RUTGERS

RDI²
Rutgers Discovery Informatics Institute

Exploring energy-performance-quality tradeoffs for scientific workflows with in-situ data analyses

Georgiana Haldeman, **Ivan Rodero***, Manish Parashar,
Sabela Ramos, Eddy Z. Zhang, Ulrich Kremer

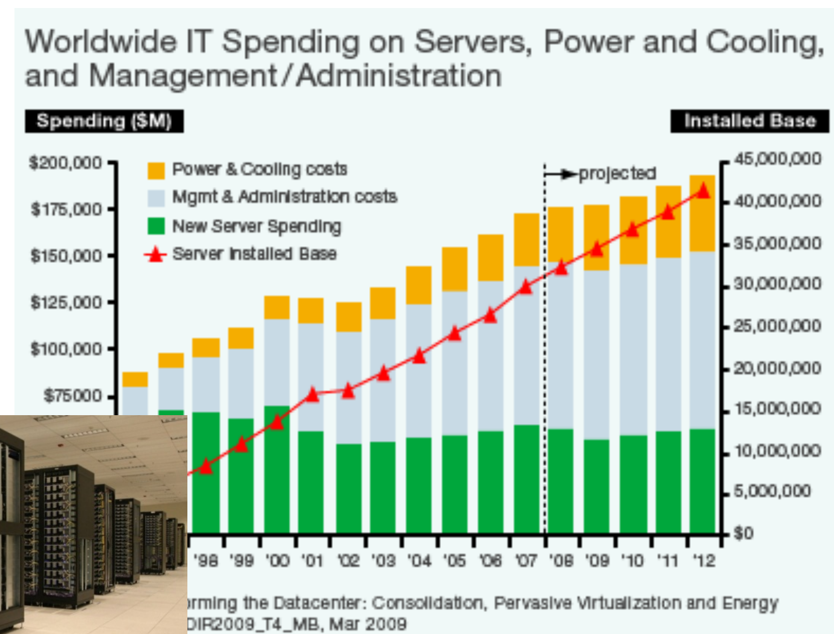
*Rutgers Discovery Informatics Institute (RDI²)
NSF Cloud and Autonomic Computing Center (CAC)
Rutgers, The State University of New Jersey
email: **irodero@rutgers.edu**

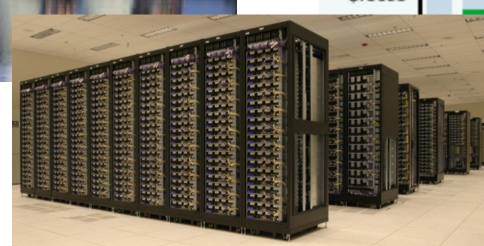
Enery-Aware High Performance Computing (EnA-HPC) – Dresden, Germany **September 1, 2014**

Power/Energy Challenge – Green HPC

- Power has become a critical concern for HPC/supercomputing
  - Impacts operational costs, reliability, correctness
  - End-to-end integrated power/energy management essential
- Increasing scale towards **exascale**
  - Using existing technology would require gigawatt??
    - Nuclear reactor scale??
    - > $2.5B annual power cost

Target < 20MW !!





Modern Science & Society Transformed by Compute & Data

- New Paradigms & Practices
  - End-to-end: Seamless access, aggregation, interactions
  - Data-driven, Data/Compute-intensive; Age of Digital Observation
  - Integrative, multi-scale, online
- Multi-disciplinary collaborations
  - Individuals, groups, teams, communities, networks
  - New global science culture
- Unprecedented opportunities, challenges

Ack. M. Parashar

Clearly, modern instruments/experiments/… are producing Big Data!!

Large Hadron Collider

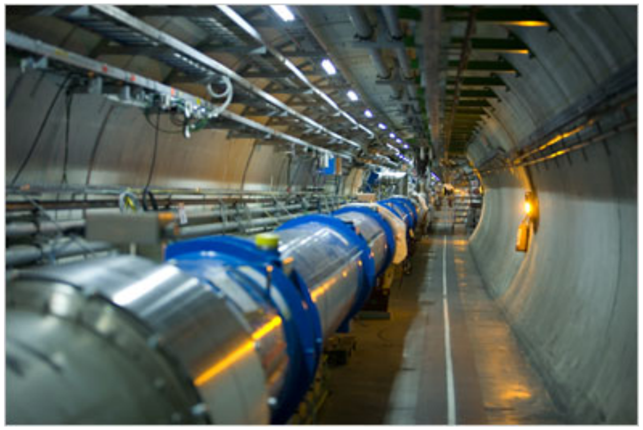


Image credit: Valerio Mezzanotti for The New York Times

Blanco 4m on Cerro Tololo

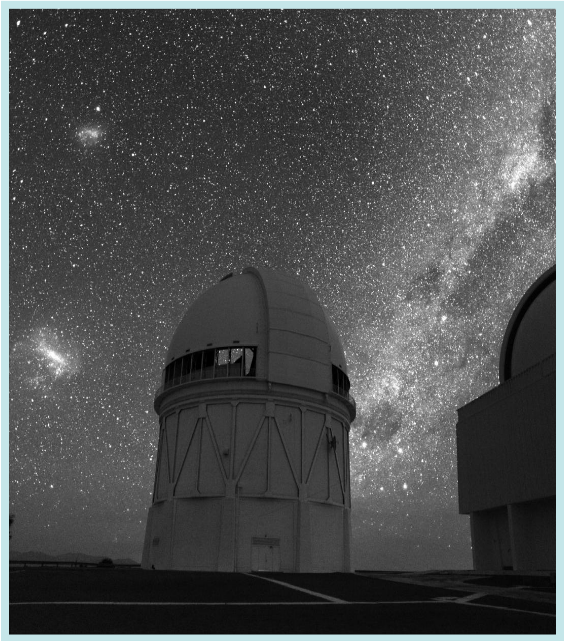
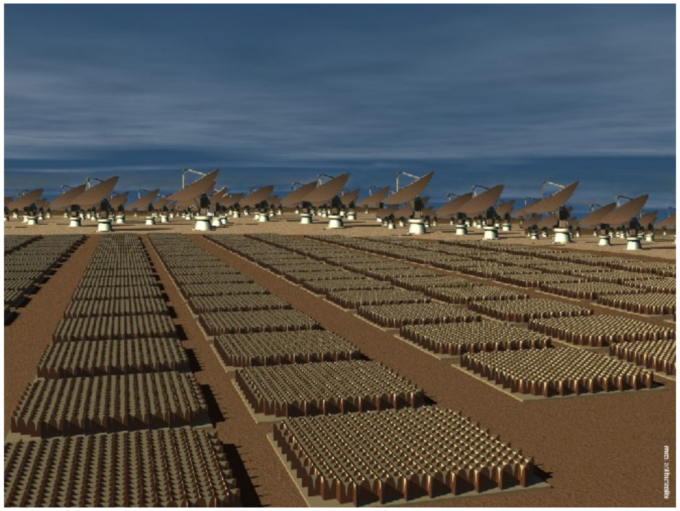


