
  - Examples:
    - Periodic reporting
    - Geographic distributions
    - Trend analysis
    - Resource optimization
- Detailed Information:
  - Customized drill-throughs
  - Geospatial analysis
  - Statistical analyses
- Analysts need to seamlessly move between these
- Leverages commercial desktop tools with HPC systems

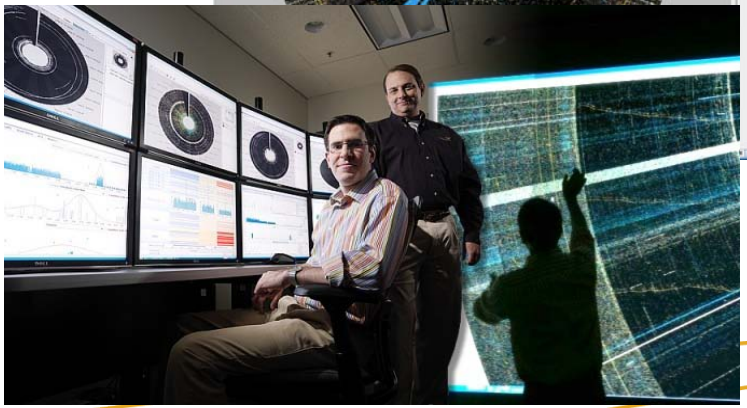


Traffic Circle

Technical POC: Bill Pike
Bill.Pike@pnl.gov

- Interactive visualization of patterns in high volume netflow data
- Scalablity to 100s of millions of network traffic header records over a single day
- Cyber analysts will gain the ability to understand patterns, trends, and characteristic behaviors of high-volume data in real time





Demos

- Triad Census
- GNITE
- Traffic Circle