Accelerating Scientific Discovery through Data-Driven Control and Real-Time Analytics

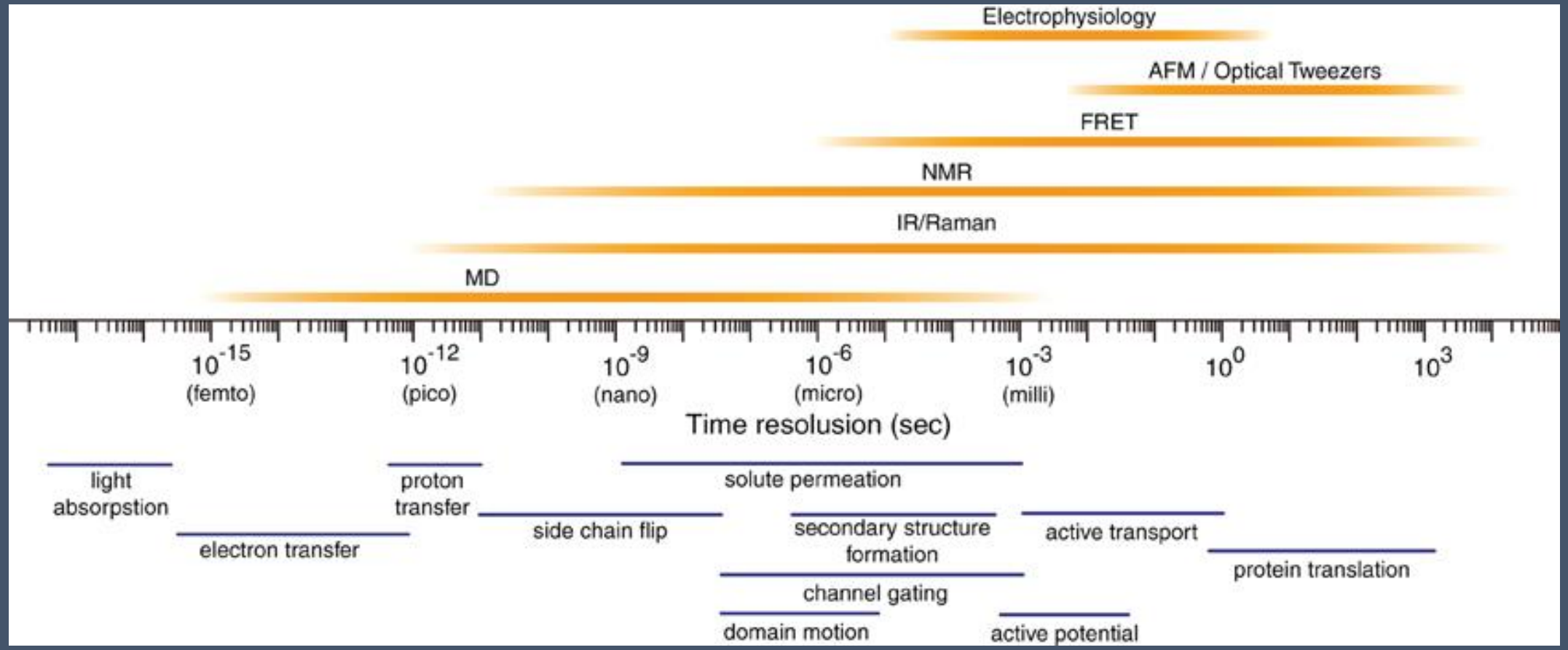


Ben Ring, Yanif Ahmad, Tom Woolf
Computer Science Department

JOHNS HOPKINS UNIVERSITY

Motivation: Exascale Scientific Exploration

Goal: Exploring rare events captured
in high-resolution simulation datasets



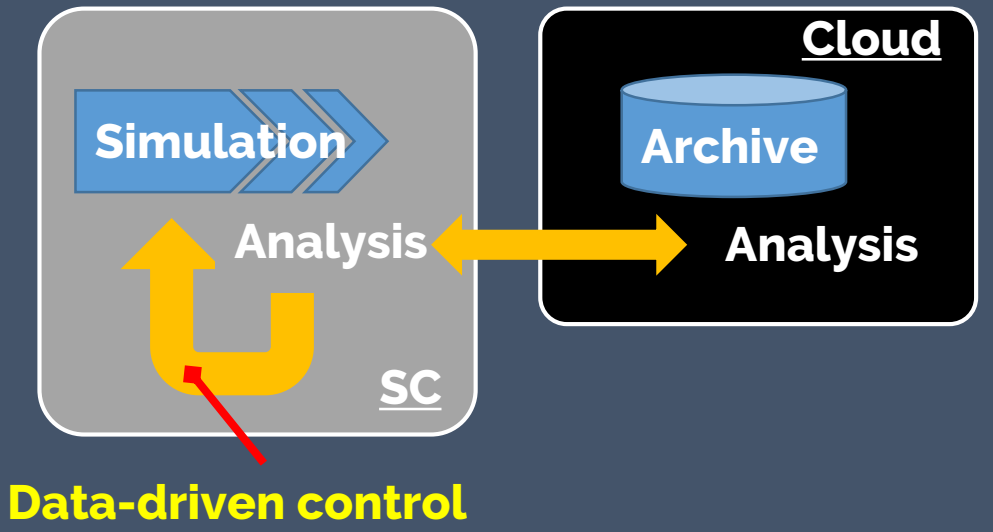
Motivation: Exascale Scientific Exploration