Image credit: Roger Smith/NOAO/AURA/NSF

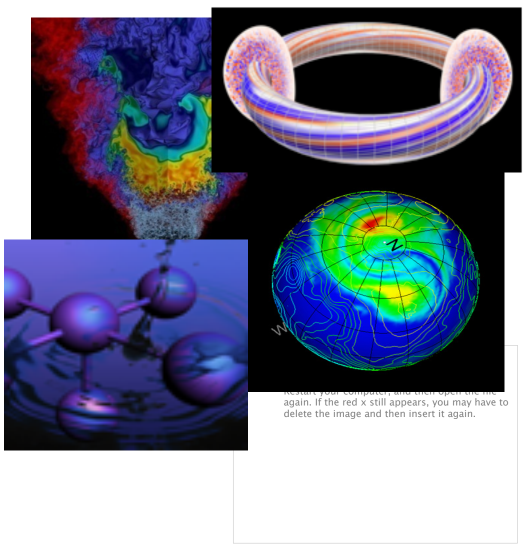
SKA project

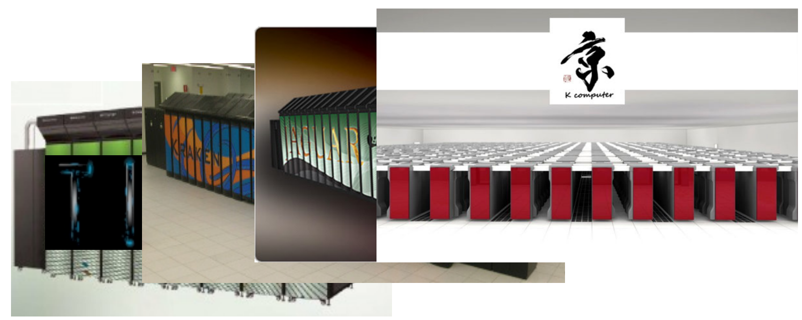


Above is proposed image

Scientific Discovery through Simulations

- Scientific simulations running on high-end computing systems generate huge amounts of data!
  - If a single core produces 2MB/minute on average, one of these machines could generate simulation data between ~**170TB** per hour -> ~**700PB** per day -> ~**1.4EB** per year
- Successful scientific discovery depends on a comprehensive understanding of this enormous simulation data





How we enable the computation scientists to efficiently manage and explore extreme scale data: "find the needles in haystack" ??

RUTGERS

RDI²
Rutgers Discovery Informatics Institute

Challenges Faced by Traditional HPC Data Pipelines

- **Data analysis challenge**
  - Can current data mining, manipulation and visualization algorithms still work effectively on extreme scale machine?
- **I/O challenge**
  - Increasing performance gap: disks are outpaced by computing speed

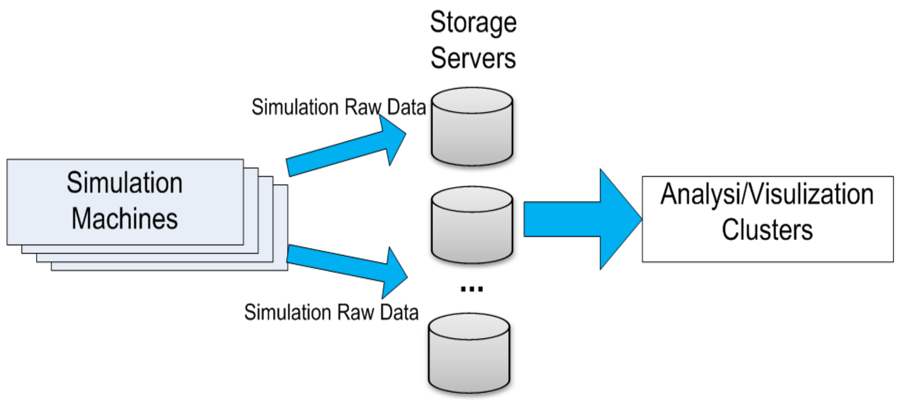


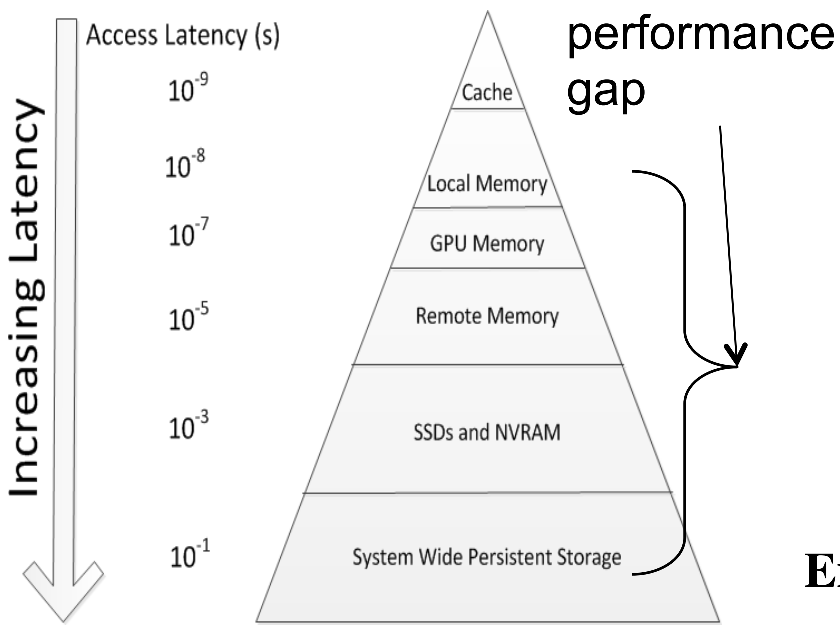
Figure. Traditional data analysis pipeline

- **Data movement challenge**
  - Lots of data movement between simulation and analysis machines, between coupled mutli-physics simulation components -> longer latencies
  - Improving data locality is critical: do work where the data resides!
- **Energy challenge**
  - Future extreme systems are designed to have low-power chips – however, much greater power consumption will be due to memory and data movement!

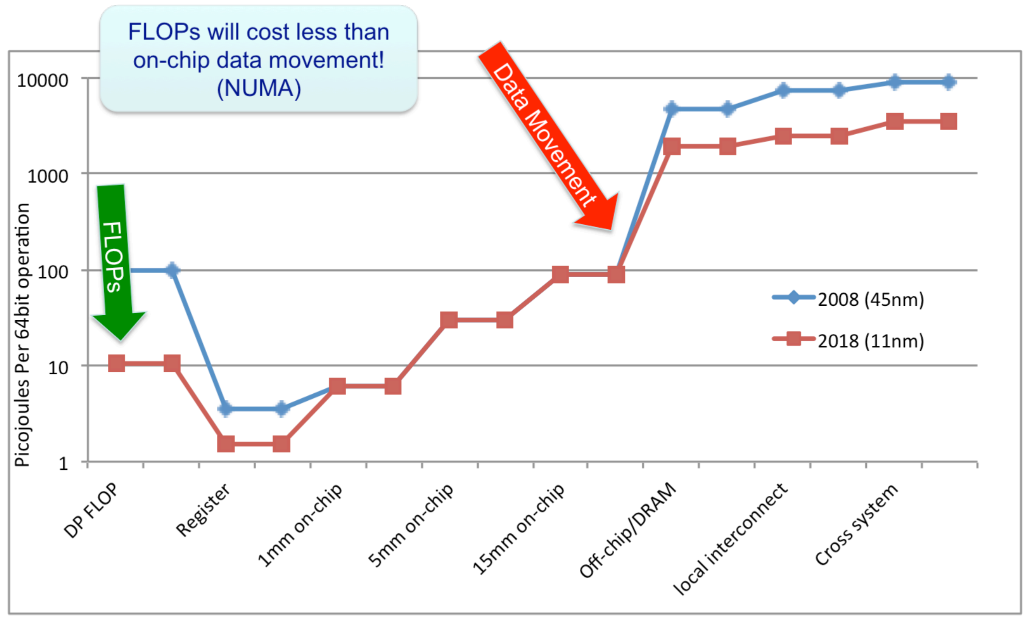
*The **costs of data movement** are increasing and dominating!*

The Cost of Data Movement

- 40-50% energy spent in off-chip memory hierarchy!! [Lefurgy, IEEE Computer'03]
- Moving data between node memory and persistent storage is slow!



