Common Challenges between Exascales and Cloud Computing

Ioan Raicu
Computer Science Department, Illinois Institute of Technology
Math and Computer Science Division, Argonne National Laboratory

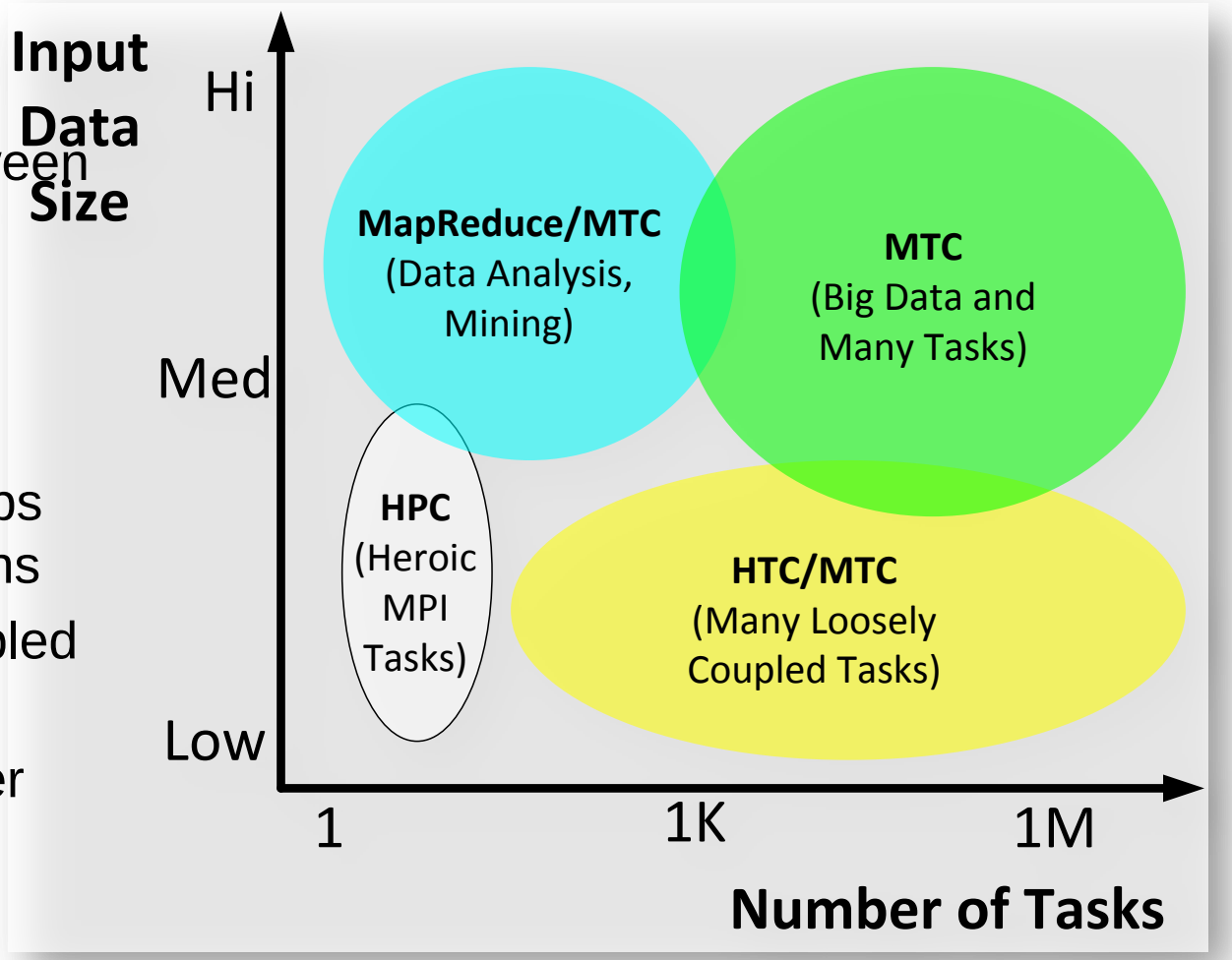
DataCloud2011 @ IPDPS 2011
May 13th, 2011

Who am I?