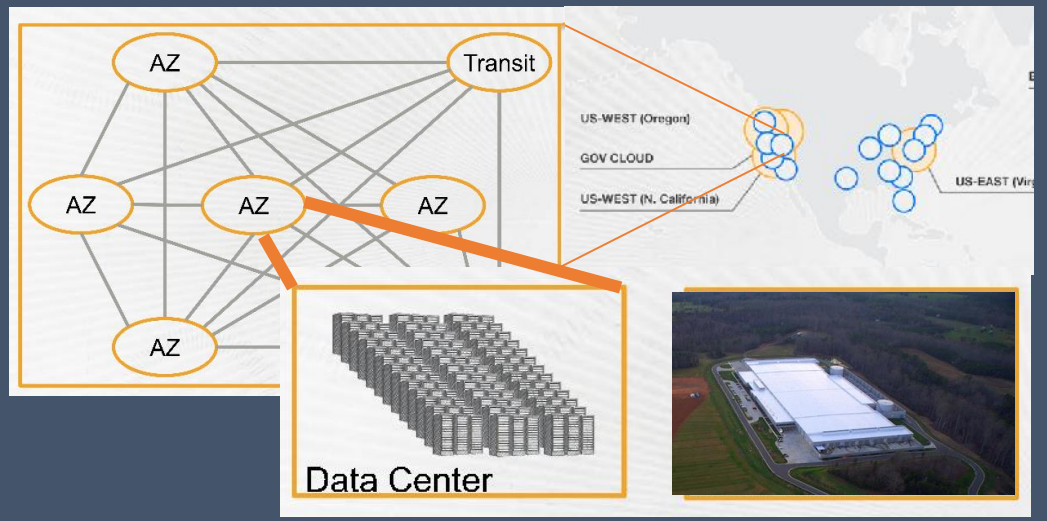
Goal: Exploring rare events captured in high-resolution simulation datasets

Exascale design principle: analysis guides simulation

Challenges:

- Data-driven control, as a stream data management problem
- Big data system design for HPC:
  i) external scheduling
  ii) symbiosis with cloud architectures





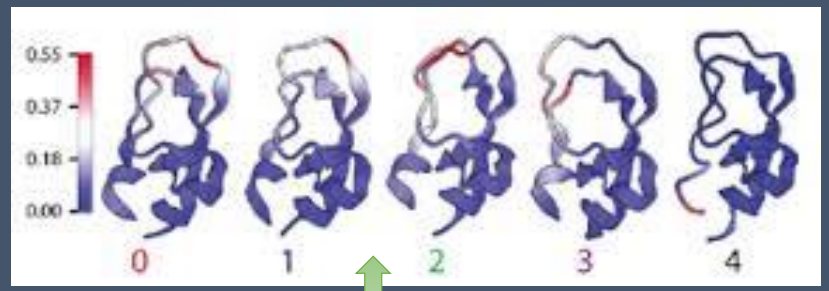
http://mvdirona.com/jrh/talksandpapers/JamesHamiltonReInvent2014.pdf

