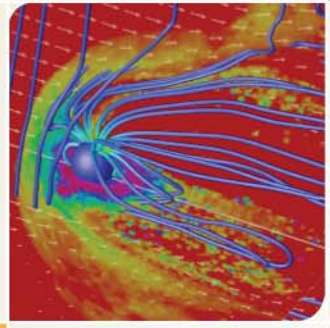
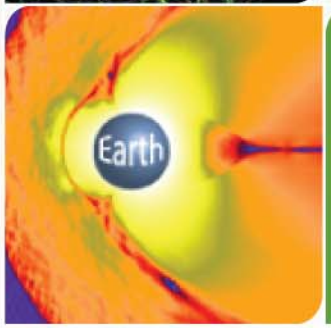
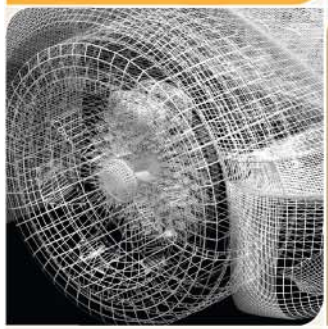





Massive-Scale Streaming Analytics

David A. Bader

Georgia Tech College of Computing
Computational Science and Engineering

Dr. David A. Bader

- Full Professor, Computational Science and Engineering
- Executive Director for High Performance Computing.
- IEEE Fellow, AAAS Fellow
- interests are at the intersection of high-performance computing and real-world applications, including computational biology and genomics and massive-scale data analytics.
- Over $165M of research awards
- Steering Committees of the major HPC conferences, IPDPS and HiPC
- Multiple editorial boards in parallel and high performance computing
  - EIC of IEEE Transactions on Parallel and Distributed Systems
- Elected chair of IEEE and SIAM committees on HPC
- 230+ publications, ≥ 5,600 citations, *h*-index ≥ 43
- National Science Foundation CAREER Award recipient
- Directed the Sony-Toshiba-IBM Center for the Cell/B.E. Processor
- Founder of the Graph500 List for benchmarking "Big Data" computing platforms
- Recognized as a "**RockStar**" of High Performance Computing by InsideHPC in 2012 and as HPCwire's **People to Watch** in 2012 and 2014.

Georgia Tech School of Computational Science and Engineering

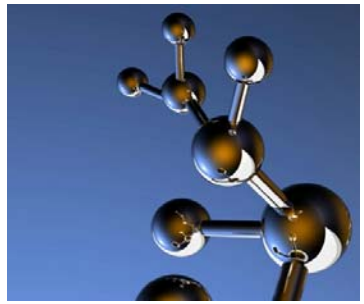
School of Computational Science and Engineering

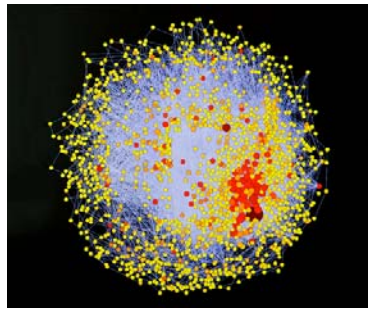
Profile of CSE: Mission

- Solve real-world problems and improve quality of life through advances in computational modeling methods and techniques.
- Apply a collaborative approach to solve hard problems in novel ways through interdisciplinary cooperation and external partnerships.
- Create leaders in government, industry, and academia who will support and advance science and engineering agendas.

Profile of CSE: History

- Founded in 2005, officially recognized as a school in 2010.
- Focus on high performance computing, big data, analytics & visualization, machine learning, cybersecurity.
- $6.00 million in research expenditures; approximately $34 million in active awards (FY 2016)
- NSF South Big Data Hub partnership: $1.25 million over 3 years to support new analysis projects in line with CSE mission.





Profile of CSE: People

- By the numbers (Fall 2016):
  - 40 faculty and staff
  (12 tenure-track faculty)
  - 68 PhD students
  - 99 masters students

- **Award-winning research teams**: ACM Gordon Bell Prize awarded to team composed mainly of CSE faculty and students.
  - Other honors include: 1 Regents' professor, 6 NSF CAREER awards , 3 IEEE fellows, 2 AAAS fellows, and 1 SIAM fellow

2015-2016 Strategic Partnership Program

NORTHROP GRUMMAN

NSA
www.IntelligenceCareers.gov/NSA

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

NVIDIA.

CRAY
THE SUPERCOMPUTER COMPANY

YAHOO! LABS

IBM

accenture
High performance. Delivered.

HPCC SYSTEMS®

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Sandia National Laboratories

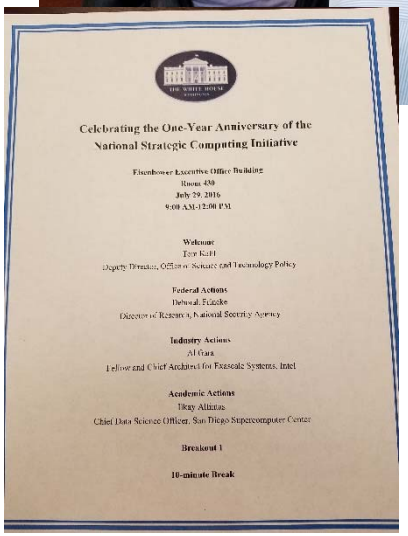
A New Home for the Future of CSE

- Georgia Tech is building Coda, a multi-story, 750,000-square-foot HPC building in the heart of Atlanta (Midtown) with a targeted opening in January 2019
- Devoted to data science and high-performance computing for centralized collaboration among industry, academia and government
- Location of CSE, IDEAS and the HPC Center, and the South BD Hub
- Georgia Tech is the anchor tenant, taking approximately one-half of the new development. Remaining space will be for corporate entities and partners.
- The Institute plans to locate academic and leading-edge research programs in computing and advanced big data analytics there.

National Strategic Computing Initiative

- (29 July 2015, The White House) The National Strategic Computing Initiative (NSCI) is an effort to create a cohesive, multi-agency strategic vision and Federal investment strategy in high-performance computing (HPC).
- This strategy will be executed in collaboration with industry and academia, maximizing the benefits of HPC for the United States.
- HPC systems, through a combination of processing capability and storage capacity, can solve computational problems that are beyond the capability of small- to medium-scale systems. They are vital to the Nation's interests in science, medicine, engineering, technology, and industry.
- The NSCI will spur the creation and deployment of computing technology at the leading edge, helping to advance Administration priorities for economic competiveness, scientific discovery, and national security.
- The National Strategic Computing Initiative has five strategic themes.
  1. Create systems that can apply exaflops of computing power to exabytes of data.
  2. Keep the United States at the forefront of HPC capabilities.
  3. Improve HPC application developer productivity.
  4. Make HPC readily available.
  5. Establish hardware technology for future HPC systems.

NSCI Anniversary Meeting, 29 July 2016



Celebrating the One-Year Anniversary of the National Strategic Computing Initiative

Eisenhower Executive Office Building
Room 450
July 29, 2016
9:00 AM-12:00 PM

Welcome
Tey Kalil
Deputy Director, Office of Science and Technology Policy

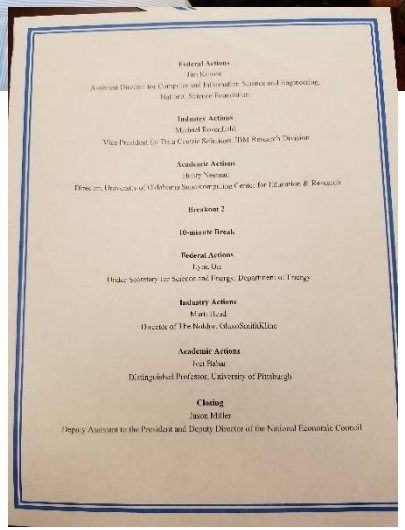
Federal Actions
Dehwal, Fiteke
Director of Research, National Security Agency

Industry Actions
Al Gara
Fellow and Chief Architect for Exascale Systems, Intel

Academic Actions
Ilkay Altintas
Chief Data Science Officer, San Diego Supercomputer Center

Breakout 1

10-minute Break



Federal Actions
Jim Kurose
Assistant Director for Computer and Information Science and Engineering,
National Science Foundation

Industry Actions
Michael Rosenfield
Vice President for Data Center Solutions, IBM Research Division

Academic Actions
Jerry Nauman
Director, University of Oklahoma Supercomputing Center for Education & Research

Breakout 2

10-minute Break

Federal Actions
Lynn Orr
Under Secretary for Science and Energy, Department of Energy

Industry Actions
Mark Head
Director of The Nobbie, GlaxoSmithKline

Academic Actions
Ivet Bahar
Distinguished Professor, University of Pittsburgh

Closing
Jason Miller
Deputy Assistant to the President and Deputy Director of the National Economic Council

Discussion Prompts

- China has deployed its first 100+petaflop system (Sunway TaihuLight) and has indicated a second is in development. What would you change about the NSCI, if anything, given these recent advancements?
- Where do you think we are missing the mark, and how can we improve the initiative to ensure mission success and U.S. leadership?
- The NSCI envisions a robust suite of public-private collaborations between the U.S. Government, industry, and academia. Which strategic objectives are most in need of collaboration, and what steps is the private sector looking for the Federal government to take toward that end?
- How do we change the HPC programming paradigm so that industry and academia can focus their workforce more on the computational problems they are trying to solve, rather than on writing, maintaining, and porting highly-specialized and architecture-dependent codes?
- How can we make HPC resources more widely available? Is cloud computing a viable HPC option, and in what contexts?
- How can we lower the barrier to entry so that converged HPC systems can be broadly deployed to support economic competitiveness and scientific discovery?
- How do we develop innovative, verifiable, scalable, and reusable application software components, architectures, and platforms to both meet mission needs and easily transition to new and evolving NSCI technologies and algorithms?
- How might the HPC and education communities better support training for the current generation and for the next generation of scientists and engineers to adopt HPC as an effective approach to solving problems of societal importance?