- The energy cost of moving data is a significant concern

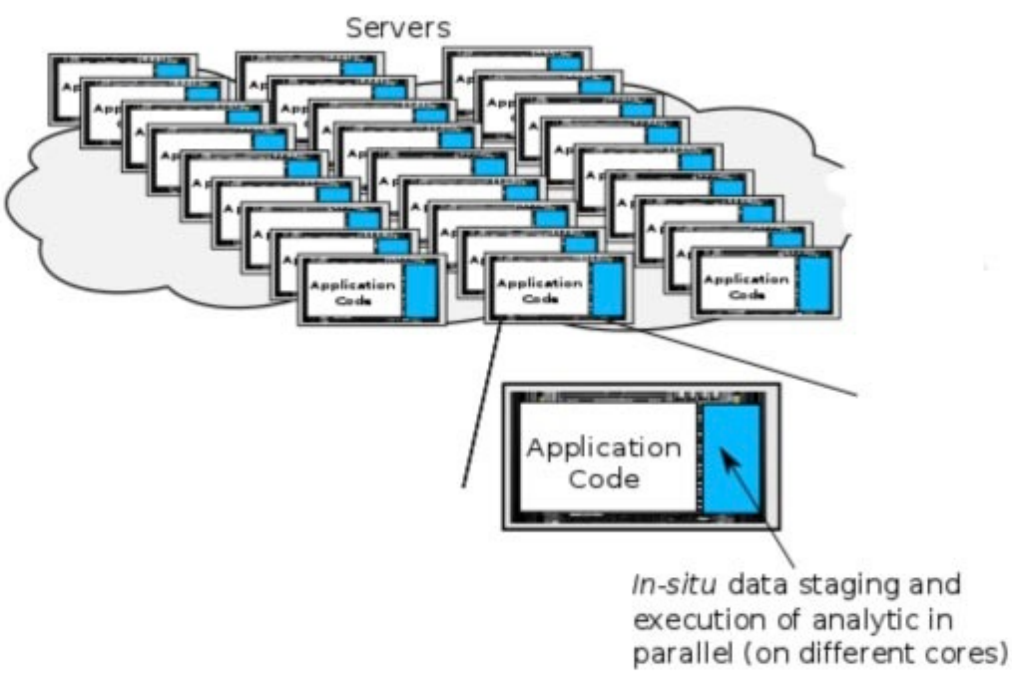


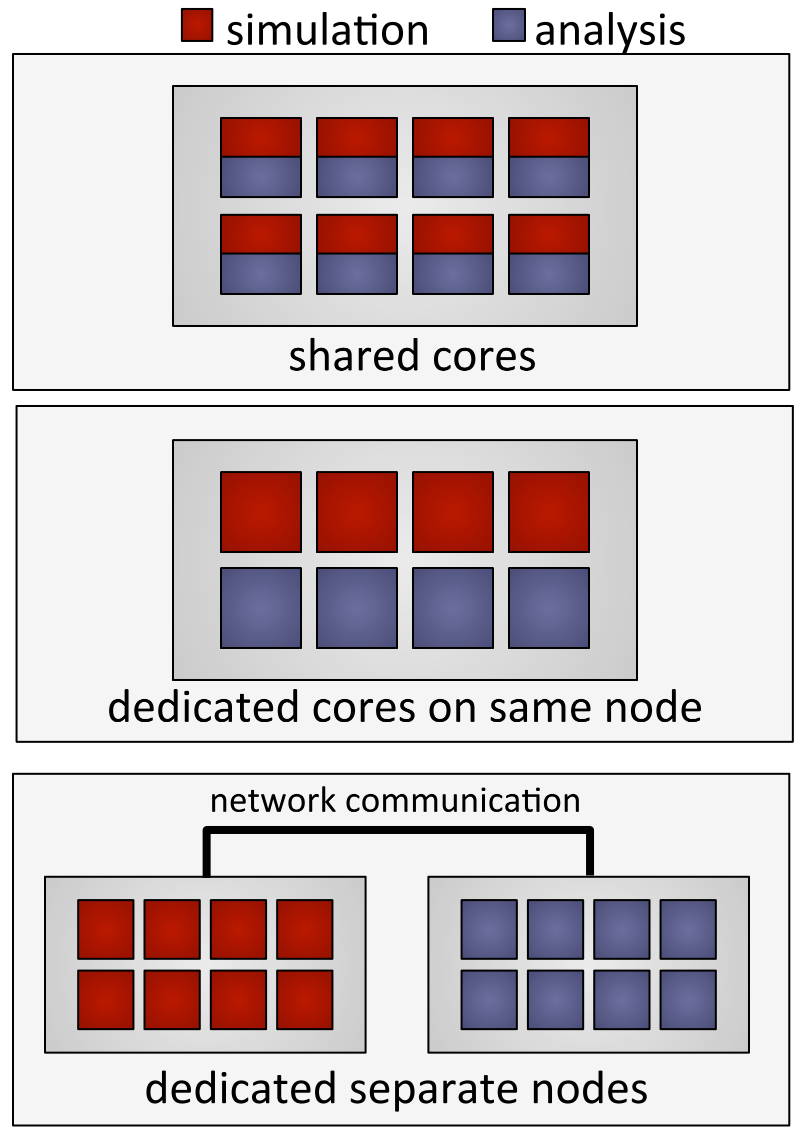
$$\mathrm{Energy\_move\_data} = \frac{\mathrm{bitrate} * \mathrm{length}^2}{\mathrm{cross\_section\_area\_of\_wire}}$$

From K. Yelick, "Software and Algorithms for Exascale: Ten Ways to Waste an Exascale Computer"

Rethinking the Data Management Pipeline – In-Situ Data Analytics

- Location of analysis compute resources
  - Same cores as the simulation
  - Dedicated cores on the same node