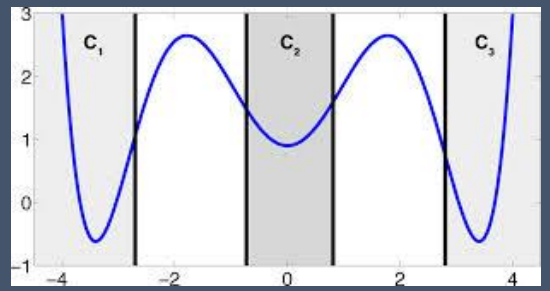
MD Simulation

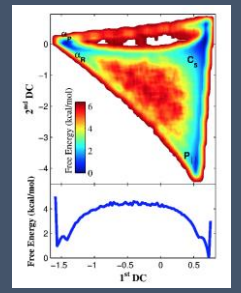
Trajectory Windowing

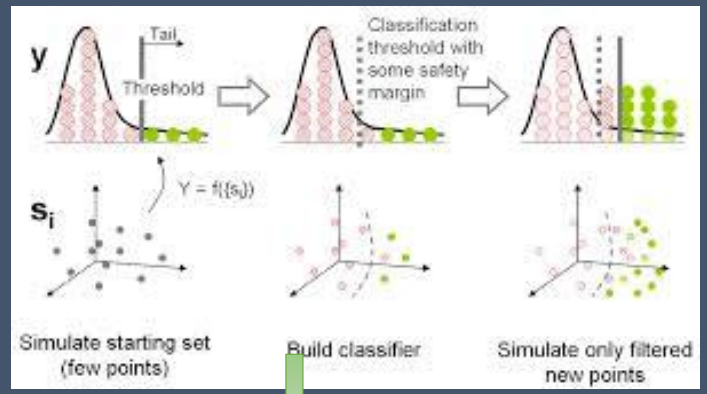
Noninear Dimensionality Reduction and Clustering

Data-driven Control







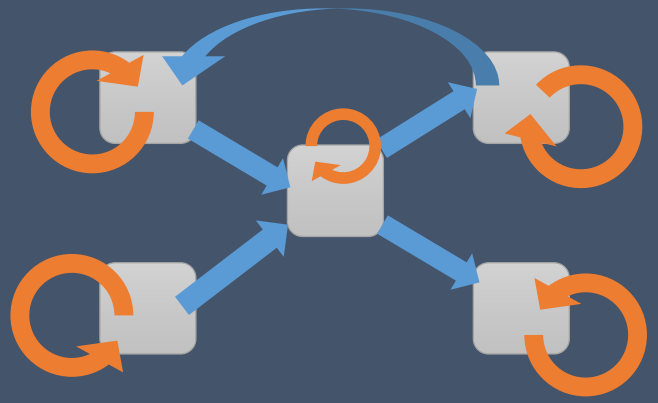


Autonomic Simulation Data Exploration

Supercomputing-specific challenge:
No queues or notification service for streaming

Our approach:
High-level **periodic**, **elastic**, "thread" abstraction for driving parallel dataflow operators via SC scheduler