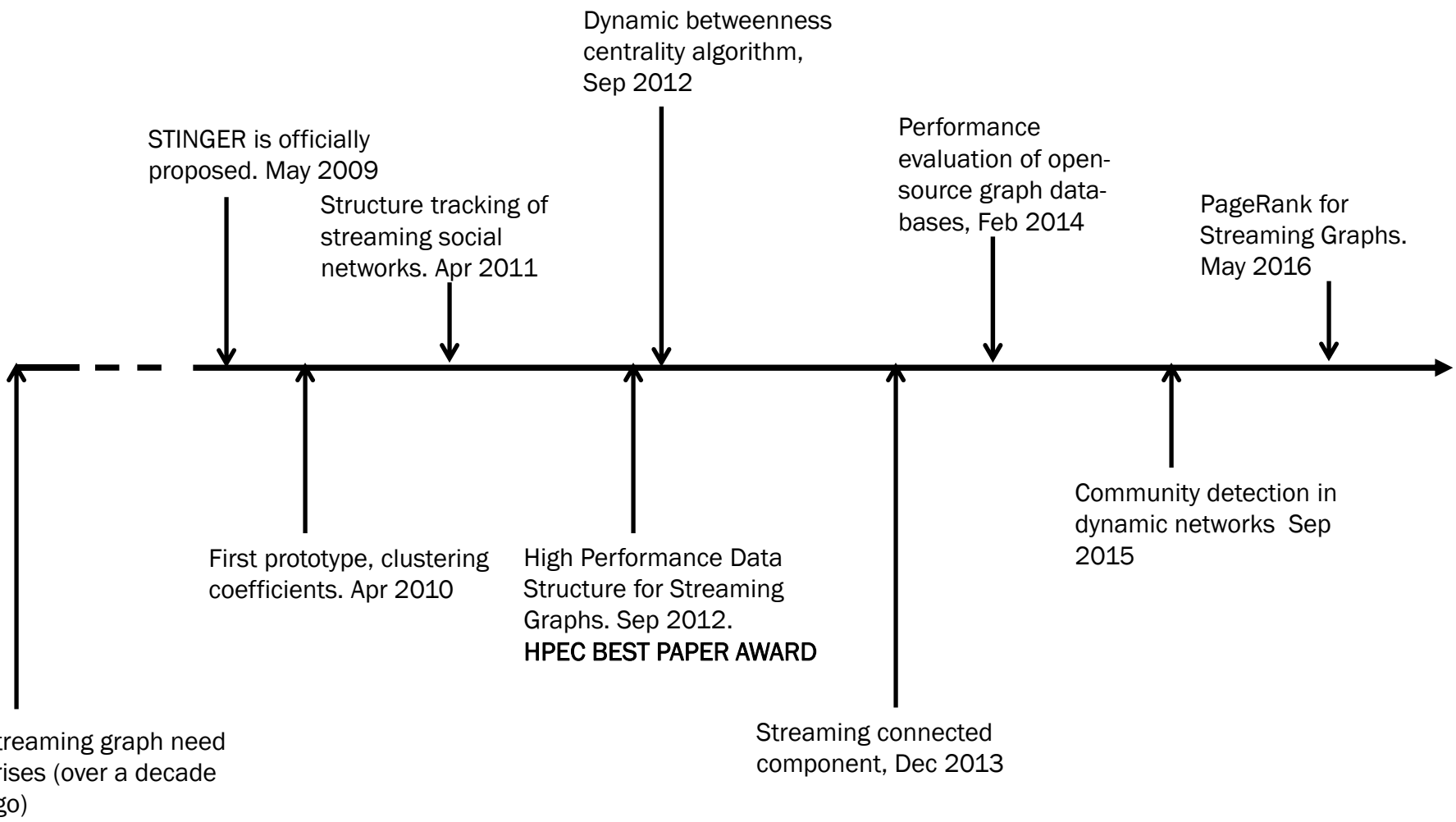
STING Initiative: Focusing on Globally Significant Grand Challenges



- Many globally-significant grand challenges can be modeled by Spatio-Temporal Interaction Networks and Graphs (or "STING").
- Emerging real-world graph problems include
  - detecting community structure in large social networks,
  - defending the nation against cyber-based attacks,
  - discovering insider threats (e.g. Ft. Hood shooter, WikiLeaks),
  - improving the resilience of the electric power grid, and
  - detecting and preventing disease in human populations.

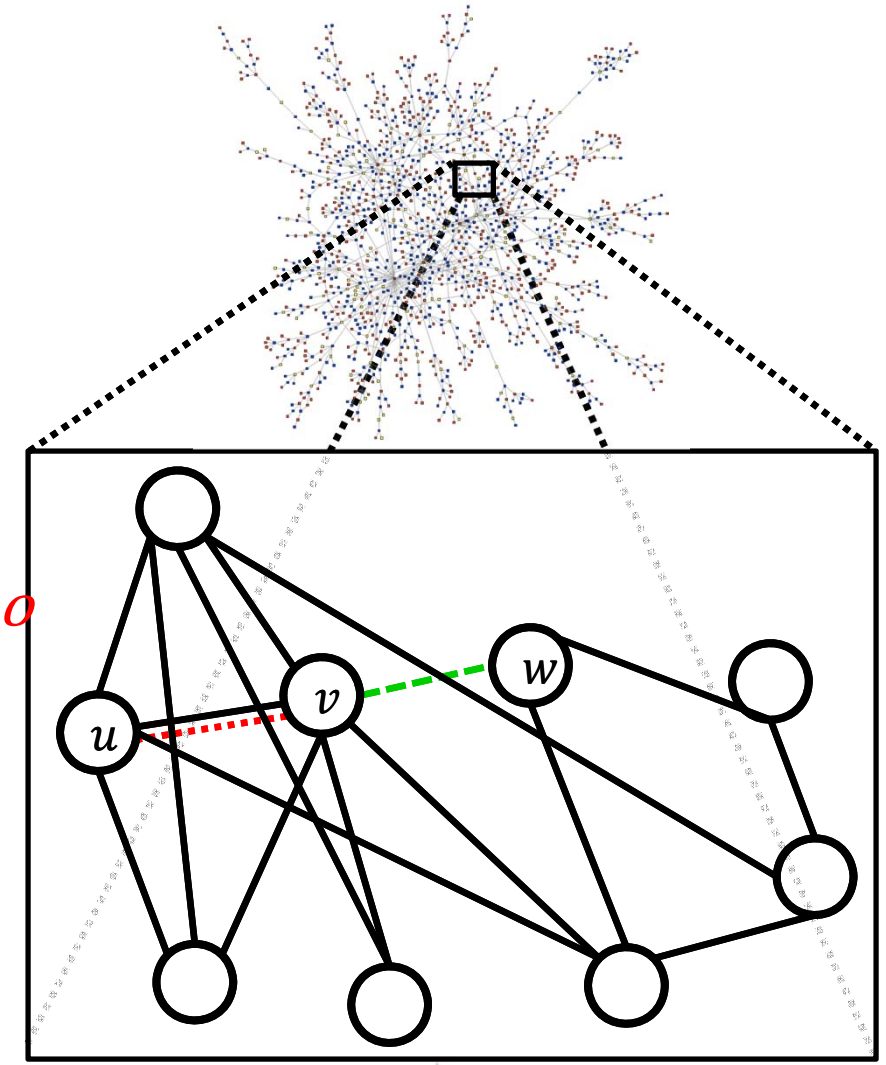
- Unlike traditional applications in computational science and engineering, solving these problems at scale often raises new research challenges because of sparsity and the lack of locality in the massive data, design of parallel algorithms for massive, streaming data analytics, and the need for new exascale supercomputers that are energy-efficient, resilient, and easy-to-program.

STINGER – Time Frame



Streaming graph example

- Dynamic/Streaming:
  - At time t:
    - v and w become friends
    - $Insert\ (v, w)$
  - At time $\hat{t}$:
    - $u\ upsets\ v.\ \ u\ and\ v\ are\ no\ longer\ friends$
    - $Delete\ (u, v)$
- small subgraph…



Traditional HPC Vs. Streaming Analytics

Traditional HPC

- Great for “static” data sets.
- Massive scalability at the cost of programmability.
- Great for dense problems.
  - Sparse problems typically underutilize the system.

Streaming Analytics

- Requires specialized analytics and data structures.
- Data is constantly changing.
- Low data re-usage.
  - Focused on memory operations and not FLOPS.

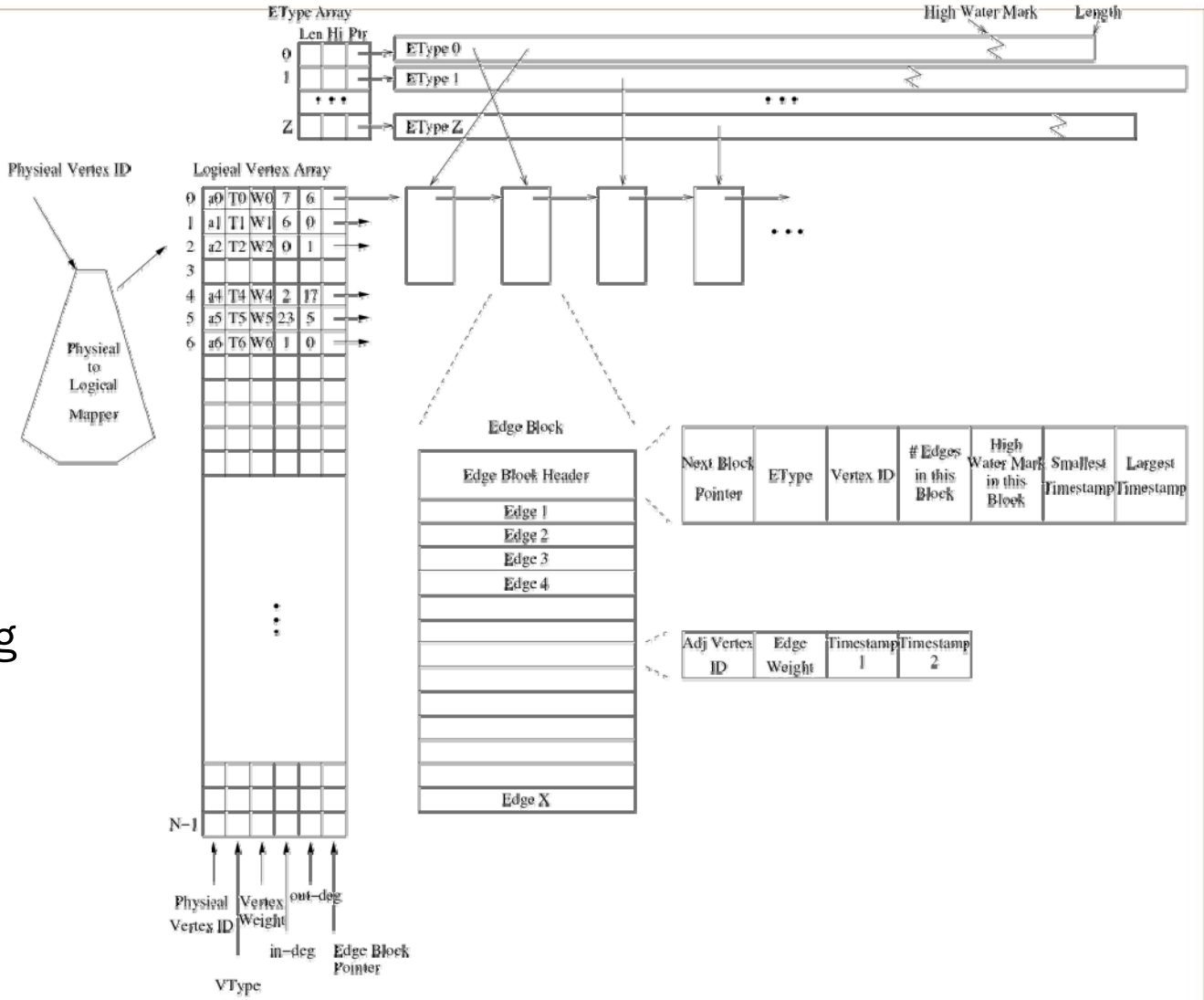


STING Extensible Representation (STINGER)

