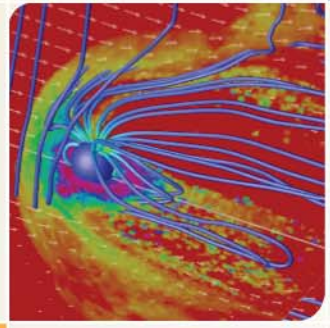
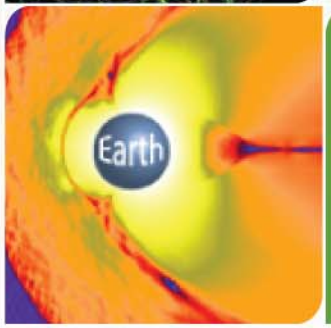
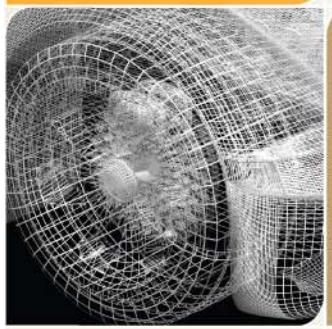





Opportunities and Challenges in Massive Data-Intensive Computing

David A. Bader

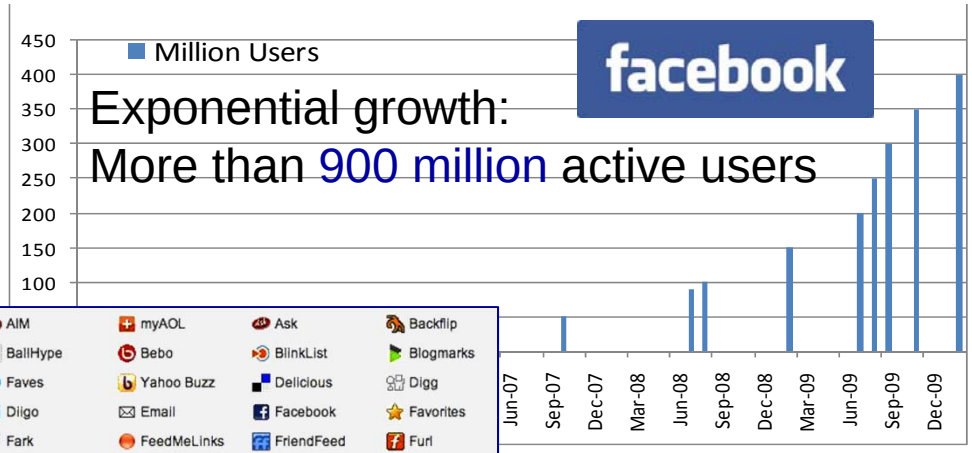
Georgia Tech College of Computing
Computational Science and Engineering

Exascale Streaming Data Analytics: Real-world challenges



All involve analyzing massive streaming complex networks:

- Health care → disease spread, detection and prevention of epidemics/pandemics (e.g. SARS, Avian flu, H1N1 "swine" flu)
- Massive social networks → understanding communities, intentions, population dynamics, pandemic spread, transportation and evacuation
- Intelligence → business analytics, anomaly detection, security, knowledge discovery from massive data sets
- Systems Biology → understanding complex life systems, drug design, microbial research, unravel the mysteries of the HIV virus; understand life, disease,
- Electric Power Grid → communication, transportation, energy, water, food supply
- Modeling and Simulation → Perform full-scale economic-social-political simulations



Ex: discovered minimal changes in O(billions)-size complex network that could hide or reveal top influencers in the community

Sample queries:

Allegiance switching: identify entities that switch communities.

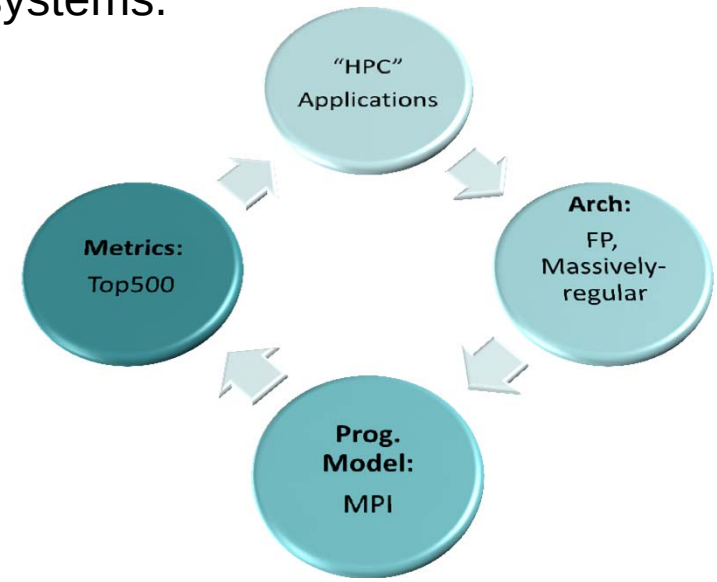
Community structure: identify the genesis and dissipation of communities