- **Current position:**
  - Assistant Professor at Illinois Institute of Technology (CS)
  - Guest Research Faculty, Argonne National Laboratory (MCS)
- **Education:** PhD, University of Chicago, March 2009
- **Funding/Awards:**
  - NSF CAREER, 2011 – 2015 ($450K)
  - NSF/CRA CIFellows, 2009 – 2010 ($140K)
  - NASA GSRP, 2006 – 2009 ($84K)
- **Over 70+ Collaborators:**
  - Ian Foster (UC/ANL), Rick Stevens (UC/ANL), Rob Ross (ANL), Marc Snir (UIUC), Arthur Barney Maccabe (ORNL), Alex Szalay (JHU), Pete Beckman (ANL), Kamil Iskra (ANL), Mike Wilde (UC/ANL), Douglas Thain (ND), Yong Zhao (UEST), Matei Ripeanu (UBC), Alok Choudhary (NU), Tevfik Kosar (SUNY), Yogesh Simhan (USC), Ewa Deelman (USC), and many more…
- **More info:** http://www.cs.iit.edu/~iraicu/index.html

Best Known For

- **MTC: Many-Task Computing**
  - Bridge the gap between HPC and HTC
  - Applied in clusters, grids, and supercomputers
  - Loosely coupled apps with HPC orientations
  - Many activities coupled by file system ops
  - Many resources over short time periods



[MTAGS08] "Many-Task Computing for Grids and Supercomputers"

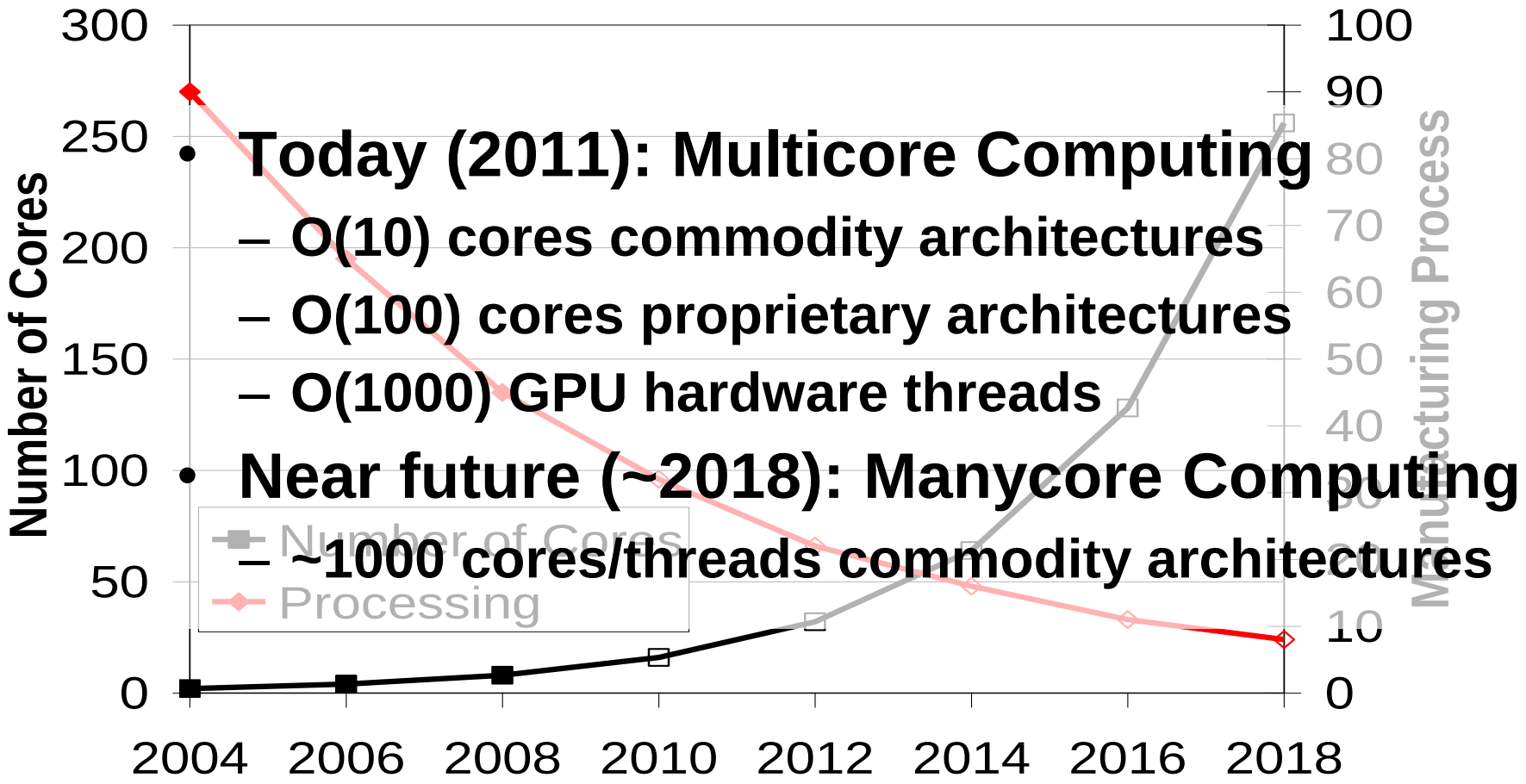
DataSys: Data-Intensive Distributed Systems Laboratory