- Design goals:
  - Enable algorithm designers to implement dynamic graph algorithms with ease.
  - Portable semantics for various platforms
  - Good performance for all types of graph problems and algorithms - static and dynamic.
  - Assumes globally addressable memory access
  - Support multiple, parallel readers and a single writer
    - One server manages the graph data structures
    - Multiple analytics run in background with read-only permissions.

STING Extensible Representation

- Semi-dense edge list blocks with free space
- Compactly stores timestamps, types, weights
- Maps from application IDs to storage IDs
- Deletion by negating IDs, separate compaction



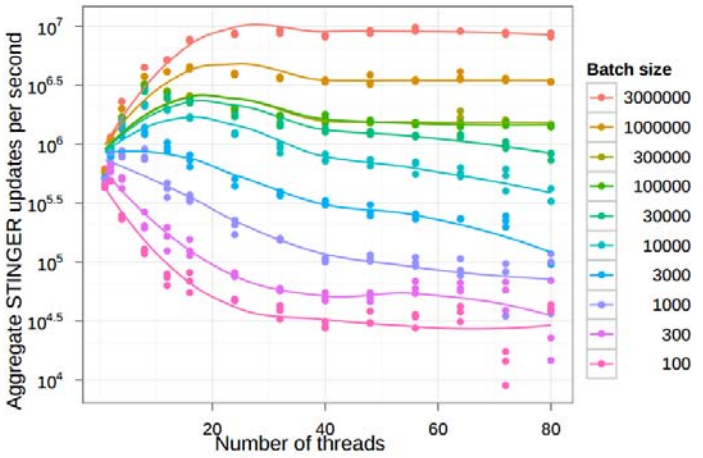
Streaming Updates

Update process

- Group updates into batches
  - Updates can include insertions and deletions
  - Big batches ⇒ Better performances

[HPEC; 2012]

Throughput rate



Experiment setup

- 4x10 Intel E7-8870 processors
- RMAT Graph
  - Vertices: 16M
  - Edges: 128M
- Various batch sizes
  - ~93% of updates are insertions
  - ~7% of updates are deletions

Takeaway

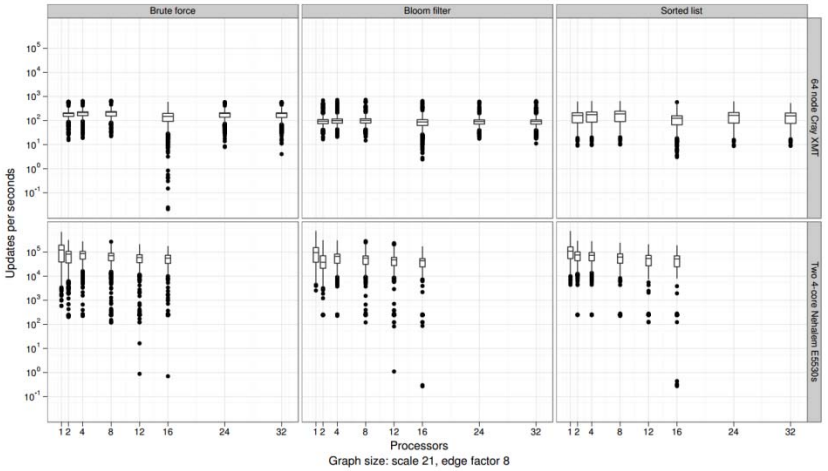
- STINGER supports extremely fast updates.
- Updates are not the bottleneck for analytics.
  - Analytic computations are the bottleneck!
- Highly scalable

Streaming Clustering Coefficients & Triangle Counting

Background

- Scores how tightly bound players are in their local community.
- Looks for common relationships for two adjacent vertices.
  - Hence the term triangle counting
- Complexity for static graph algorithm (intersection based): $O(v \cdot d_{max}^2)$

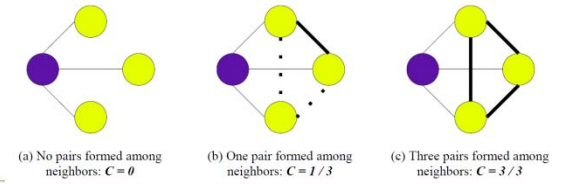
[MTAAP; 2010]



Multiple streaming implementations

- Brute-force – straightforward and exact
- Bloom-filter – approximate yet extremely fast
- Sorted-list – uses intersections. Fast and exact.

Larger batches give faster speedups.



Experiment setup

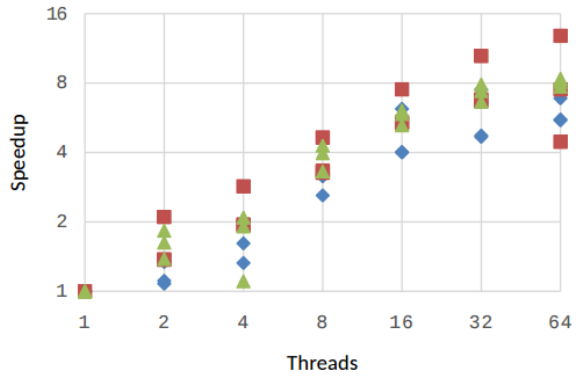
- Executed on two systems
  - Cray XMT – 64 nodes
  - 2x4 Intel E5530 system
    - 8 cores, 16 threads
- Used RMAT synthetic graphs
  - 2M vertices, 16 edges
- Hundreds of thousands updates per second

Streaming Connected Components

Background

- Tracks connected components in high velocity networks.
- Connected components imply that players are connected to each other some sequence of relationships

[HiPC; 2013]



Average edge degree: ◆8 ■16 ▲32

Our algorithm

- Takes into account small-world property
  - Diameter is a small.
  - Most players have numerous relationships within the connected component.
  - Edge insertions are always easy.
  - Very edge deletions are complex.

Takeaway

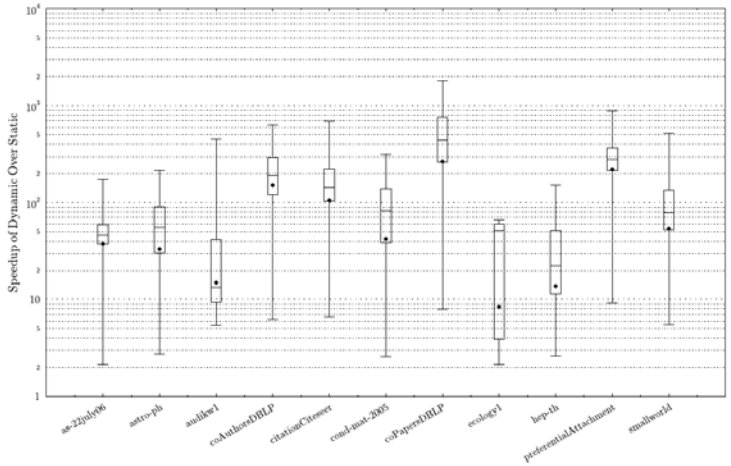
- Up to 1.26 million updates per second on 4×16 AMD (Opteron 6282)
- Hundreds of time faster than static computation.
- Great for social networks.
- STINGER requires only 10% of execution time. Rest of time - analytic update.
- Scalability similar to BFS.

Dynamic Betweenness Centrality

Background

- Used for finding key players in network based on the number of relationships that go through them.
- Fastest known algorithm by Brandes (2002) is still computationally expensive for large networks: $O\big(V \cdot (V + E)\big)$.

[Social Computing; 2012]



Our dynamic graph algorithm

- Supports optimizations:
  - Approximation: reduces accuracy, significantly faster.
  - Parallelization: utilizes many core systems
- Supports: vertex insertions & deletions and edge insertions & deletions.

Experimental setup and takeaways

- 4x10 Intel E7-8870 processors
- Thousands of times faster than static recomputations.
- Significantly reduces the amount of necessary computations.
- Only small percentage of the graph is affected due to update

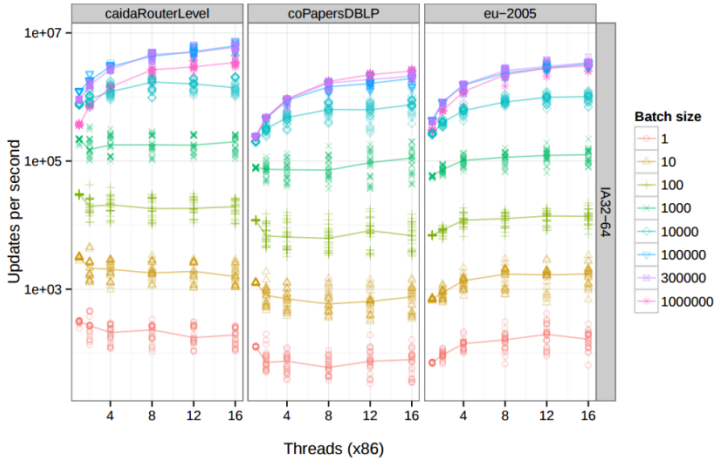
Streaming Community Detection and Monitoring



Background

- Communities are typically defined by groups of vertices with more intra-relationships than inter-relationships.
  - More formally: $Q(C) = \frac{Intra_C}{|E|} - \frac{Inter_C^2}{|E|^2}$
- In addition to the graph, an additional *community* network is maintained.
  - Significantly smaller than full network!
  - Updates applied to community network.