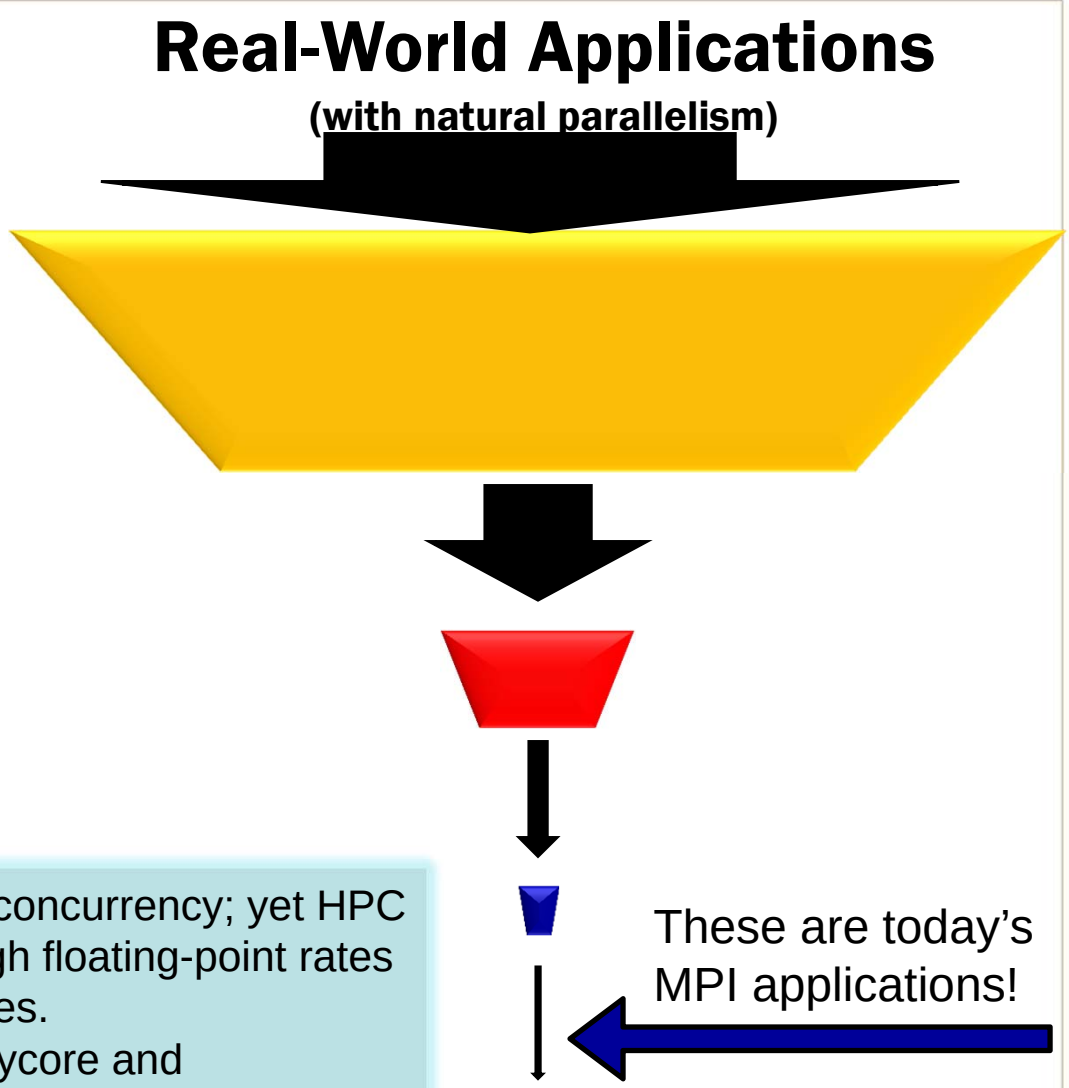
Phase change: identify significant change in the network structure

REQUIRES PREDICTING / INFLUENCE CHANGE IN REAL-TIME AT SCALE

Flywheel has driven HPC into a corner

For **decades**, HPC has been on a vicious cycle of enabling applications that run well on HPC systems.





➔Data-intensive computing has natural concurrency; yet HPC architectures are designed to achieve high floating-point rates by exploiting spatial and temporal localities.

- For the first time in **decades**, manycore and multithreaded can let us rethink architecture.