System optimizer:
Manage "thread" elasticity, and minimize I/O



Operator thread → Dataflow

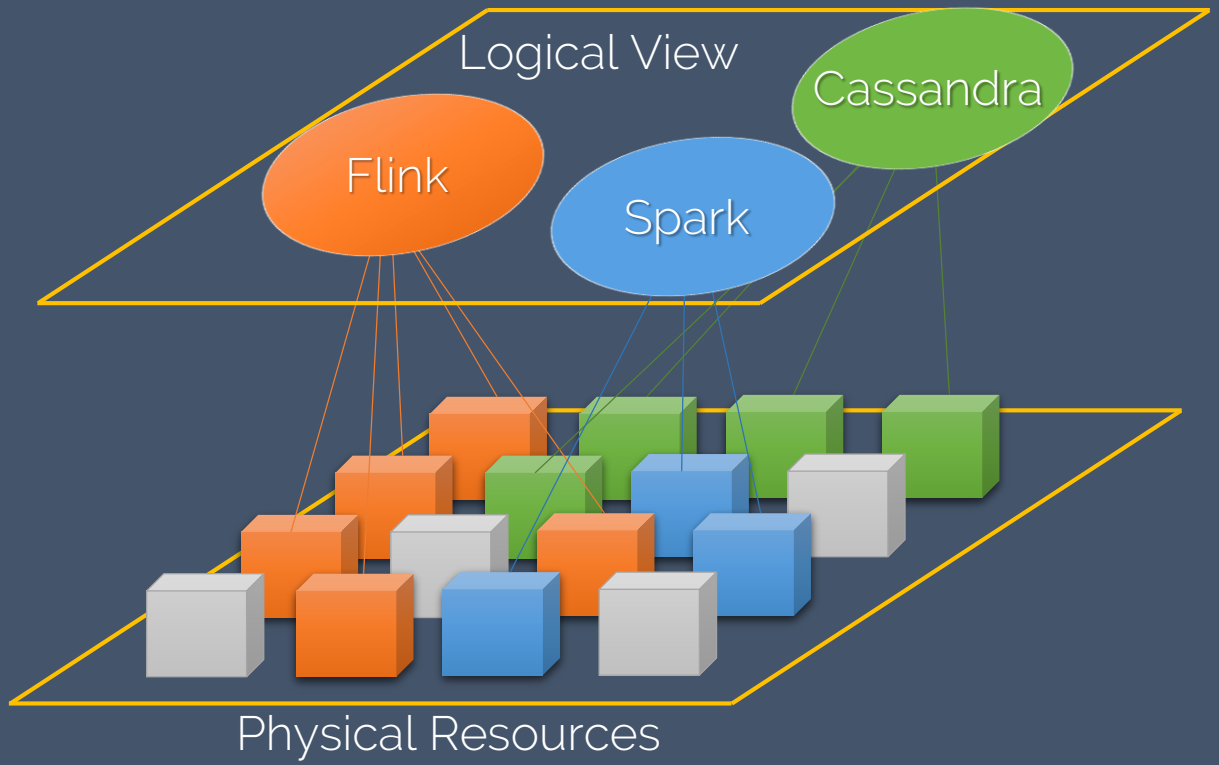
Building Block: Microservice Overlays

Operators as microservices (as inspired by containers):

- Index retrieval
- Analysis
- Data transformation

Execution engine:

- Fixed-core, persistently scheduled catalog
- In-memory shared state as cluster FS bypass



Query & Analysis Model

Continuous analytics, alongside interactive query processor implemented as an operator