[MTAAP; 2013]



An agglomerative approach

- Certain types of updates do not change the community structure.
  - We only need to process updates that *"might"* cause change.
  - Few updates require a lot of process time.

Experimental setup and takeaways

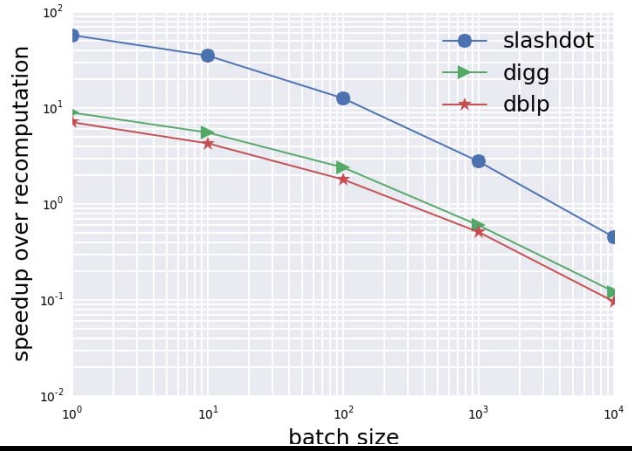
- 4x8 Intel E7-4820 system
  - 32 cores, 64 threads
- Easily supports millions of updates per second.
- Bigger batches offer improved performance.
- Dynamic algorithm is 1000s of time faster than static graph algorithm.
- Real-time tracking of communities with a network.

Streaming Seed-Set Expansion

Background

- Seeds are vertices of interest.
- Seed Set Expansion is the process of detecting a community around a seed.
- Streaming SSE – allows tracking vertices of interest over time
  - Important events such as community split and merging can be reported

[ASONAM;2015]



Algorithm details

- Greedy algorithm.
  - Vertices are inserted one at a time into the community.
- An edge update checks for possible changes in the community.
- Pruning can be applied when an update causes a big change in the community.
  - Pruning makes things slower
  - Pruning offers more accurate results in comparison with static graph algorithm.

Takeaways

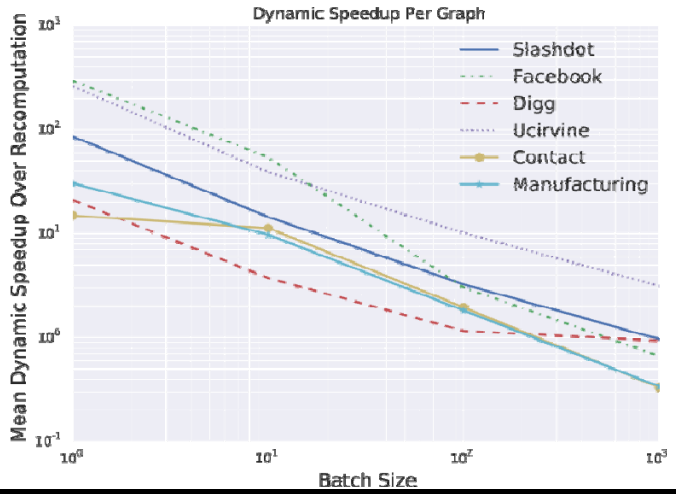
- Highly accurate in comparison to static graph algorithm.
  - Precision and recall typically above 90%.
- Larger batches require more work ⇒ smaller speedups

Incremental Page-Rank

Background

- Pagerank is used by measuring the importance of vertices by the number and weight of links going through it.
- Works like a propagation algorithm.
  - Algorithm continues until no changes are detected.

[GABB;2016]



Algorithm

- Uses STINGER to perform linear algebra operations.
- Supports both insertions and deletions.
- Incremental implies that only a small subset of the graph is traversed.
  - Does only the necessary amount of work

Takeaway

- Large batches: reduce lower latency by > 2× over restarting on average.
- Small batches: potentially hundreds of time faster than restart.
- Improved power performance (modeled).
- Can deal with several thousand updates per second.

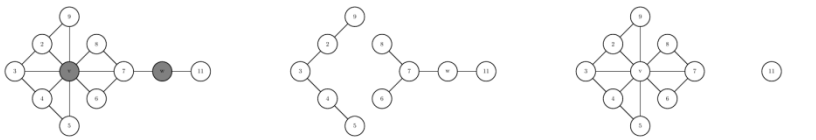
Community Centric Analysis (in process)

Background

- Focuses on finding key players in communities
  - Might be overlooked by network wide analytics.
  - Computationally less expensive.
  - Highly scalable
  - Key players detected due to change to their community upon extraction.
- We modify several widely used analytics for this new type of computation.
- Starts off with an initial exploration of static graphs

Modified metrics of interest

- Change in community modularity
- Change to the community diameter
- Change in the number of connected components in the community



(a) Initial community. (b) Vertex v is removed from community. (c) Vertex w is removed from community.

Community Centric Approach

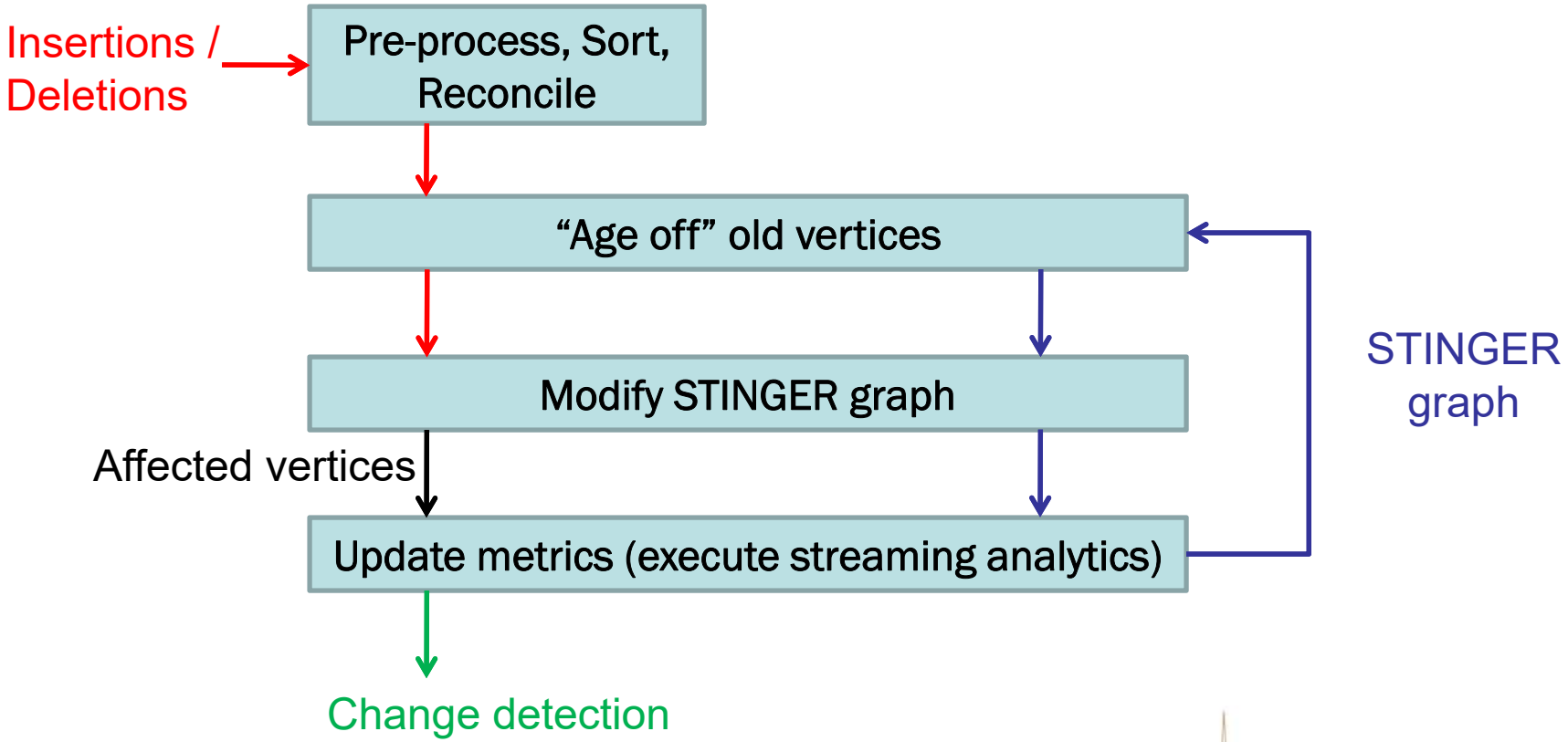
- Given a community C and a metric M, for each vertex u in each community C:
  - Calculate initial metric $M_{initial}$ on community (left figure) using static graph algorithm **(done once)**
  - Remove vertex u and links using **STINGER**
  - Calculate changed metric M_{after} using $dynamic$ graph algorithm using **STINGER**
  - Look at change to community: $\Delta M_u = \frac{M_{after}}{M_{initial}}$
  - Insert vertex u and links using **STINGER**