- **Research Focus**
  - Emphasize designing, implementing, and evaluating systems, protocols, and middleware with the goal of supporting **data-intensive applications on extreme scale distributed systems**, from many-core systems, clusters, grids, clouds, and supercomputers
- **People**
  - Dr. Ioan Raicu (Director)
  - Tonglin Li (PhD Student)
  - Xi Duan (MS Student)
  - Raman Verma (Research Staff)
  - 3 PhD and 1 UG students joining over the next several months
- **More information**
  - http://datasys.cs.iit.edu/

Overview

- **This talk covers material from my NSF CAREER award:**
  - Ioan Raicu, Arthur Barney Maccabe, Marc Snir, Rob Ross, Mike Wilde, Kamil Iskra, Jacob Furst, Mary Cummane. “Avoiding Achilles’ Heel in Exascale Computing with Distributed File Systems”, NSF OCI CAREER Award #1054974
- **And from a recent invited position paper (to appear):**
  - Ioan Raicu, Pete Beckman, Ian Foster. “Making a Case for Distributed File Systems at Exascale”, ACM Workshop on Large-scale System and Application Performance (LSAP), 2011

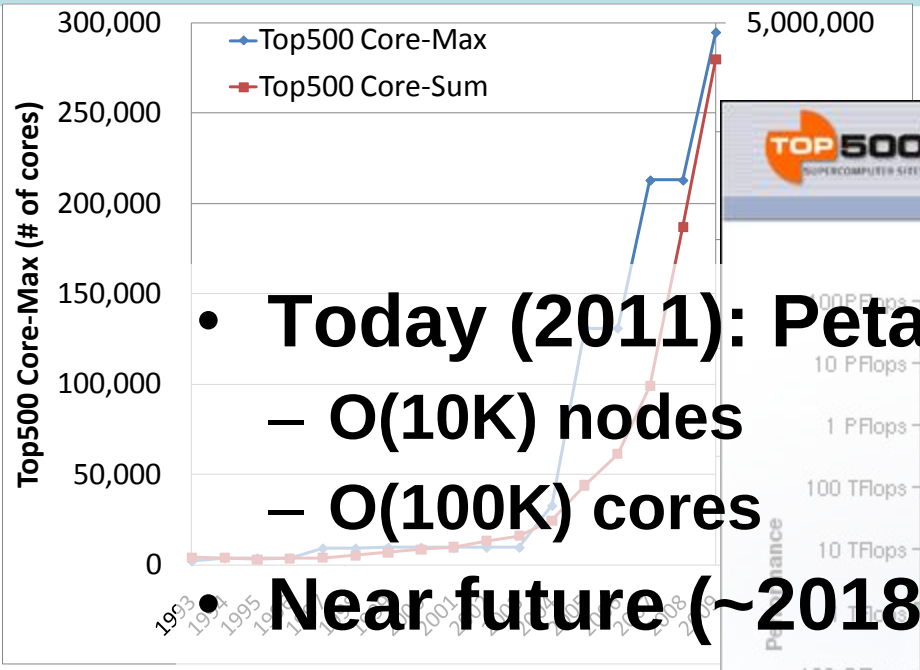
Manycore Computing

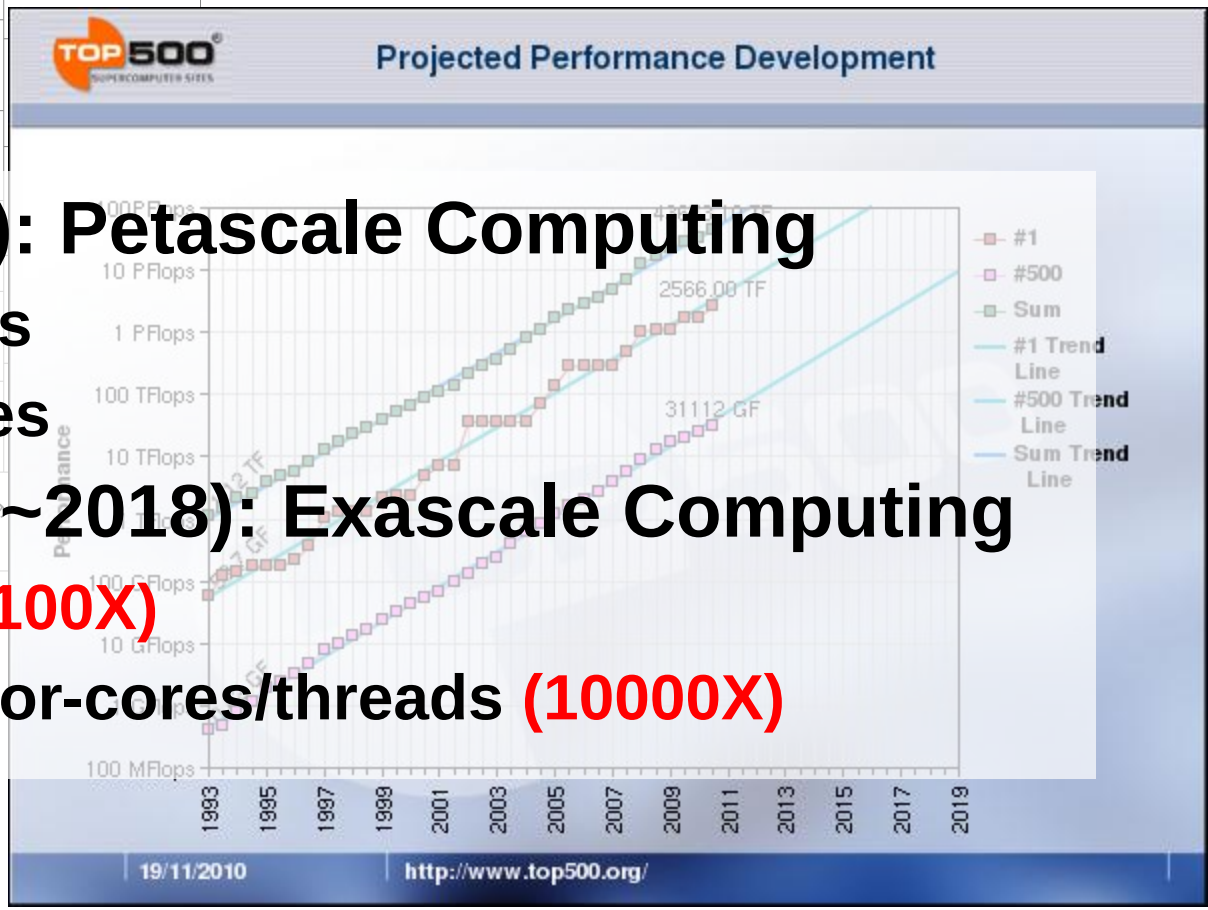


- **Today (2011): Multicore Computing**
  - **O(10) cores commodity architectures**
  - **O(100) cores proprietary architectures**
  - **O(1000) GPU hardware threads**
- **Near future (~2018): Manycore Computing**
  - **~1000 cores/threads commodity architectures**

Pat Helland, Microsoft, The Irresistible Forces Meet the Movable Objects, November 9th, 2007

Exascale Computing





- **Today (2011): Petascale Computing**
  - **O(10K) nodes**
  - **O(100K) cores**
- **Near future (~2018): Exascale Computing**
  - **~1M nodes (100X)**
  - **~1B processor-cores/threads (10000X)**