Workload: **Combinatorial data reduction**

Use cases: i) Multiresolution analysis

ii) Feature analysis and representation learning

iii) Subspace exploration and resimulation

Optimization in Data-Driven Control

Offline objective

$$\max_{D} \sum_{d \in D} U(sim_d) \ / \ cost_d$$

Online selection

$$w_{d,t_{i+1}} = w_{d,t} \ + \ U(sim_d)/cost_d$$

$$sim_{t+1} = \max_{D} w_{d,t}$$

User-defined control objective on coarse-grained analysis

- Decomposable, to determine contribution of individual simulations

Simulation parameter selection

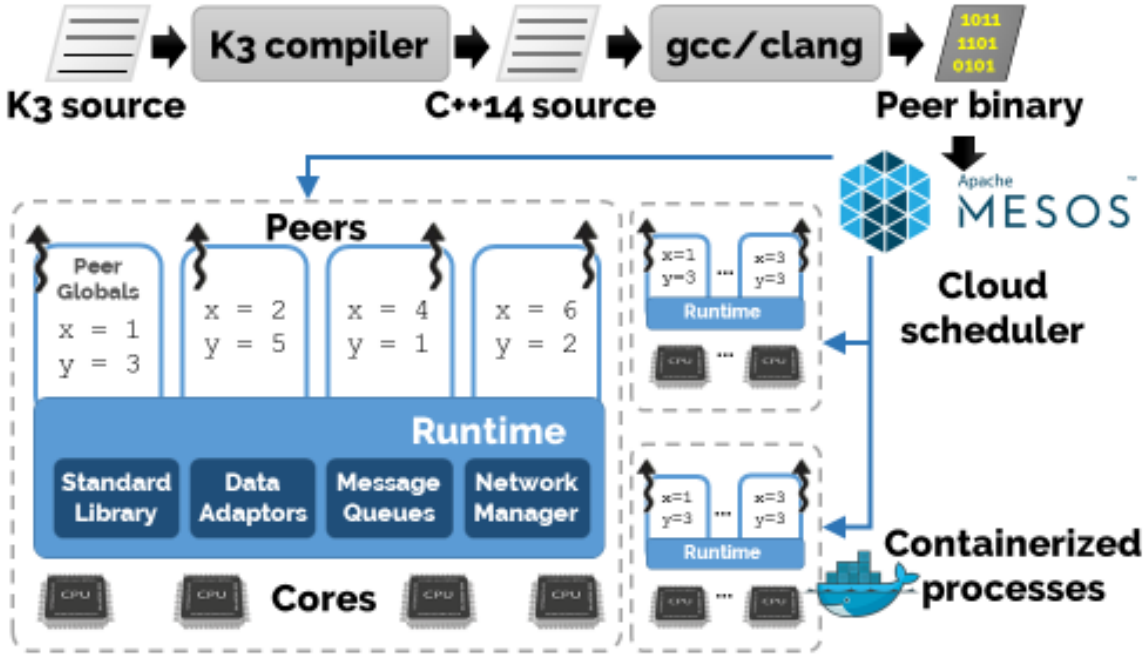
- Extend with resource-oriented objective for joint optimization
- Evaluate by competitive analysis

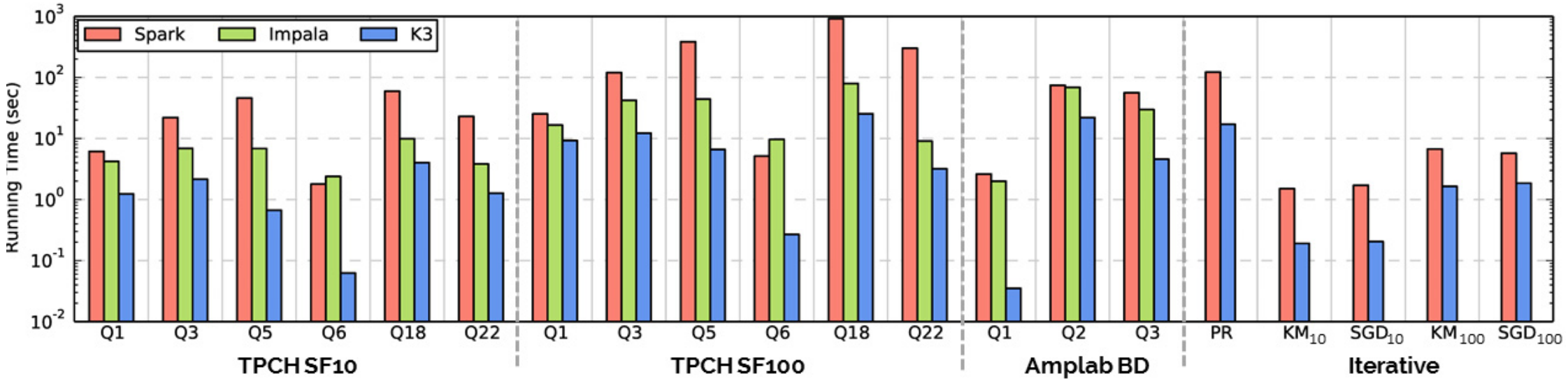
Project Teasers

K3: https://github.com/damsl/k3

A language, compiler and runtime for building big data systems

- Automatic compiled memory management for in-memory DBMS
- Declarative systems programming
- Speedup: 2.4x-74x (Spark)





Cluster-Scale View Engine

https://github.com/damsl/k3-mosaic

Cloud Factory: Specialization as a Service