Initial Findings

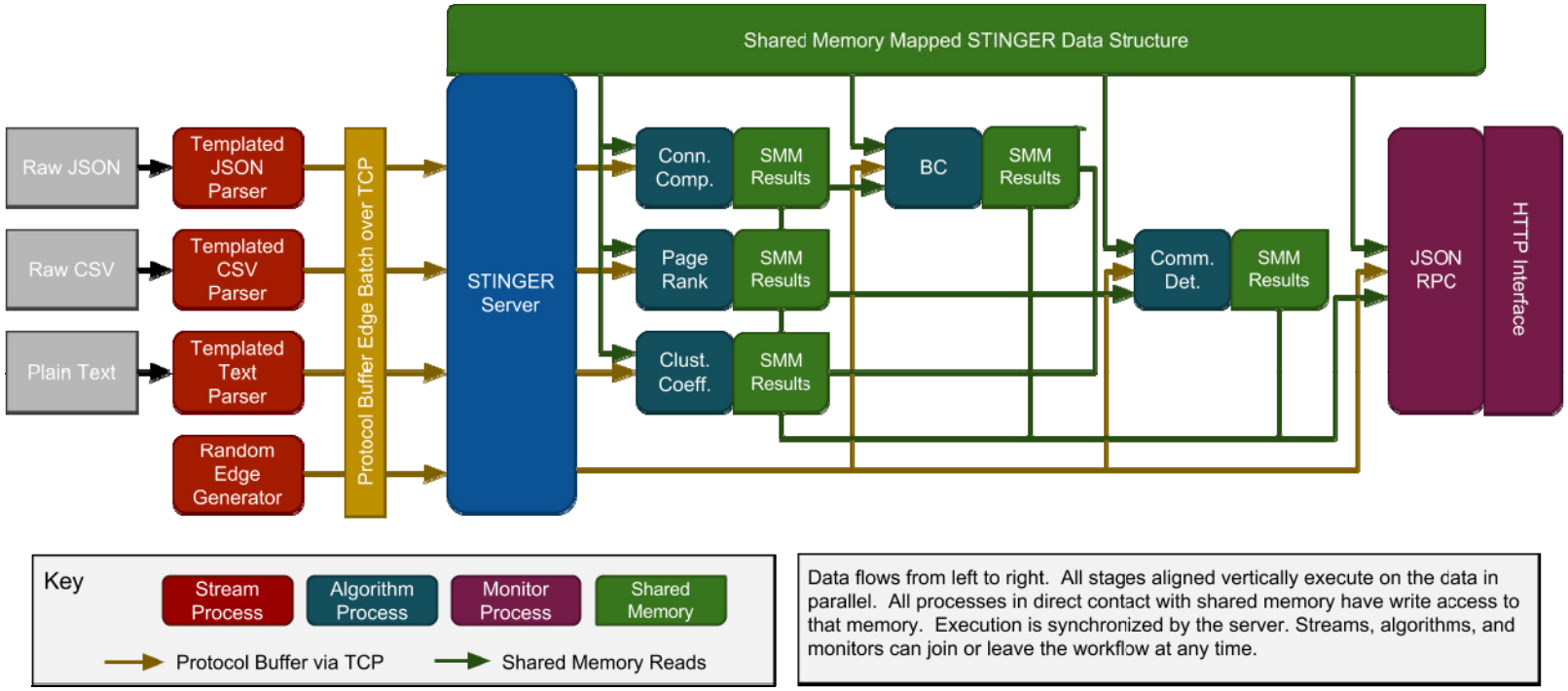
- A different way to use streaming analytics: "$vertex\ delta\ computations$"
- Multiple metrics pinpoint same key vertices.
- Computationally efficient
  - Over $20X$ faster than networks approach.
  - Highly scalable
- Applicable to other metrics as well

STINGER Graph & Analytic Update Process

Accumulate recent graph updates in main memory and create a batch.



STING: High-level architecture



- Server: Graph storage, kernel orchestration
- OpenMP + sufficiently POSIX-ish
- Multiple processes for resilience



STINGER: as an analysis package

- **Streaming edge insertions and deletions:**
  Performs new edge insertions, updates, and deletions in batches or individually.
  Optimized to update at rates of over 3 million edges per second on graphs of one billion edges.
- **Streaming clustering coefficients:**
  Tracks the local and global clustering coefficients of a graph.
- **Streaming connected components:**
  Real time tracking of the connected components.
- **Streaming Betweenness Centrality:**
  Find the key points within information flows and structural vulnerabilities.
- **Streaming community detection:**
  Track and update the community structures within the graph as they change.

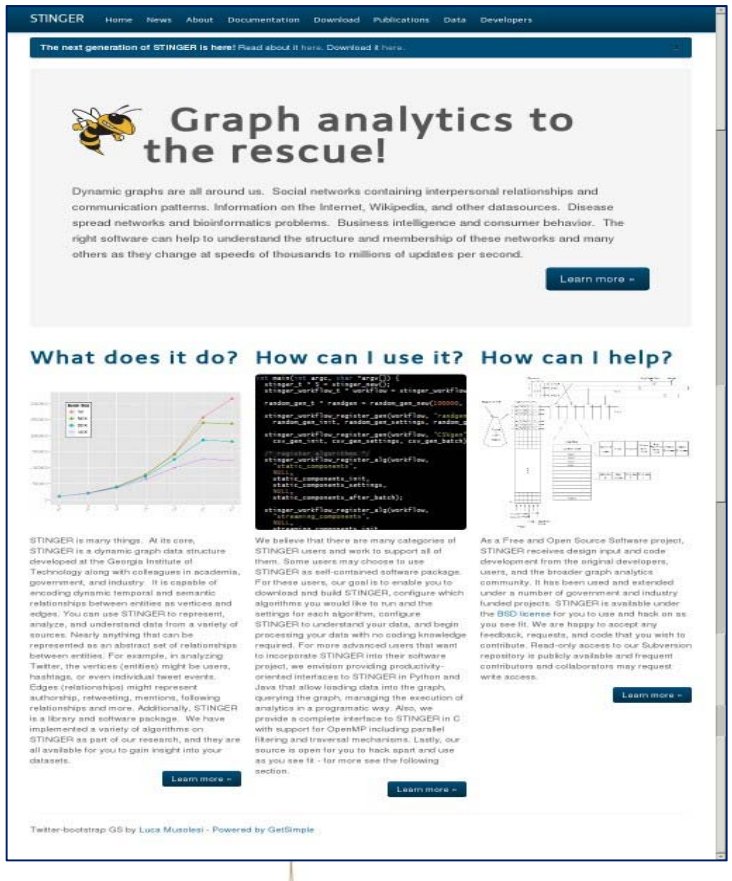
- **Anything that a static graph package can do (and a whole lot more):**
  - **Parallel agglomerative clustering:**
    Find clusters that are optimized for a user-defined edge scoring function.
  - **K-core Extraction:**
    Extract additional communities and filter noisy high-degree vertices.
  - **Classic breadth-first search:**
    Performs a parallel breadth-first search of the graph starting at a given source vertex to find shortest paths.
  - **Parallel connected components:**
    Finds the connected components in a static network.

http://www.stingergraph.com/

STINGER: Where do you get it?

http://www.stingergraph.com/

- Gateway to
  - code,
  - development,
  - documentation,
  - presentations…
- Users / contributors / questioners: Georgia Tech, PNNL, CMU, Berkeley, Intel, Cray, NVIDIA, IBM, Federal Government, Ionic Security, Citi, Accenture



STINGER Development

Academic

- Maintained by Georgia Tech
- Ideal for prototyping.
- Sandbox for developing new concepts
- When software matures…

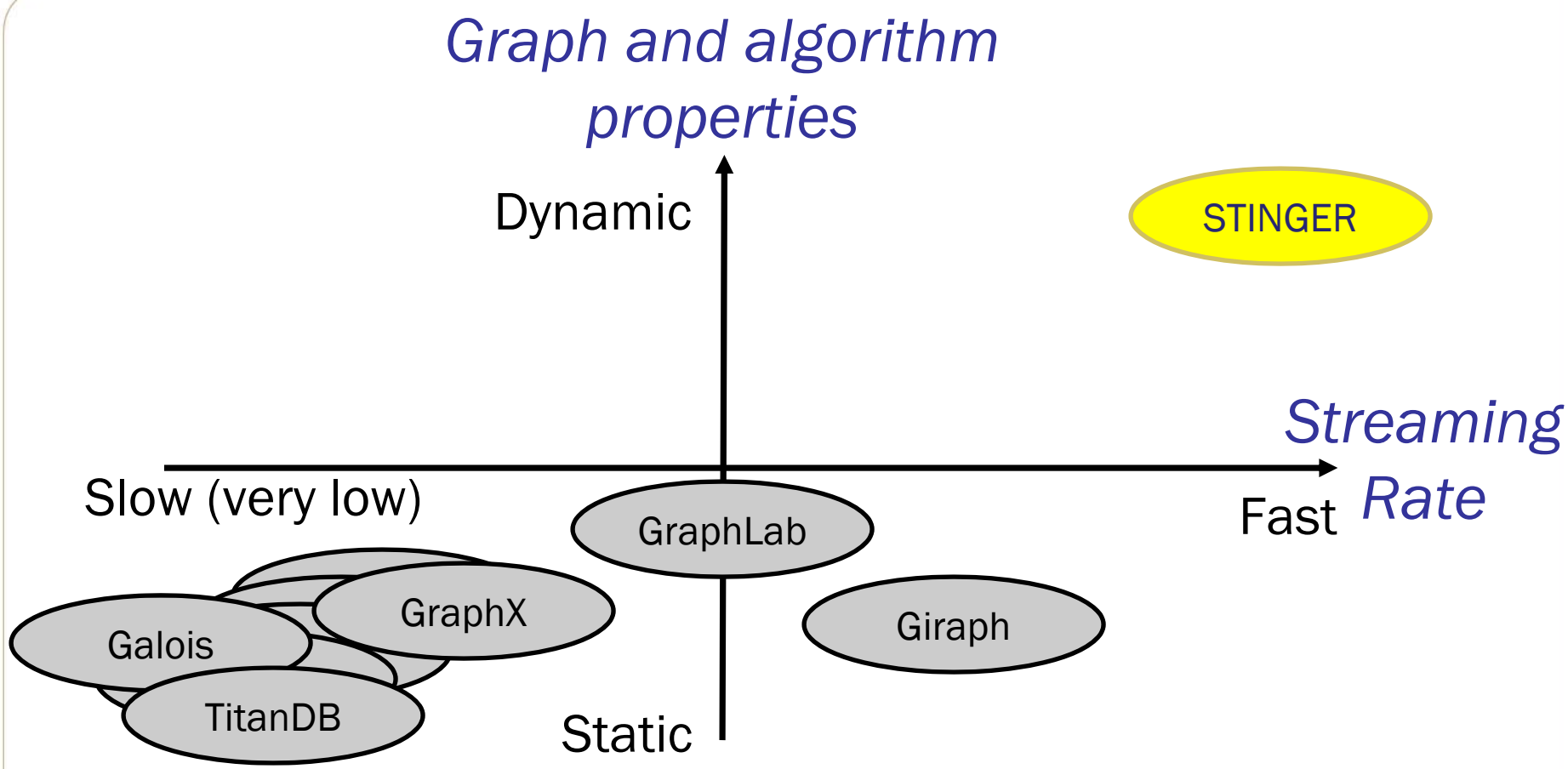
Enterprise

- Tech transfer for GTRI
- Enterprise software integrity
  - Nightly builds
  - Unit testing required