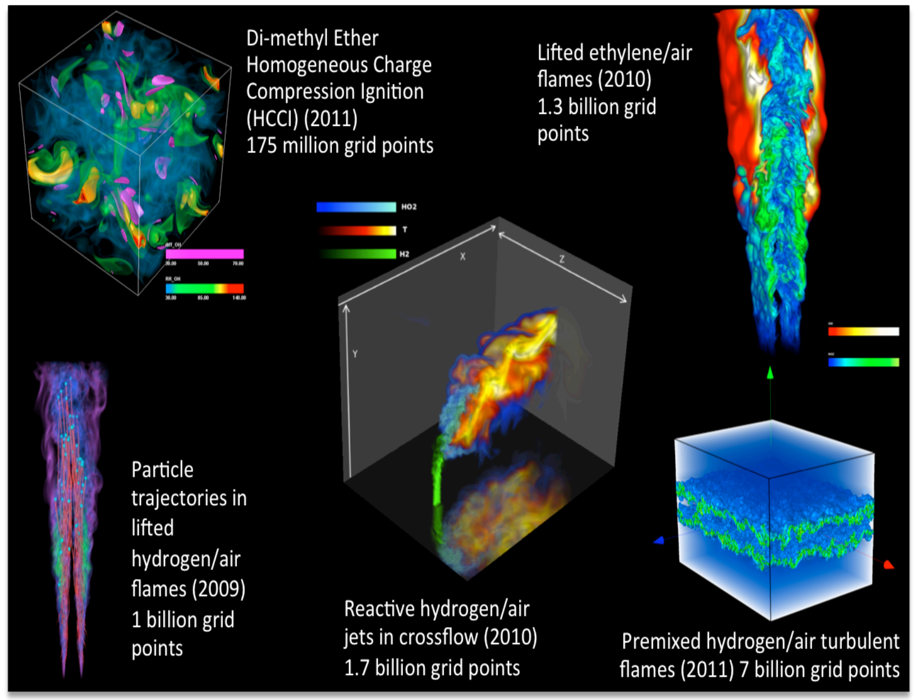


Power Behavior of In-situ Analytics Pipeline

- Combustion simulation workflow with an **in-situ data analytics** pipeline

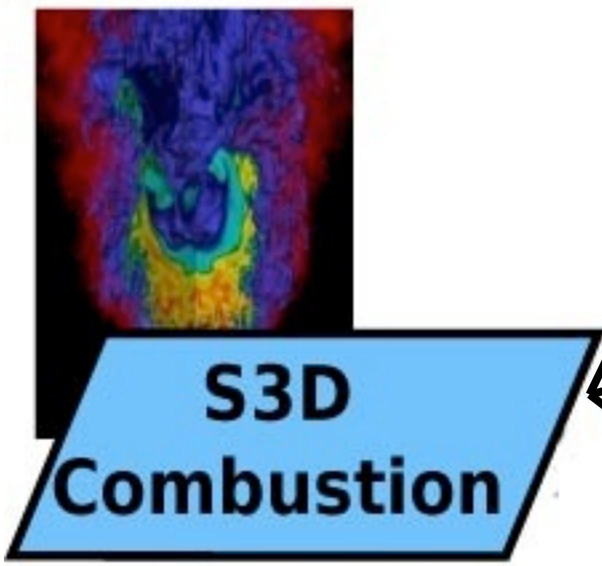
- With research groups from

Sandia National Laboratories
Lawrence Livermore National Laboratory
Los Alamos NATIONAL LABORATORY
OAK RIDGE National Laboratory
Pacific Northwest NATIONAL LABORATORY
THE UNIVERSITY OF UTAH



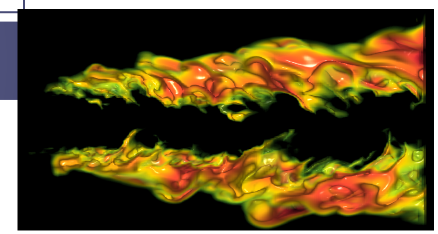
Recent data sets generated by S3D, developed at the Combustion Research Facility, Sandia National Laboratories

In-situ DataAnalysis as Part of S3D

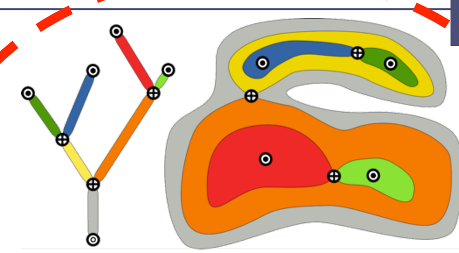


$$
\begin{aligned}
M_{p,\mathscr{S}} &= M_{p,\mathscr{S}_1} + M_{p,\mathscr{S}_2} \\
&+ \sum_{k=1}^{p-2} \binom{k}{p} \left[\left(-\frac{n_2}{n} \right)^k M_{p-k,\mathscr{S}_1} + \left(\frac{n_1}{n} \right)^k M_{p-k,\mathscr{S}_2} \right] \delta_{2,1}^k \\
&+ \left(\frac{n_1 n_2}{n} \delta_{2,1} \right)^p \left[\frac{1}{n_2^{p-1}} - \left(\frac{-1}{n_1} \right)^{p-1} \right]. \qquad \text{(III.1)}
\end{aligned}
$$