Top500 Projected Development,
http://www.top500.org/lists/2010/11/performance_development

Cloud Computing

- Relatively new paradigm… 3~4 years old
- Amazon in 2009
  - 40K servers split over 6 zones
    - 320K-cores, 320K disks
    - $100M costs + $12M/year in energy costs
    - Revenues about $250M/year
    - http://www.siliconvalleywatcher.com/mt/archives/2009/10/meausuring_amaz.php
- Amazon in 2018
  - Will likely look similar to exascale computing
    - 100K~1M nodes, ~1B-cores, ~1M disks
    - $100M~$200M costs + $10M~$20M/year in energy
    - Revenues 100X~1000X of what they are today

Common Challenges

- Power efficiency
  - Will limit the number of cores on a chip (Manycore)
  - Will limit the number of nodes in cluster (Exascale and Cloud)
  - Will dictate a significant part of the cost of ownership
- Programming models/languages
  - Automatic parallelization
  - Threads, MPI, workflow systems, etc
  - Functional, imperative
  - Languages vs. Middleware

Programming Models Work

- **Falkon**
  - **Fast and Lightweight T Execution Framework**
  - http://dev.globus.or ubator/Falkon
- **Swift**
  - **Parallel Programming System**
  - http://www.ci.uchica wift/index.php

Field	Description	Characteristics	Status
Astronomy	Creation of montages from many digital images	Many 1-core tasks, much communication, complex dependencies	Experimental
Astronomy	Stacking of cutouts from digital sky surveys	Many 1-core tasks, much communication	Experimental
Biochemistry*	Analysis of mass-spectrometer data for post-translational protein modifications	10,000-100 million jobs for proteomic searches using custom serial codes	In development
Biochemistry*	Protein structure prediction using iterative fixing algorithm; exploring other biomolecular interactions	Hundreds to thousands of 1- to 1,000-core simulations and data analysis	Operational
Biochemistry*	Identification of drug targets via computational docking/screening	Up to 1 million 1-core docking operations	Operational
Bioinformatics*	Metagenome modeling	Thousands of 1-core integer programming problems	In development
Business economics	Mining of large text corpora to study media bias	Analysis and comparison of over 70 million text files of news articles	In development
Climate science	Ensemble climate model runs and analysis of output data	Tens to hundreds of 100- to 1,000-core simulations	Experimental
Economics*	Generation of response surfaces for various eco-nomic models	1,000 to 1 million 1-core runs (10,000 typical), then data analysis	Operational
Neuroscience*	Analysis of functional MRI datasets	Comparison of images; connectivity analysis with structural equation modeling, 100,000+ tasks	Operational
Radiology	Training of computer-aided diagnosis algorithms	Comparison of images; many tasks, much communication	In development
Radiology	Image processing and brain mapping for neuro-surgical planning research	Execution of MPI application in parallel	In development

Note: Asterisks indicate applications being run on Argonne National Laboratory's Blue Gene/P (Intrepid) and/or the TeraGrid Sun Constellation at the University of Texas at Austin (Ranger).