http://git.cc.gatech.edu/git/u/eriedy3/stinger.git/

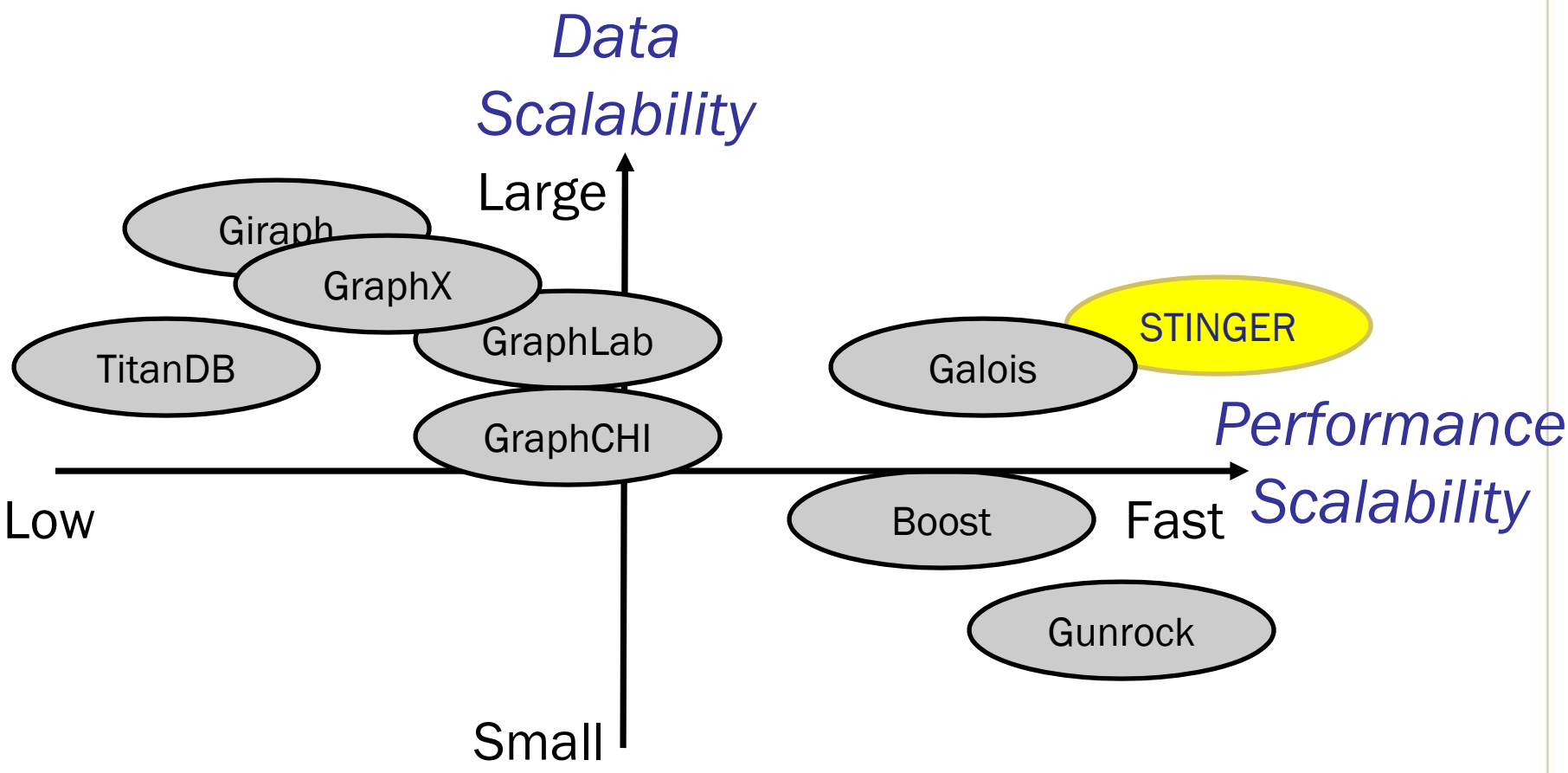
https://github.com/stingergraph

Graph Analytics (Dynamic vs. Streaming)



- Many libraries support updates to graph
  - They do not have dynamic graph analytics

Scalability (Volume vs. Performance)



- Many libraries support updates to graph
  - They do not have dynamic graph analytics

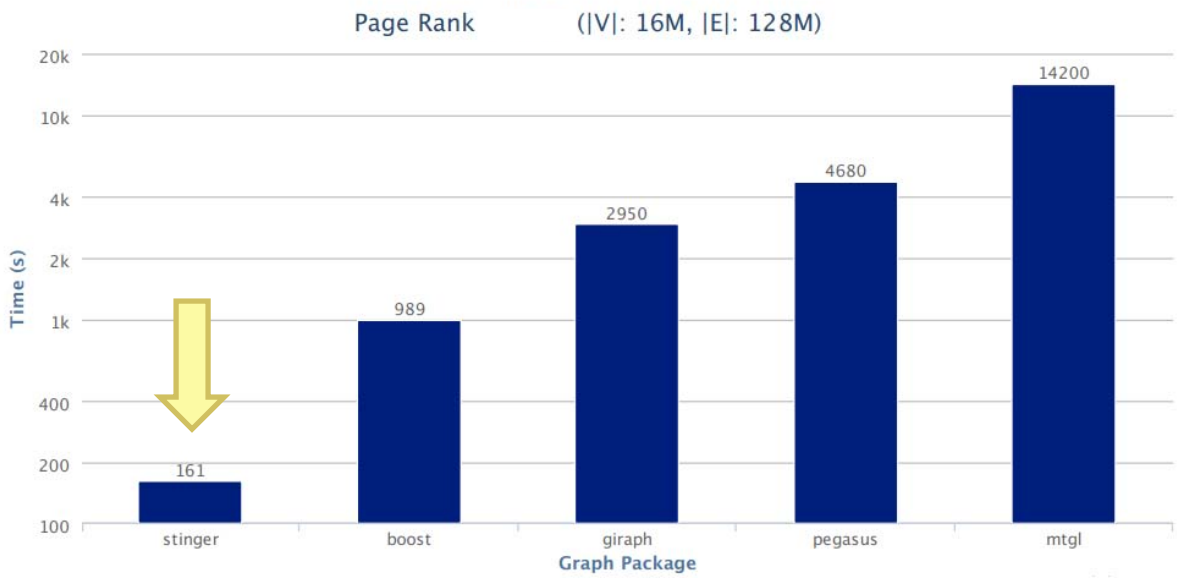


Graph Library Comparison

Framework	Platform	Dynamic Data Structure/Insertions/Deletions			Dynamic Algorithms	Semantic Support	Utilization	Data scalability	Absolute Performance
STINGER	Shared	✓	✓	✓	●	●	●	◔	●
Galois	Shared	X	---	---	○	○	●	◔	●
Ligra	Shared	X	---	---	○	○	●	◔	●
Boost	Shared/ Distributed	X	---	---	○	○	◔	◑	◔
Gunrock	Shared/ Distributed	X	---	---	○	○	●	◕	●
Giraph	Distributed				○	○	◕	●	◕
GraphX	Distributed	X	---	---	○	○	◕	●	◕
GraphLab	Distributed	X	---	---	○	○	◕	●	◕
Llama	Shared	✓	✓	✓	◑	○	◑	◑	◑

PageRank - Performance Analysis

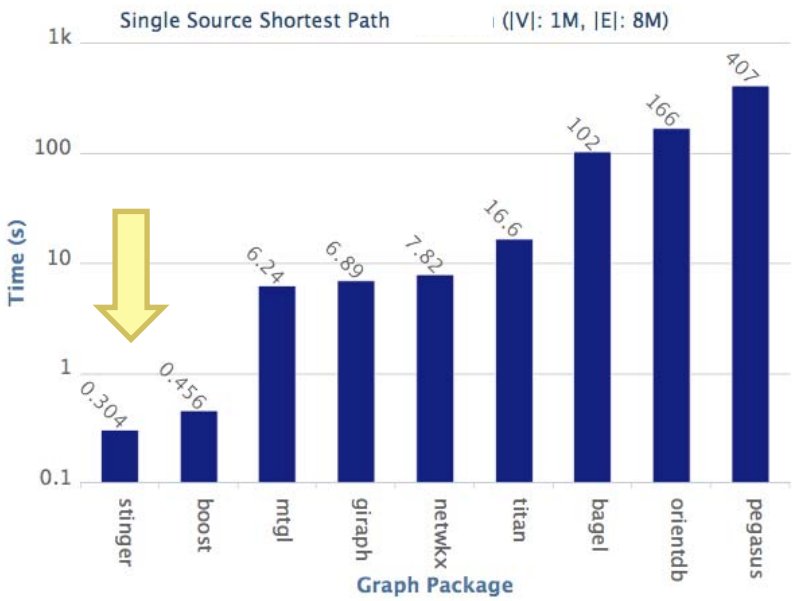
R-MAT Graph
- Vertices: **16M**
- Edges: **128M**



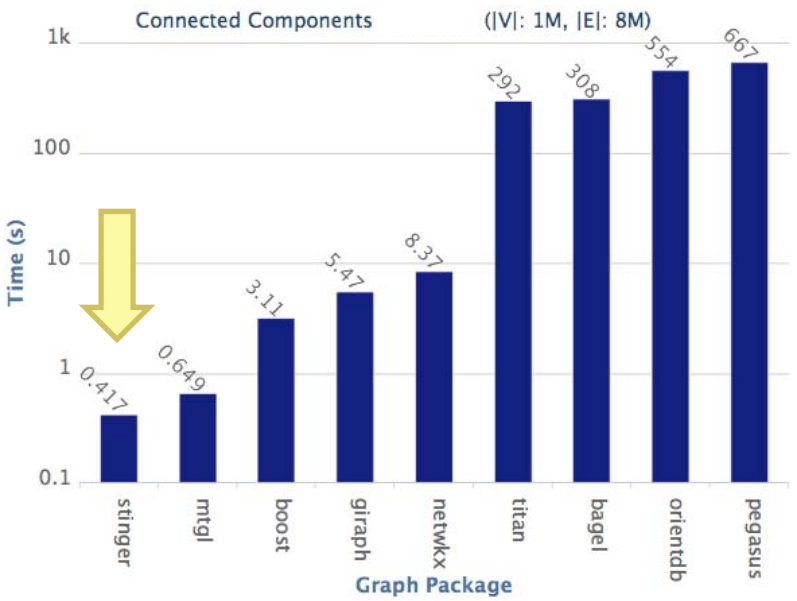
- Lower is better.
- 2 orders of magnitude difference in performance.
- Still outperforms other static-only graph packages.
- Outperforms the distributed systems even for large networks with plenty of computational demand!
- Some platforms did not complete in reasonable amount of time.