Opportunity 1: High performance computing for massive graphs

- Traditional HPC has focused primarily on solving large problems from chemistry, physics, and mechanics, using dense linear algebra.
- HPC faces new challenges to deal with:
  - time-varying interactions among entities, and
  - massive-scale graph abstractions where the vertices represent the nouns or entities and the edges represent their observed interactions.
- Few parallel computers run well on these problems because
  - they often lack locality required to get high performance from distributed-memory cache-based supercomputers.
- Case study: Massively threaded architectures are shown to run several orders of magnitude faster than the fastest supercomputers on these types of problems!

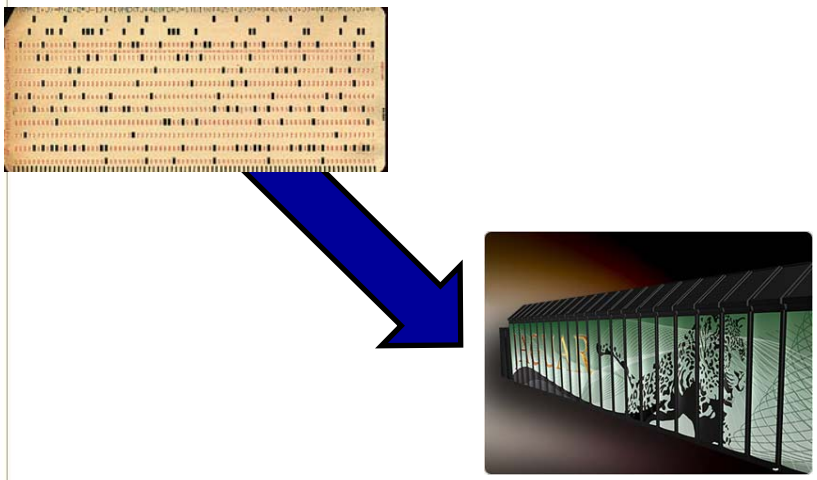
➔ A focused research agenda is needed to design algorithms that scale on these new platforms.

Opportunity 2: Streaming analytics

- While our high performance computers have delivered a sustained petaflop, they have done so using the same antiquated batch processing style where a program and a static data set are scheduled to compute in the next available slot.
- Today, data is overwhelming in volume *and rate*, and we struggle to keep up with these streams.

➔ Fundamental computer science research is needed in:

- ➔ the design of streaming architectures, and
- ➔ data structures and algorithms that can compute important analytics while sitting in the middle of these torrential flows.



VS.

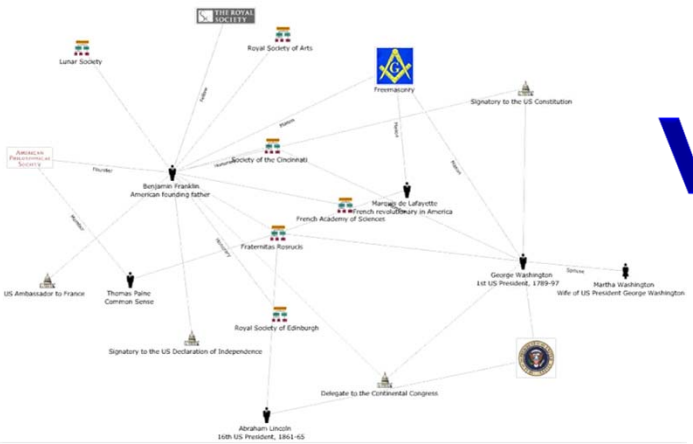
Opportunity 3: Information Visualization techniques for massive graphs

- Information Visualization today
  - addresses traditional scientific computing (fluid flow, molecular dynamics), or
  - when handling discrete data, scale to only hundreds of vertices at best.

➔ However, there is a strong need for visualization in the data sciences so that analytics can gain understanding from data sets with **from millions to billions** of interacting non-planar discrete entities.