Common Challenges

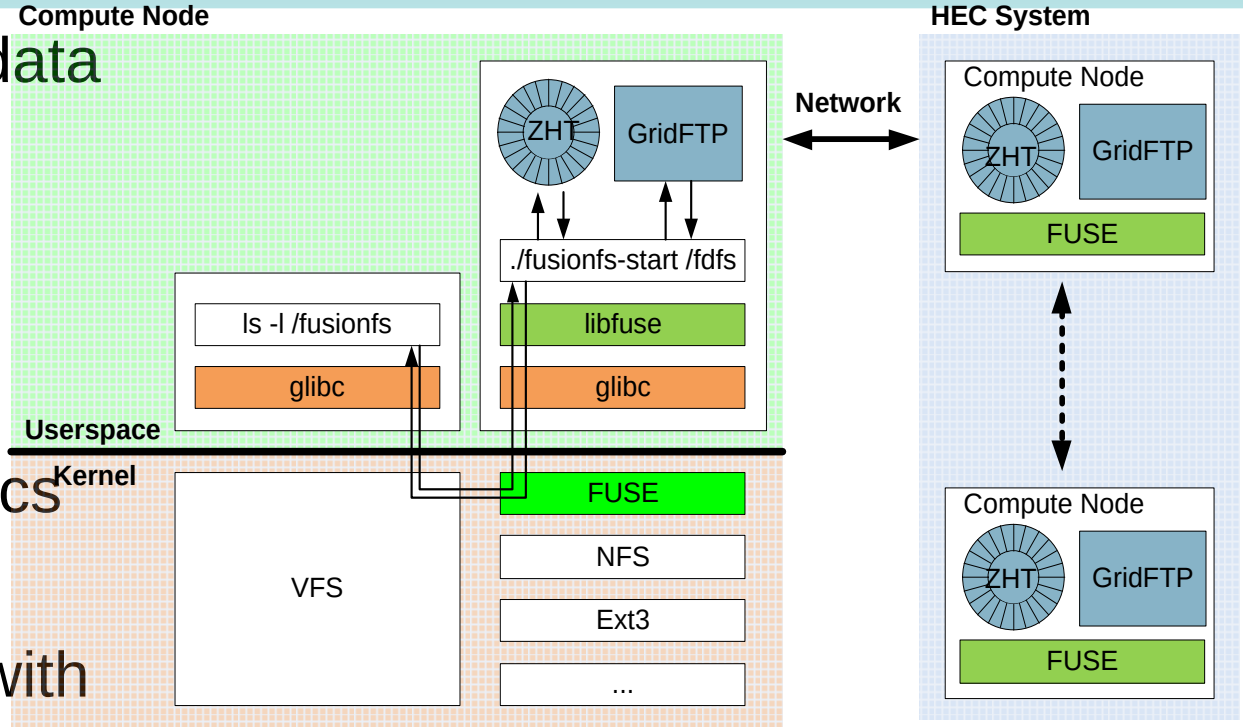
- Bottlenecks in scarce resources
  - Storage (Exascale and Clouds)
  - Memory (Manycore)
- Reliability
  - How to keep systems operational in face of failures
  - Checkpointing (Exascale)
  - Node-level replication enabled by virtualization (Exascale and Clouds)
  - Hardware redundancy and hardware error correction (Manycore)

FusionFS: Fusion Distributed File System

- ***Building on my own research (e.g. data-diffusion), parallel file systems (PVFS), and distributed file systems (e.g. GFS) ➔ FusionFS, a distributed file system for HEC***
  - It should complement parallel file systems, not replace them
- **Critical issues:**
  - Must mimic parallel file systems interfaces and features in order to get wide adoption (e.g. POSIX)
  - Must handle some workloads currently run on parallel file systems significantly better

FusionFS Details

- Distributed Metadata Management
- Distributed Data Management
- Data Indexing
- Relaxed Semantics
- Data Locality
- Overlapping I/O with Computations
- POSIX



Main Message

- ***Preserving locality is critical!***
- *Segregating storage from compute resources is **BAD***
- *Parallel file systems + distributed file systems + distributed hash tables + nonvolatile memory*
  ➔ new storage architecture for extreme-scale HEC
- *Co-locating storage and compute is **GOOD***
  - *Leverage the abundance of processing power, bisection bandwidth, and local I/O*

More Information

- **More information:**
  - http://www.cs.iit.edu/~iraicu/index.html
  - http://datasys.cs.iit.edu/
- **Relevant upcoming workshops and journals**
  - **DataCloud:** IEEE Int. Workshop on Data-Intensive Computing in the Clouds (at IPDPS), 2011
  - **HPDC/SigMetrics:** HPDC/SIGMETRICS 2011 Student Poster Session, 2011
  - **JGC:** Springer Journal of Grid Computing, Special Issue on Data Intensive Computing in the Clouds, 2011
  - **MTAGS:** ACM Workshop on Many-Task Computing on Grids and Supercomputers (at SC), 2011
  - **ScienceCloud:** ACM Workshop on Scientific Cloud Computing (at HPDC), 2010, 2011
  - **SPJ:** Scientific Programming Journal, Special Issue on Science-driven Cloud Computing, 2011
  - **TPDS:** IEEE Transactions on Parallel and Distributed Systems, Special Issue on Many-Task Computing, 2011