Other algorithms

Static Single-Source Shortest Path



Static Connected Components



R-MAT Graph

- Vertices: **1M**
- Edges: **8M**

- STINGER is orders of magnitude faster.
- Still outperforms other static-only graph packages.



STINGER Summary

- Massive-Scale Streaming Analytics require
  - Simple programming model
    - Simple API.
    - CSR-like in concept.
    - STINGER has a lot more under the hood.
  - Extremely fast updates
    - Millions of updates per second.
    - These must not be bottlenecks for updating an analytic.
    - STINGER offers these
- STINGER has major performance benefits
  - Thousands of times faster than static graph computation.
  - Hundreds of thousands of updates per second for numerous analytics.
  - Real-time monitoring of underlying network.



Conclusions

- Massive-Scale Streaming Analytics will require new
  - High-performance computing platforms
  - Streaming algorithms
  - Energy-efficient implementations

  and are promising to solve **real-world challenges**!
- Mapping applications to high performance architectures may yield 6 or more orders of magnitude performance improvement



Acknowledgments

- Jason Riedy, Research Scientist
- Oded Green, Research Scientist
- Graduate Students (Georgia Tech):
  - Rob McColl
  - Eisha Nathan
  - Anita Zakrzewska
- Bader Alumni:
  - Dr. David Ediger (GTRI)
  - Dr. James Fairbanks (GTRI)
  - Dr. Oded Green (GT)
  - Dr. Seunghwa Kang (Pacific Northwest National Lab)
  - Prof. Kamesh Madduri (Penn State)
  - Dr. Guojing Cong (IBM TJ Watson Research Center)



Bader, Related Recent Publications (2005-2009)

- D.A. Bader, G. Cong, and J. Feo, "On the Architectural Requirements for Efficient Execution of Graph Algorithms," *The 34th International Conference on Parallel Processing* (ICPP 2005), pp. 547-556, Georg Sverdrups House, University of Oslo, Norway, June 14-17, 2005.
- D.A. Bader and K. Madduri, "Design and Implementation of the HPCS Graph Analysis Benchmark on Symmetric Multiprocessors," *The 12th International Conference on High Performance Computing* (HiPC 2005), D.A. Bader *et al.*, (eds.), Springer-Verlag LNCS 3769, 465-476, Goa, India, December 2005.
- D.A. Bader and K. Madduri, "Designing Multithreaded Algorithms for Breadth-First Search and st-connectivity on the Cray MTA-2," *The 35th International Conference on Parallel Processing* (ICPP 2006), Columbus, OH, August 14-18, 2006.
- D.A. Bader and K. Madduri, "Parallel Algorithms for Evaluating Centrality Indices in Real-world Networks," *The 35th International Conference on Parallel Processing* (ICPP 2006), Columbus, OH, August 14-18, 2006.
- K. Madduri, D.A. Bader, J.W. Berry, and J.R. Crobak, "Parallel Shortest Path Algorithms for Solving Large-Scale Instances," *9th DIMACS Implementation Challenge -- The Shortest Path Problem*, DIMACS Center, Rutgers University, Piscataway, NJ, November 13-14, 2006.
- K. Madduri, D.A. Bader, J.W. Berry, and J.R. Crobak, "An Experimental Study of A Parallel Shortest Path Algorithm for Solving Large-Scale Graph Instances," *Workshop on Algorithm Engineering and Experiments* (ALENEX), New Orleans, LA, January 6, 2007.
- J.R. Crobak, J.W. Berry, K. Madduri, and D.A. Bader, "Advanced Shortest Path Algorithms on a Massively-Multithreaded Architecture," *First Workshop on Multithreaded Architectures and Applications* (MTAAP), Long Beach, CA, March 30, 2007.
- D.A. Bader and K. Madduri, "High-Performance Combinatorial Techniques for Analyzing Massive Dynamic Interaction Networks," *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, DIMACS Center, Rutgers University, Piscataway, NJ, September 24-25, 2007.
- D.A. Bader, S. Kintali, K. Madduri, and M. Mihail, "Approximating Betewenness Centrality," The *5th Workshop on Algorithms and Models for the Web-Graph* (WAW2007), San Diego, CA, December 11-12, 2007.
- David A. Bader, Kamesh Madduri, Guojing Cong, and John Feo, "Design of Multithreaded Algorithms for Combinatorial Problems," in S. Rajasekaran and J. Reif, editors, *Handbook of Parallel Computing: Models, Algorithms, and Applications*, CRC Press, Chapter 31, 2007.
- Kamesh Madduri, David A. Bader, Jonathan W. Berry, Joseph R. Crobak, and Bruce A. Hendrickson, "Multithreaded Algorithms for Processing Massive Graphs," in D.A. Bader, editor, *Petascale Computing: Algorithms and Applications*, Chapman & Hall / CRC Press, Chapter 12, 2007.
- D.A. Bader and K. Madduri, "SNAP, Small-world Network Analysis and Partitioning: an open-source parallel graph framework for the exploration of large-scale networks," *22nd IEEE International Parallel and Distributed Processing Symposium* (IPDPS), Miami, FL, April 14-18, 2008.
- S. Kang, D.A. Bader, "An Efficient Transactional Memory Algorithm for Computing Minimum Spanning Forest of Sparse Graphs," 14th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP), Raleigh, NC, February 2009.
- Karl Jiang, David Ediger, and David A. Bader. "Generalizing k-Betweenness Centrality Using Short Paths and a Parallel Multithreaded Implementation." The 38th International Conference on Parallel Processing (ICPP), Vienna, Austria, September 2009.
- Kamesh Madduri, David Ediger, Karl Jiang, David A. Bader, Daniel Chavarría-Miranda. "A Faster Parallel Algorithm and Efficient Multithreaded Implementations for Evaluating Betweenness Centrality on Massive Datasets." 3rd Workshop on Multithreaded Architectures and Applications (MTAAP), Rome, Italy, May 2009.
- David A. Bader, et al. "STINGER: Spatio-Temporal Interaction Networks and Graphs (STING) Extensible Representation." 2009.



Bader, Related Recent Publications (2010-2011)

- David Ediger, Karl Jiang, E. Jason Riedy, and David A. Bader. "Massive Streaming Data Analytics: A Case Study with Clustering Coefficients," Fourth Workshop in Multithreaded Architectures and Applications (MTAAP), Atlanta, GA, April 2010.
- Seunghwa Kang, David A. Bader. "Large Scale Complex Network Analysis using the Hybrid Combination of a MapReduce cluster and a Highly Multithreaded System:," Fourth Workshop in Multithreaded Architectures and Applications (MTAAP), Atlanta, GA, April 2010.
- David Ediger, Karl Jiang, Jason Riedy, David A. Bader, Courtney Corley, Rob Farber and William N. Reynolds. "Massive Social Network Analysis: Mining Twitter for Social Good," The 39th International Conference on Parallel Processing (ICPP 2010), San Diego, CA, September 2010.
- Virat Agarwal, Fabrizio Petrini, Davide Pasetto and David A. Bader. "Scalable Graph Exploration on Multicore Processors," *The 22nd IEEE and ACM Supercomputing Conference* (SC10), New Orleans, LA, November 2010.
- Z. Du, Z. Yin, W. Liu, and D.A. Bader, "On Accelerating Iterative Algorithms with CUDA: A Case Study on Conditional Random Fields Training Algorithm for Biological Sequence Alignment," IEEE International Conference on Bioinformatics & Biomedicine, Workshop on Data-Mining of Next Generation Sequencing Data (NGS2010), Hong Kong, December 20, 2010.
- D. Ediger, J. Riedy, H. Meyerhenke, and D.A. Bader, "Tracking Structure of Streaming Social Networks," 5th Workshop on Multithreaded Architectures and Applications (MTAAP), Anchorage, AK, May 20, 2011.
- D. Mizell, D.A. Bader, E.L. Goodman, and D.J. Haglin, "Semantic Databases and Supercomputers," 2011 Semantic Technology Conference (SemTech), San Francisco, CA, June 5-9, 2011.
- P. Pande and D.A. Bader, "Computing Betweenness Centrality for Small World Networks on a GPU," *The 15th Annual High Performance Embedded Computing Workshop* (HPEC), Lexington, MA, September 21-22, 2011.
- David A. Bader, Christine Heitsch, and Kamesh Madduri, "Large-Scale Network Analysis," in J. Kepner and J. Gilbert, editor, *Graph Algorithms in the Language of Linear Algebra*, SIAM Press, Chapter 12, pages 253-285, 2011.
- Jeremy Kepner, David A. Bader, Robert Bond, Nadya Bliss, Christos Faloutsos, Bruce Hendrickson, John Gilbert, and Eric Robinson, "Fundamental Questions in the Analysis of Large Graphs," in J. Kepner and J. Gilbert, editor, *Graph Algorithms in the Language of Linear Algebra*, SIAM Press, Chapter 16, pages 353-357, 2011.