Statistics



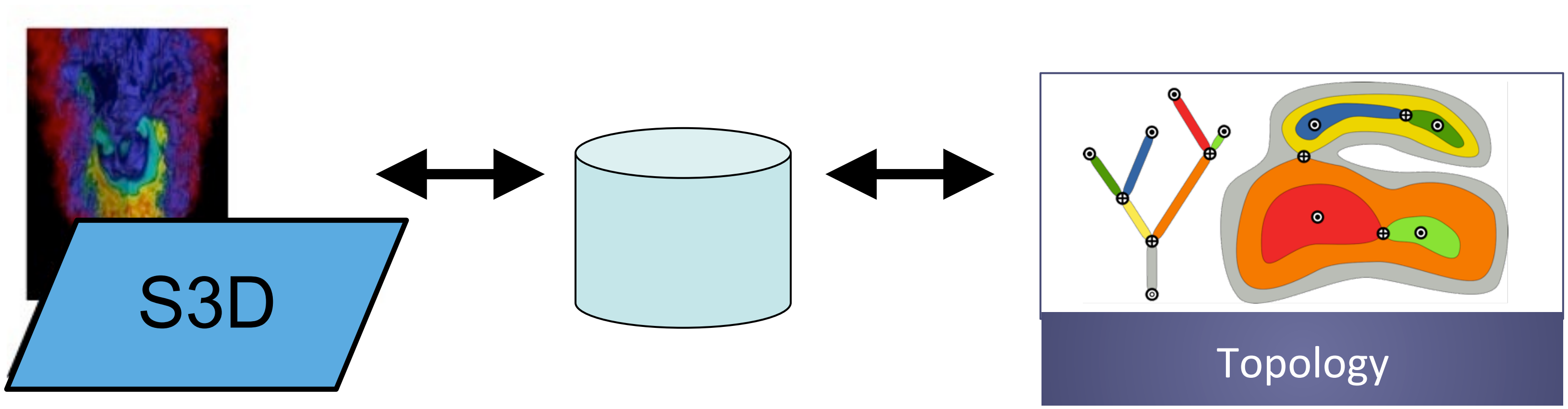
Volume Rendering



Topology

Identify features of interest

Example: Simulation + Data Analysis Workflow



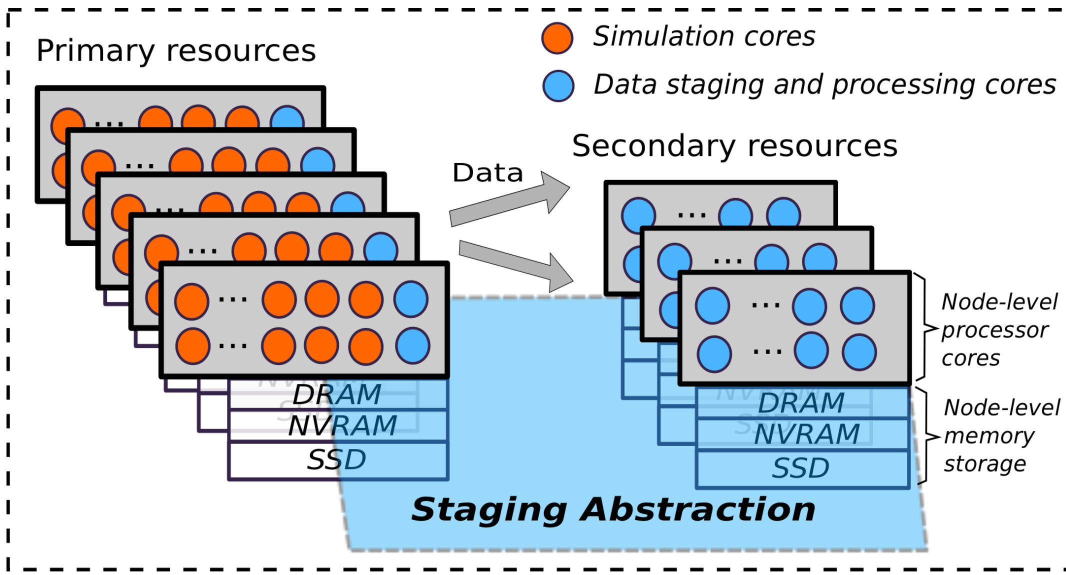
- Modeling data placement and data paths
  - **Deep memory hierarchy** 1
- Modeling in-situ analysis choices (cores sharing)
- **Opportunities for speculation** 2

1 Data Staging over Deep Memory Hierarchy

- Small DRAM capacity per core – even aggregated memory on dedicated nodes can hardly keep all coupled data (given the ratio of resource allocations for compute nodes and dedicated nodes)

Hybrid Staging

- Spans horizontally across compute nodes
- Spans vertically across the multi-level memory hierarchy, e.g. DRAM/NVRAM/SSD, to extend the capacity of in-memory data staging



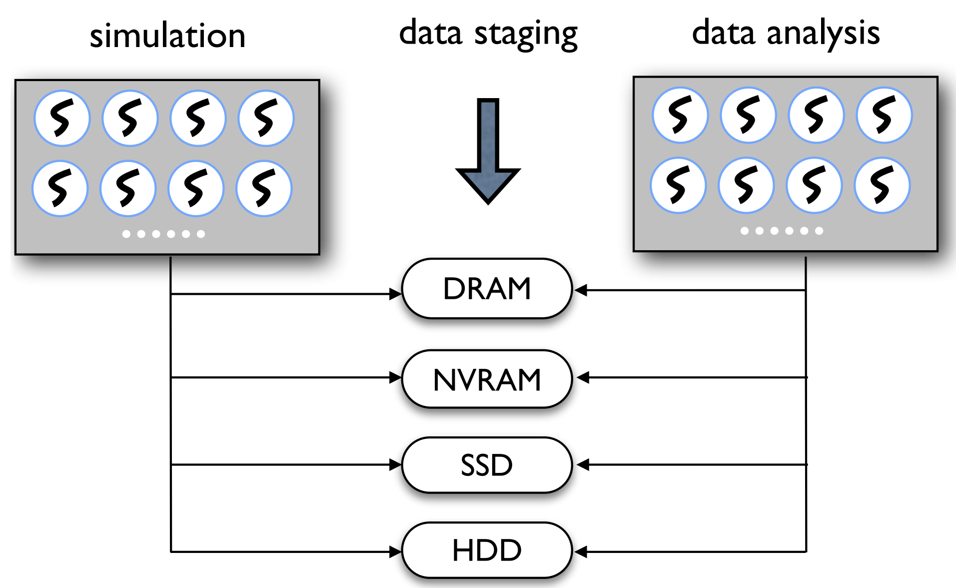
HPC core computation resource

1 Data Staging over Deep Memory Hierarchy

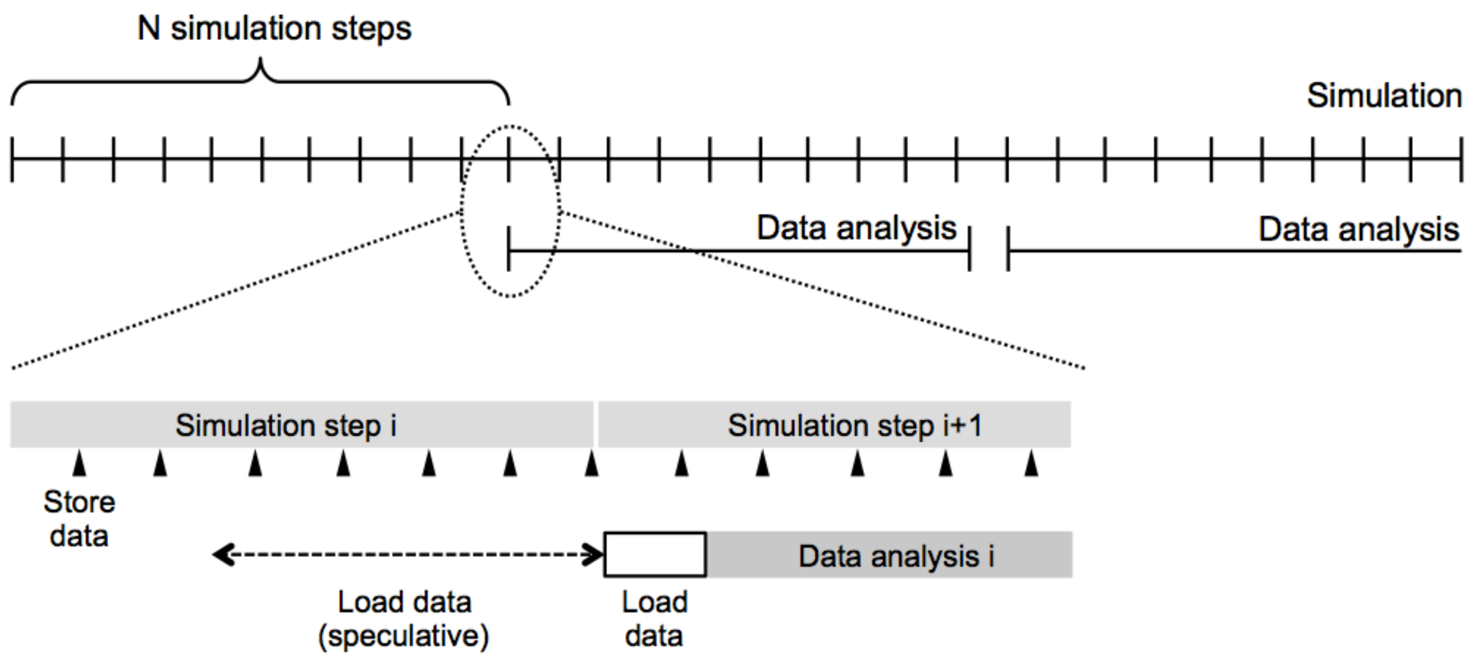
- Small DRAM capacity per core – even aggregated memory on dedicated nodes can hardly keep all coupled data (given the ratio of resource allocations for compute nodes and dedicated nodes)

Hybrid Staging

- Spans horizontally across compute nodes
- Spans vertically across the multi-level memory hierarchy, e.g. DRAM/NVRAM/SSD, to extend the capacity of in-memory data staging