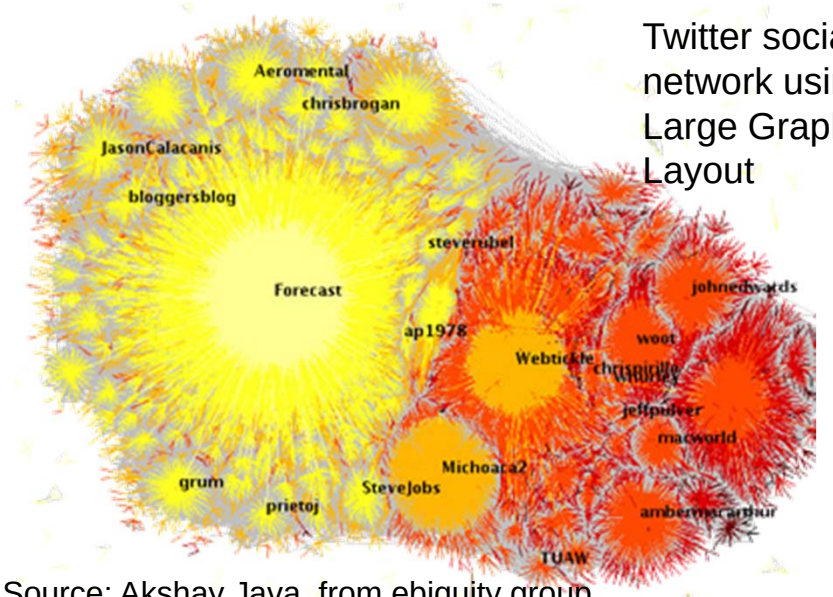
  - Applications include: data mining, intelligence, situational awareness

NNDB
tracking the entire world



NNDB Mapper of George Washington

VS.



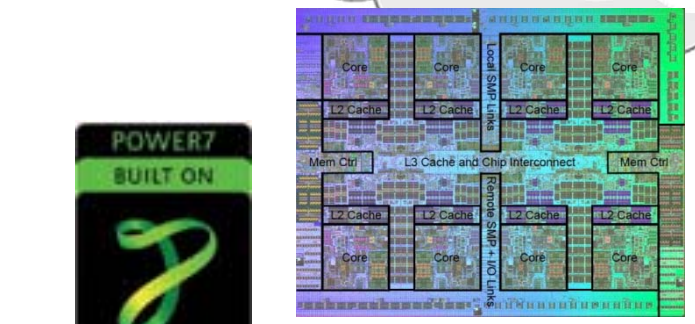
Twitter social network using Large Graph Layout

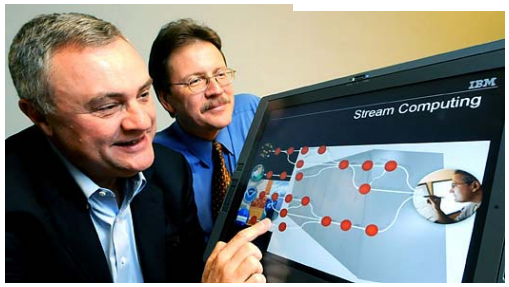
Source: Akshay Java, from ebiquity group

Opportunity 4: Heterogeneous Systems: Methodologies for combining the use of the Cloud and Manycore for high-performance computing.

- Today, there is a dichotomy between using clouds (e.g. Hadoop, map-reduce) for massive data storage, filtering, summarization, and massively parallel/multithreaded systems for data-intensive computation.

➔ We must develop methodologies for employing these complementary systems for solving grand challenges in data analysis.



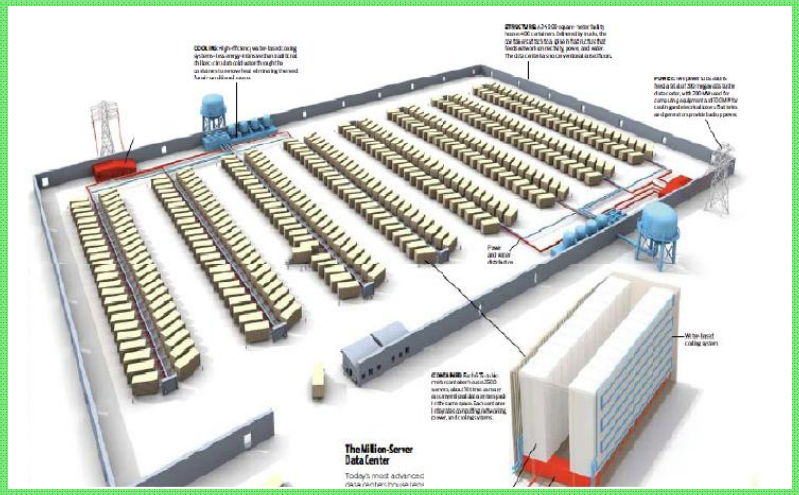


Steve Mills, SVP of IBM Software (left), and Dr. John Kelly, SVP of IBM Research, view Stream Computing technology

Opportunity 5: Energy-efficient high-performance computing

- The main constraint for our ability to compute has changed
  - from availability of compute resources
  - to the ability to power and cool our systems within budget.

➔ Holistic research is needed that can permeate from the architecture and systems up to the applications AND DATA CENTERS, whereby energy use is a first-class object that can be optimized at all levels.



Microsoft's Chicago Million Server DataCenter

Acknowledgment of Support