Bader, Related Recent Publications (2012)

- E.J. Riedy, H. Meyerhenke, D. Ediger, and D.A. Bader, "**Parallel Community Detection for Massive Graphs**," The 9th International Conference on Parallel Processing and Applied Mathematics (PPAM 2011), Torun, Poland, September 11-14, 2011. Lecture Notes in Computer Science, 7203:286-296, 2012.
- E.J. Riedy, D. Ediger, D.A. Bader, and H. Meyerhenke, "**Parallel Community Detection for Massive Graphs**," 10th DIMACS Implementation Challenge -- Graph Partitioning and Graph Clustering, Atlanta, GA, February 13-14, 2012.
- E.J. Riedy, H. Meyerhenke, D.A. Bader, D. Ediger, and T. Mattson, "**Analysis of Streaming Social Networks and Graphs on Multicore Architectures**," The 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan, March 25-30, 2012.
- J. Riedy, H. Meyerhenke, and D.A. Bader, "**Scalable Multi-threaded Community Detection in Social Networks**," 6th Workshop on Multithreaded Architectures and Applications (MTAAP), Shanghai, China, May 25, 2012.
- H. Meyerhenke, E.J. Riedy, and D.A. Bader, "**Parallel Community Detection in Streaming Graphs**," Minisymposium on Parallel Analysis of Massive Social Networks, *15th SIAM Conference on Parallel Processing for Scientific Computing* (PP12), Savannah, GA, February 15-17, 2012.
- D. Ediger, E.J. Riedy, H. Meyerhenke, and D.A. Bader, "**Analyzing Massive Networks with GraphCT**," Poster Session, *15th SIAM Conference on Parallel Processing for Scientific Computing* (PP12), Savannah, GA, February 15-17, 2012.
- R.C. McColl, D. Ediger, and D.A. Bader, "**Many-Core Memory Hierarchies and Parallel Graph Analysis**," Poster Session, *15th SIAM Conference on Parallel Processing for Scientific Computing* (PP12), Savannah, GA, February 15-17, 2012.
- E.J. Riedy, D. Ediger, H. Meyerhenke, and D.A. Bader, "**STING: Software for Analysis of Spatio-Temporal Interaction Networks and Graphs,**" Poster Session, *15th SIAM Conference on Parallel Processing for Scientific Computing* (PP12), Savannah, GA, February 15-17, 2012.
- Y. Chai, Z. Du, D.A. Bader, and X. Qin, "**Efficient Data Migration to Conserve Energy in Streaming Media Storage Systems,**" *IEEE Transactions on Parallel & Distributed Systems*, 2012.
- M. S. Swenson, J. Anderson, A. Ash, P. Gaurav, Z. Sükösd, D.A. Bader, S.C. Harvey and C.E Heitsch, "**GTfold: Enabling parallel RNA secondary structure prediction on multi-core desktops**," *BMC Research Notes*, 5:341, 2012.
- D. Ediger, K. Jiang, E.J. Riedy, and D.A. Bader, "**GraphCT: Multithreaded Algorithms for Massive Graph Analysis**," *IEEE Transactions on Parallel & Distributed Systems*, 2012.
- D.A. Bader and K. Madduri, "**Computational Challenges in Emerging Combinatorial Scientific Computing Applications**," in O. Schenk, editor, *Combinatorial Scientific Computing*, Chapman & Hall / CRC Press, Chapter 17, pages 471-494, 2012.
- O. Green, R. McColl, and D.A. Bader, "**GPU Merge Path -- A GPU Merging Algorithm**," *26th ACM International Conference on Supercomputing* (ICS), San Servolo Island, Venice, Italy, June 25-29, 2012.
- O. Green, R. McColl, and D.A. Bader, "**A Fast Algorithm for Streaming Betweenness Centrality**," *4th ASE/IEEE International Conference on Social Computing* (SocialCom), Amsterdam, The Netherlands, September 3-5, 2012.
- D. Ediger, R. McColl, J. Riedy, and D.A. Bader, "**STINGER: High Performance Data Structure for Streaming Graphs**," *The IEEE High Performance Extreme Computing Conference* (HPEC), Waltham, MA, September 20-22, 2012. **Best Paper Award.**
- J. Marandola, S. Louise, L. Cudennec, J.-T. Acquaviva and D.A. Bader, "**Enhancing Cache Coherent Architecture with Access Patterns for Embedded Manycore Systems**," *14th IEEE International Symposium on System-on-Chip* (SoC), Tampere, Finland, October 11-12, 2012.
- L.M. Munguía, E. Ayguade, and D.A. Bader, "**Task-based Parallel Breadth-First Search in Heterogeneous Environments**," *The 19th Annual IEEE International Conference on High Performance Computing* (HiPC), Pune, India, December 18-21, 2012.



Bader, Related Recent Publications (2013)

- X. Liu, P. Pande, H. Meyerhenke, and D.A. Bader, "**PASQUAL: Parallel Techniques for Next Generation Genome Sequence Assembly,**" *IEEE Transactions on Parallel & Distributed Systems*, 24(5):977-986, 2013.
- David A. Bader, Henning Meyerhenke, Peter Sanders, and Dorothea Wagner (eds.), ***Graph Partitioning and Graph Clustering***, American Mathematical Society, 2013.
- E. Jason Riedy, Henning Meyerhenke, David Ediger and David A. Bader, "**Parallel Community Detection for Massive Graphs**," in David A. Bader, Henning Meyerhenke, Peter Sanders, and Dorothea Wagner (eds.), *Graph Partitioning and Graph Clustering*, American Mathematical Society, Chapter 14, pages 207-222, 2013.
- S. Kang, D.A. Bader, and R. Vuduc, "**Energy-Efficient Scheduling for Best-Effort Interactive Services to Achieve High Response Quality**," *27th IEEE International Parallel and Distributed Processing Symposium* (IPDPS), Boston, MA, May 20-24, 2013.
- J. Riedy and D.A. Bader, "**Multithreaded Community Monitoring for Massive Streaming Graph Data**," *7th Workshop on Multithreaded Architectures and Applications* (MTAAP), Boston, MA, May 24, 2013.
- D. Ediger and D.A. Bader, "**Investigating Graph Algorithms in the BSP Model on the Cray XMT**," *7th Workshop on Multithreaded Architectures and Applications* (MTAAP), Boston, MA, May 24, 2013.
- O. Green and D.A. Bader, "**Faster Betweenness Centrality Based on Data Structure Experimentation**," *International Conference on Computational Science* (ICCS), Barcelona, Spain, June 5-7, 2013.
- Z. Yin, J. Tang, S. Schaeffer, and D.A. Bader, "**Streaming Breakpoint Graph Analytics for Accelerating and Parallelizing the Computation of DCJ Median of Three Genomes**," *International Conference on Computational Science* (ICCS), Barcelona, Spain, June 5-7, 2013.
- T. Senator, D.A. Bader, et al., "**Detecting Insider Threats in a Real Corporate Database of Computer Usage Activities**," *19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD), Chicago, IL, August 11-14, 2013.
- J. Fairbanks, D. Ediger, R. McColl, D.A. Bader and E. Gilbert, "**A Statistical Framework for Streaming Graph Analysis**," *IEEE/ACM International Conference on Advances in Social Networks Analysis and Modeling* (ASONAM), Niagara Falls, Canada, August 25-28, 2013.
- A. Zakrzewska and D.A. Bader, "**Measuring the Sensitivity of Graph Metrics to Missing Data**," *10th International Conference on Parallel Processing and Applied Mathematics* (PPAM), Warsaw, Poland, September 8-11, 2013.
- O. Green and D.A. Bader, "**A Fast Algorithm for Streaming Betweenness Centrality**," *5th ASE/IEEE International Conference on Social Computing* (SocialCom), Washington, DC, September 8-14, 2013.
- R. McColl, O. Green, and D.A. Bader, "**A New Parallel Algorithm for Connected Components in Dynamic Graphs**," *The 20th Annual IEEE International Conference on High Performance Computing* (HiPC), Bangalore, India, December 18-21, 2013.



Bader, Related Recent Publications (2014-2015)

- R. McColl, D. Ediger, J. Poovey, D. Campbell, and D.A. Bader, "A Performance Evaluation of Open Source Graph Databases," *The 1st Workshop on Parallel Programming for Analytics Applications* (PPAA 2014) held in conjunction with the *19th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP 2014), Orlando, Florida, February 16, 2014.
- O. Green, L.M. Munguia, and D.A. Bader, "Load Balanced Clustering Coefficients," *The 1st Workshop on Parallel Programming for Analytics Applications* (PPAA 2014) held in conjunction with the *19th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming* (PPoPP 2014), Orlando, Florida, February 16, 2014.
- A. McLaughlin and D.A. Bader, "Revisiting Edge and Node Parallelism for Dynamic GPU Graph Analytics," *8th Workshop on Multithreaded Architectures and Applications* (MTAAP), held in conjuntion with *The IEEE International Parallel and Distributed Processing Symposium (IPDPS 2014)*, Phoenix, AZ, May 23, 2014.
- Z. Yin, J. Tang, S. Schaeffer, D.A. Bader, "A Lin-Kernighan Heuristic for the DCJ Median Problem of Genomes with Unequal Contents," *20th International Computing and Combinatorics Conference* (COCOON), Atlanta, GA, August 4-6, 2014.
- Y. You, D.A. Bader and M.M. Dehnavi, "Designing an Adaptive Cross-Architecture Combination for Graph Traversal," *The 43rd International Conference on Parallel Processing* (ICPP 2014), Minneapolis, MN, September 9-12, 2014.
- A. McLaughlin, J. Riedy, and D.A. Bader, "Optimizing Energy Consumption and Parallel Performance for Betweenness Centrality using GPUs," *The 18th Annual IEEE High Performance Extreme Computing Conference* (HPEC), Waltham, MA, September 9-11, 2014.
- A. McLaughlin and D.A. Bader, "Scalable and High Performance Betweenness Centrality on the GPU," *The 26th IEEE and ACM Supercomputing Conference* (SC14), New Orleans, LA, November 16-21, 2014. **Best Student Paper Finalist.**
- D. Dauwe, E. Jonardi, R. Friese, S. Pasricha, A.A. Maciejewski, D.A. Bader, and H.J. Siegel, "**A Methodology for Co-Location Aware Application Performance Modeling in Multicore Computing**," 17th Workshop on Advances on Parallel and Distributed Processing Symposium (APDCM), Hyderabad, India, May 25, 2015.
- A. Zakrzewska and D.A. Bader, "**Fast Incremental Community Detection on Dynamic Graphs**," 11th International Conference on Parallel Processing and Applied Mathematics (PPAM), Krakow, Poland, September 6-9, 2015.
- A. McLaughlin, J. Riedy, and D.A. Bader, "**An Energy-Efficient Abstraction for Simultaneous Breadth-First Searches**," The 19th Annual IEEE High Performance Extreme Computing Conference (HPEC), Waltham, MA, September 15-17, 2015.
- A. McLaughlin, D. Merrill, M. Garland and D.A. Bader, "**Parallel Methods for Verifying the Consistency of Weakly-Ordered Architectures**," The 24th International Conference on Parallel Architectures and Compilation Techniques (PACT), San Francisco, CA, October 18-21, 2015.

Acknowledgment of Support