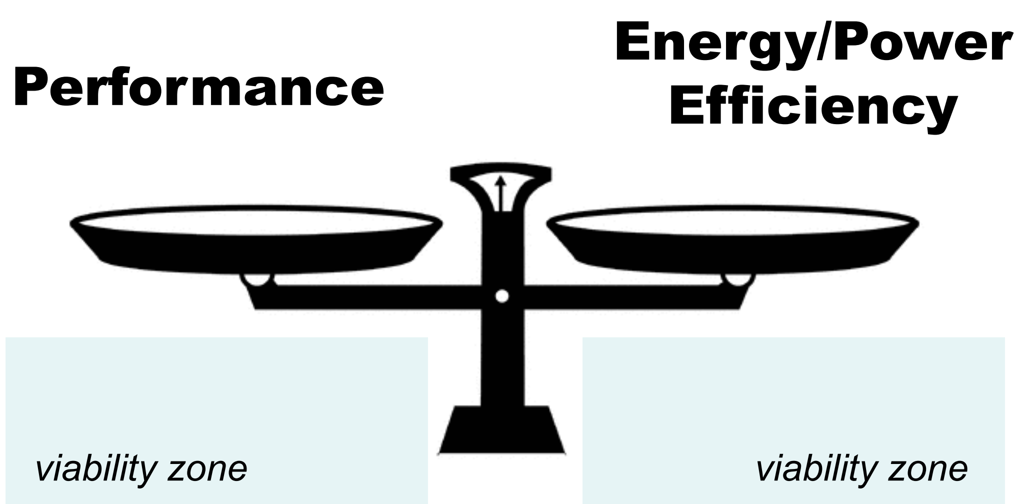
2 Synthetic Workflow for Understand Relative Behaviors



- Data analysis pipeline (e.g., S3D+Topology analysis)
- Synthetic kernels to evaluation relative behaviors
- Potential use of **speculative** data movement
  - Out-of-the-core data movement vs. traditional speculation at CPU level

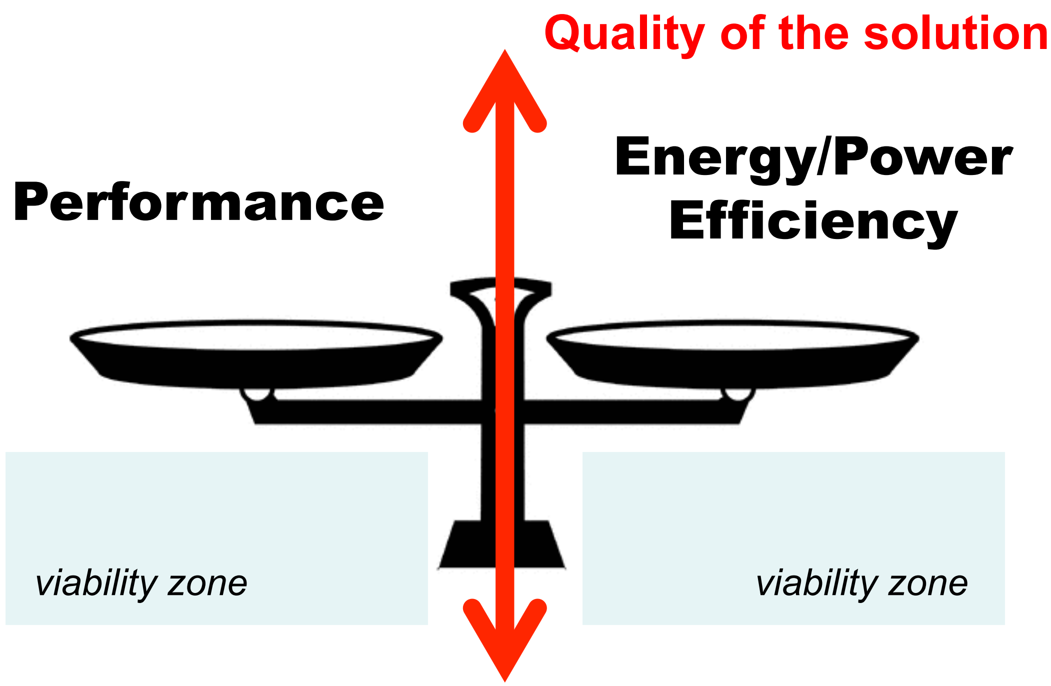
Understanding Behaviors and Tradeoffs

- Performance and Energy/Power
- Limitations when “viability zones” are exclusive



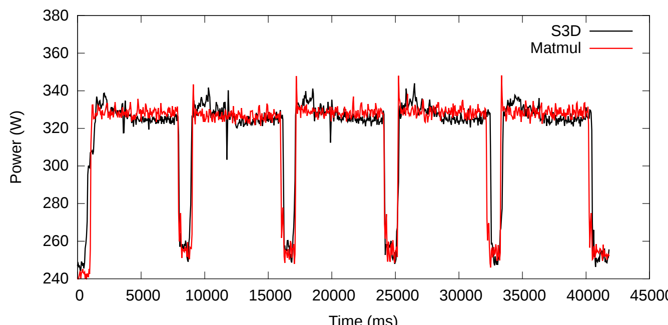
Understanding Behaviors and Tradeoffs

- Quality/accuracy of the solution
  - E.g., single/double precision, convergence values, AMR codes, etc.
- **Frequency of analysis** to represent quality of the solution



Evaluation Methodology

- Evaluation framework (single node)
  - Customizable multi-threaded framework which takes care of the workflow synchronization
  - Can run different kernel/applications as the simulation and analysis are customizable
- Synthetic kernels to evaluation relative behaviors
  - Matrix multiplication in simulation steps
  - Word finding in analysis steps



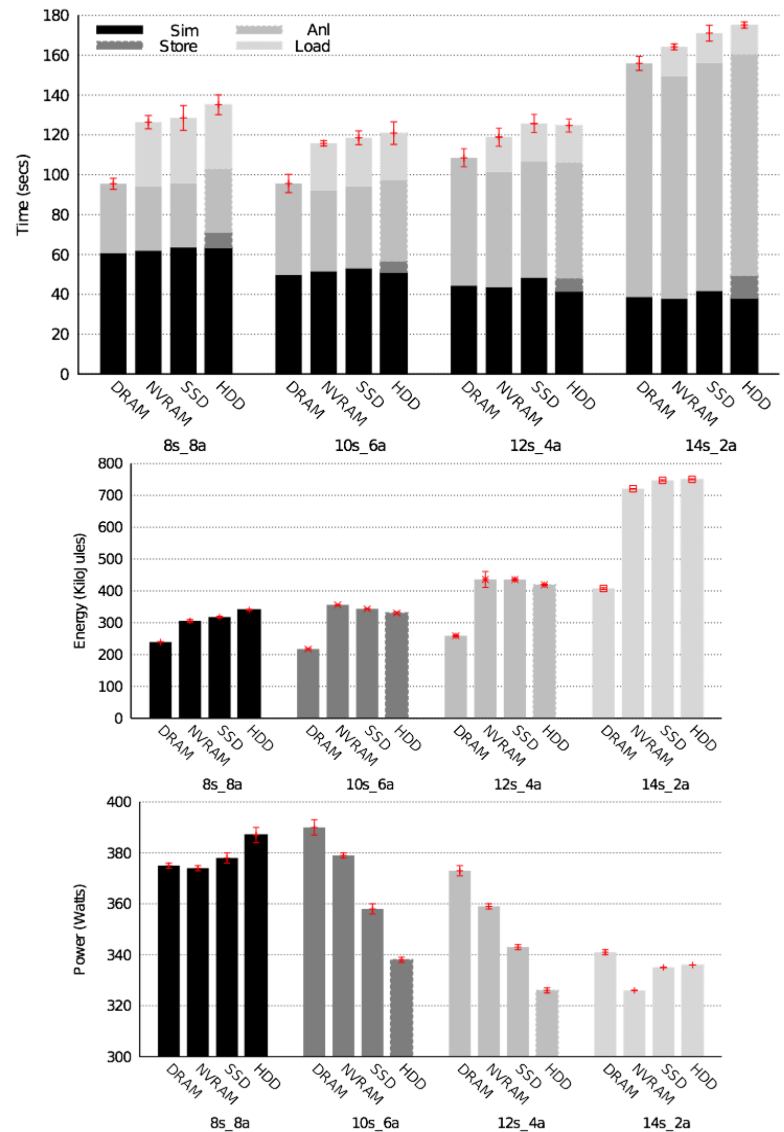
- Multiple customizable input parameters
  - Number of cores assigned for simulation/analysis
  - Data path (HDD, SSD, etc.)
  - Frequency of analysis, i.e. every x number of steps

Computational And data-enabled Platform for Energy efficiency Research (CAPER)

- NSF funded research instrument
- SyperMicro SYS-4027GR-TRT (support up to 8 GPUs concurrently)
- Phase 1
  - 8 servers with 2 Intel Xeon Ivybridge E5-2650v2 (16 cores)
  - 128GB of DRAM
  - 1TB of PCIe Flash-based NVRAM (Fusion-io iodrive2)
  - 2TB of SSD (RAID)
  - 4TB of hard disk (RAID)
  - Intel Xeon Phi 7120P
  - Infiniband FDR and 10G Ethernet
- Phase 2
  - NVIDIA K40
- Instrumentation
  - Coarse grain: PDU (1Hz)
  - **Fine grain: Yokogawa DL850E ScopeCorder (1KHz) – from modules at 10Ms/s**

1 Data Staging over (Deep) Memory Hierarchy

- Empirical evaluation on CAPER
- Execution time, energy consumption, and average power of the workflow's execution using different configurations and devices for data staging
- Each group of columns represents one configuration (number of cores for simulation/analysis)
- **Goal: finding sweet spots for in-situ data analysis**
  - In this example 12 core for simulation and 4 for analysis
  - Using very few cores for analysis delays simulation tasks
  - Higher power with power demanding devices (e.g., DRAM)



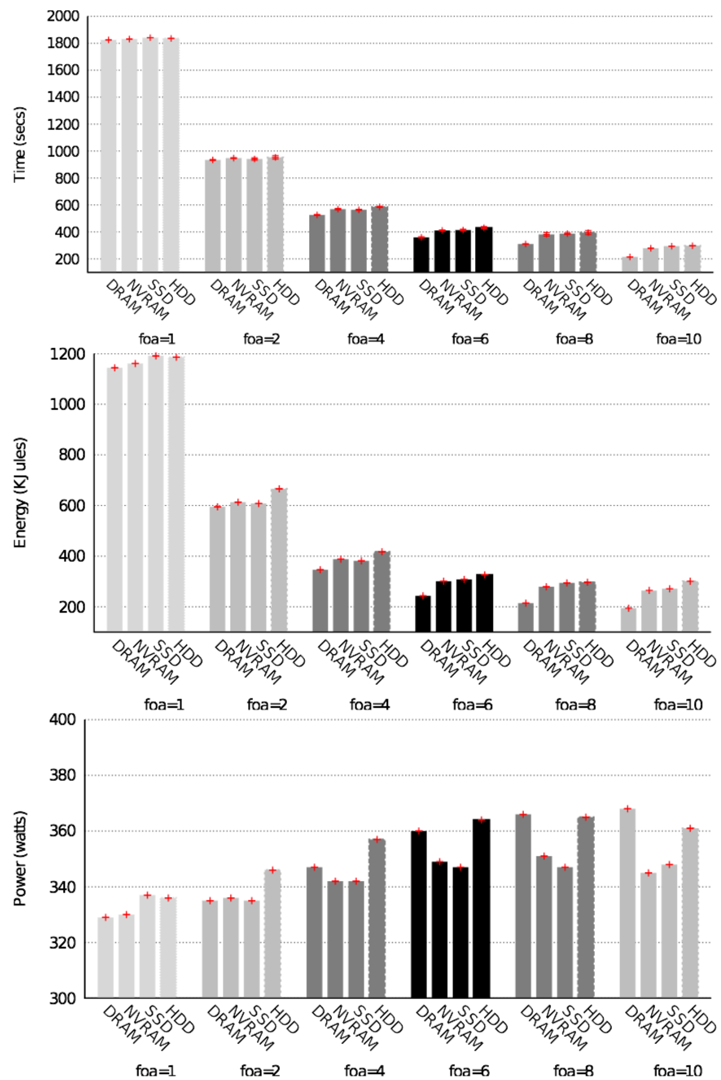
Rutgers

RDI²

Rutgers Discovery Informatics Institute

Tradeoffs with the Quality of Solution (Freq. of analysis)

- Execution time, energy consumption, and average power of the workflow's execution for different frequency of analysis ("foa") and different devices for data staging
- Frequency of analysis foa = k means that the data analysis is performed every k simulation steps
- **Execution time and energy consumption decreases as the frequency of analysis decreases**
- **However, average power increases as more computation/data movement is performed in parallel**

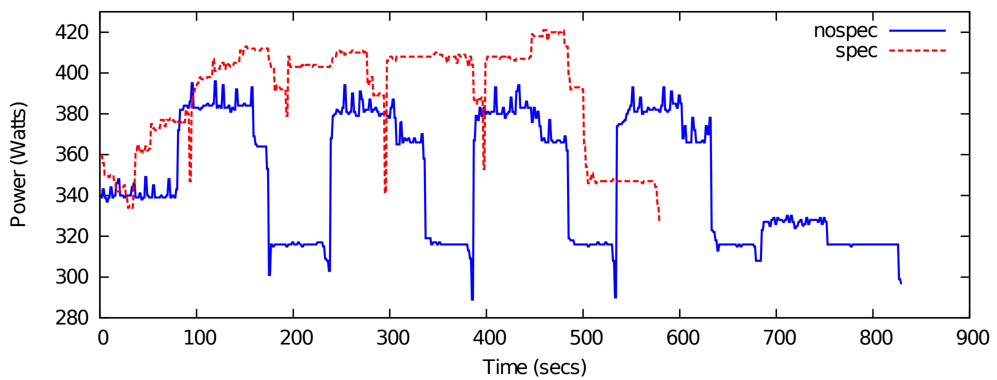


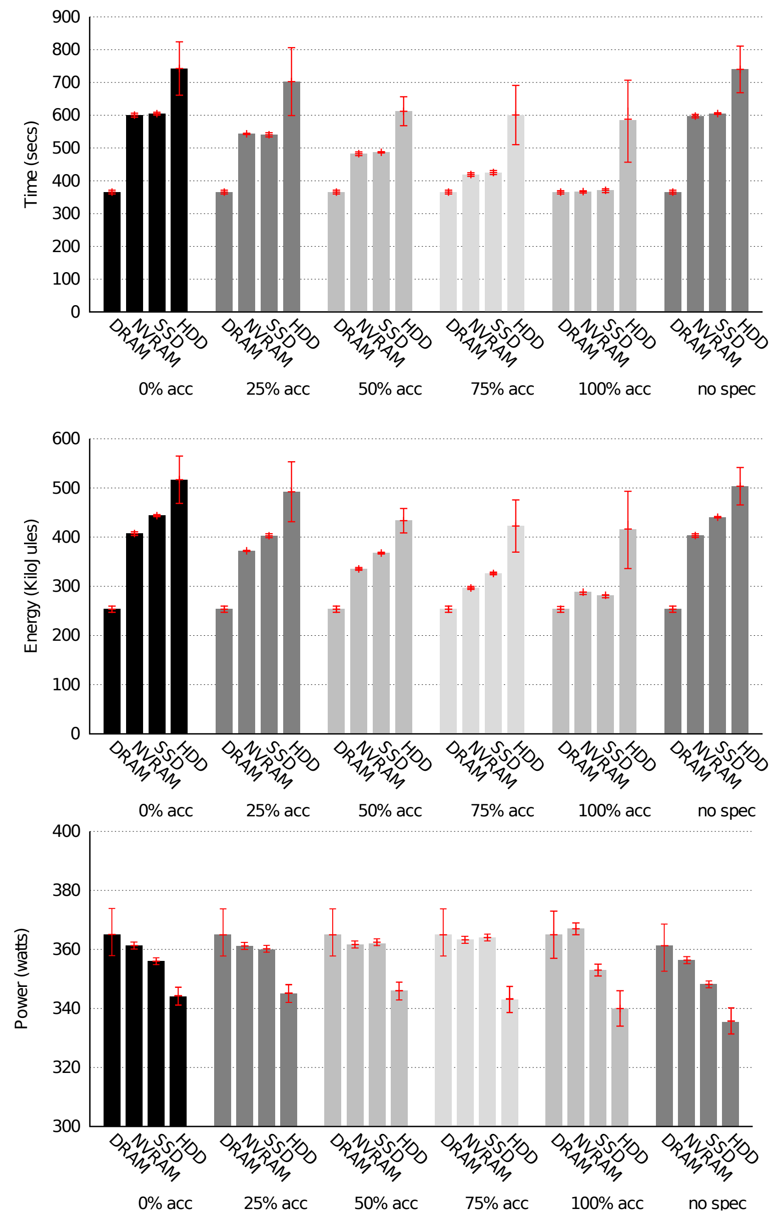
RUTGERS

RDI²
Rutgers Discovery Informatics Institute

2 Speculative Data Movement

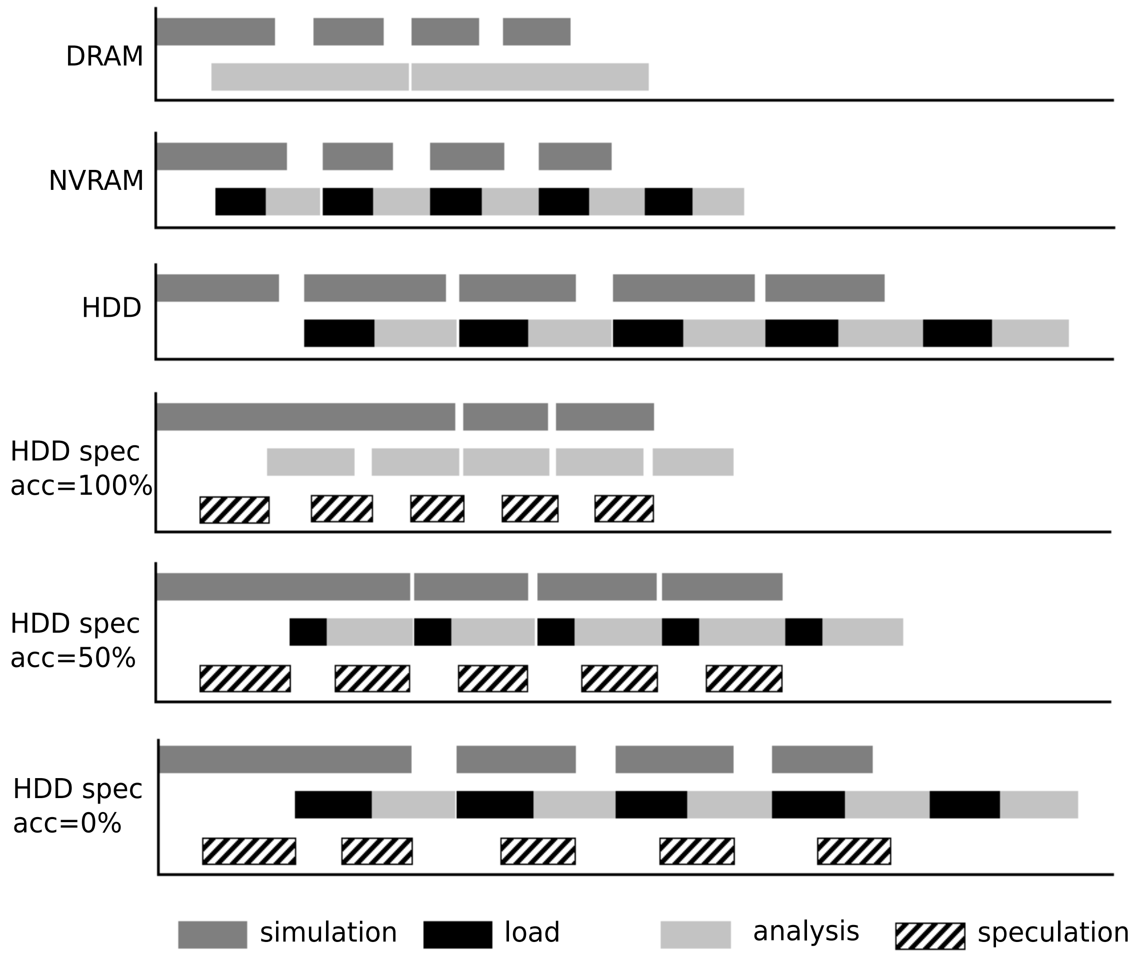
- Data speculation incurs little overhead when it is 0% accurate vs. no speculation both in terms of time and energy consumption
- The average power is higher when performing data speculation because it shares resources with the simulation and the analysis





Speculative Data Movement (cont.)

- Results present tradeoffs that can be exploited at runtime
  - E.g., Execution behavior with NVRAM is similar with HDD when data speculation is accurate
  - However, overall energy consumption is a bit higher with NVRAM (device power requirements)



Conclusions

- Costs (energy, latency) related to transporting, processing and analyzing increasing **data volumes and rates** are limiting the insights from extreme scale applications
- Energy/power-efficiency in combination with other objectives – understanding tradeoffs are important
  - **Quality of solution**, Performance, Resiliency, etc.
- Using **data speculation** in data-intensive workflows can positively impact energy consumption without much negative impact on performance or the quality of the solution
- Co-design is essential along multiple dimensions
  - E.g., **runtime system** to balance these tradeoffs

RUTGERS

RDI[2]

Rutgers Discovery Informatics Institute

Thank You!

Ivan Rodero, Ph.D.

Rutgers Discovery Informatics Institute
NSF Cloud and Autonomic Computing Center
Rutgers, The State University of New Jersey

Email: **irodero@rutgers.edu**
WWW: **http://nsfcac.rutgers.edu/people/irodero/**