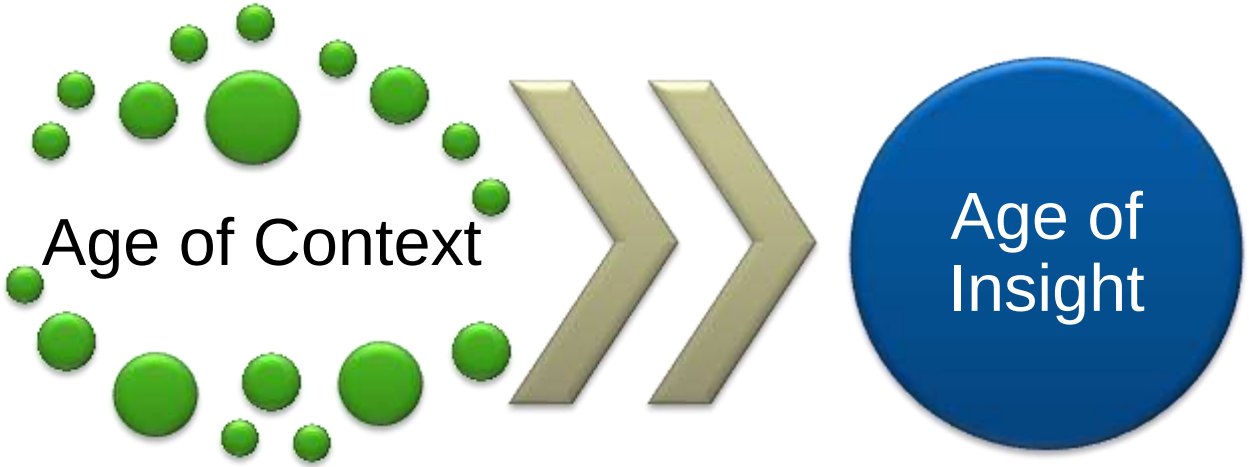
HPC + D + A = HPDA?

Barry Bolding, PhD
SVP & Chief Strategy Officer
Cray Inc.
bbolding@cray.com

Era of Mobile

https://medium.com/crossing-the-pond/into-the-age-of-context-f0aed15171d7
Age of Context - Robert Scoble and Shel Israel

Context can be powerful, but…

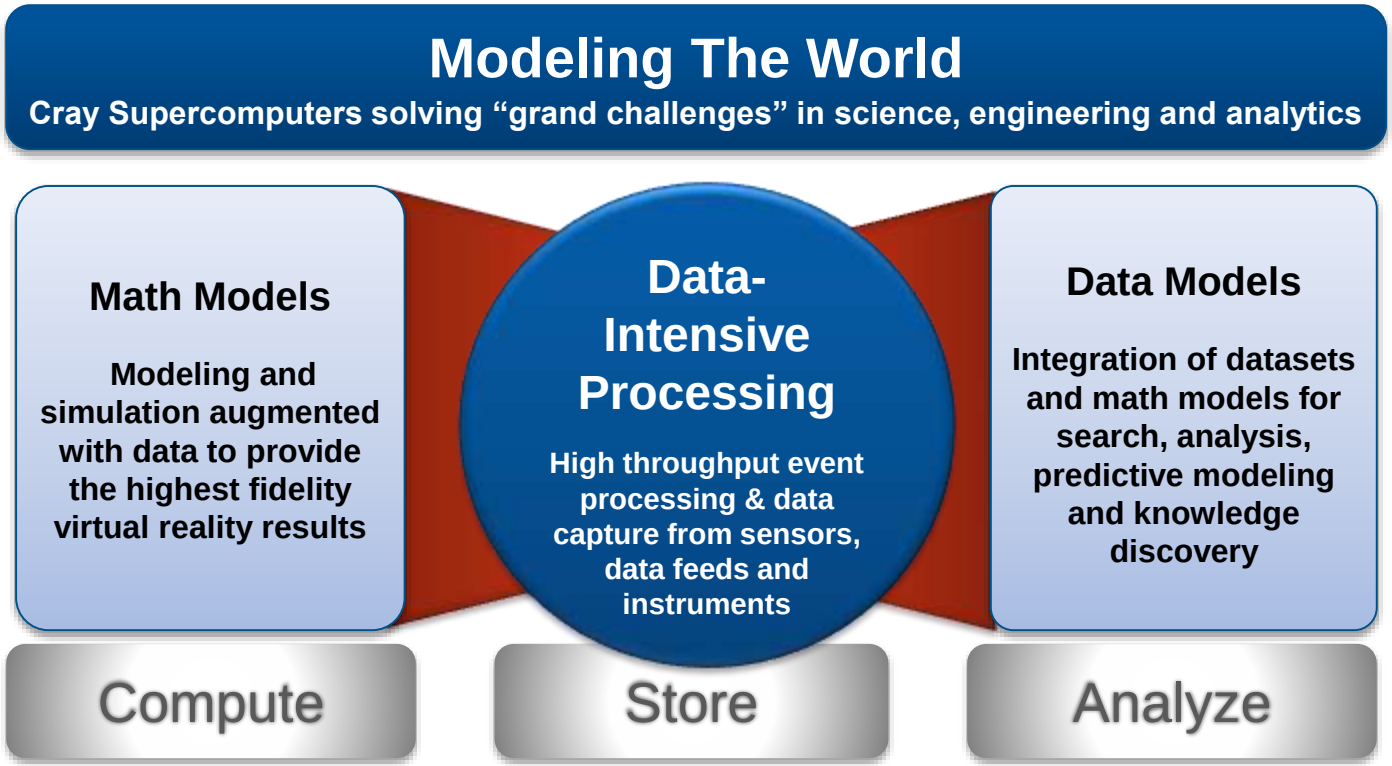


Contextual interpolation

Insightful extrapolation

Cray's Vision:

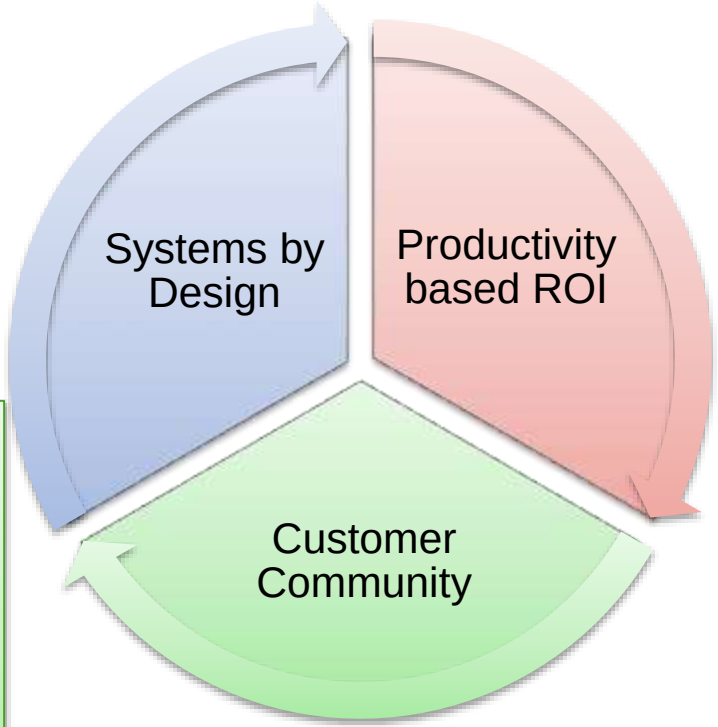
The Fusion of Supercomputing and Big & Fast Data



Cray's Unique Strengths in Supercomputing + Analytics

We design our systems for **production computing and analytics at scale**.

The Cray community of customers and partners are the most demanding and enabling. Allows Cray to push the boundaries of scalable systems.

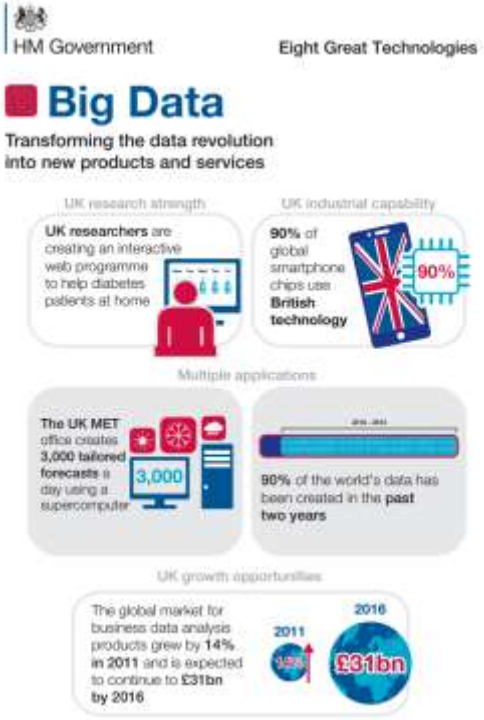


Cray takes the productive applications view of ROI:

- Match system designs to the applications
- Enable highly productive programming environments.
- Build solutions that are flexible and upgradeable.
- Lower TCO than commodity clusters and cloud for data-intensive applications

The Alan Turing Institute

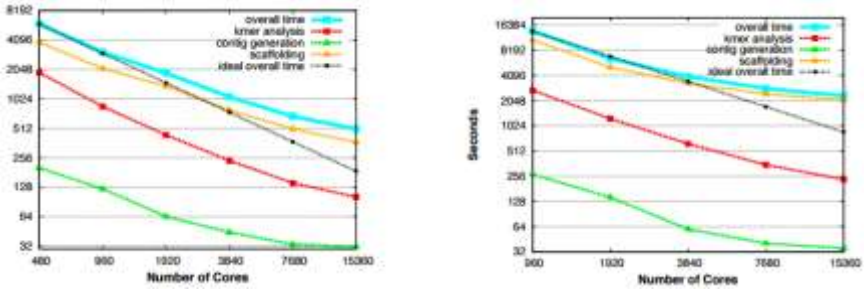
- **UK national institute for data sciences**
  - *to break new boundaries in how we use big data in a fast moving, competitive world*
- **Joint venture: founder Universities and EPSRC**
  - Cambridge, Edinburgh, Oxford, UCL, Warwick
- **Launch partners**
  - Lloyds Register Foundation, GCHQ and Cray
- **UK Government investment in "Eight Great Technologies"**
- **https://turing.ac.uk/**



“We don’t just want insights, we want actionable insights.”

Peter Grindrod, University of Oxford

Human Genome Assembly – Under 9 Minutes on "Edison"



The end-to-end scaling of the team's HipMer genome assembler approach showing the human genome scalability on Edison on the left and the more complex wheat genome on the right (both axes are in log scale).

- Cray XC30 "Edison" reduces the assembly time of a complete human genome to 8.4 minutes
- A UPC modified version of the Meraculous code (Hip-Mer) was 170x faster.
- Eliminating the analytics back-log and fueling the drive towards precision medicine
- Cray's CX30 with Aries interconnect coupled with high memory bandwidth and memory on each node provided the I/O bandwidth required for genome assembly work

References: http://www.theplatform.net/2015/08/20/supercomputer-force-knocks-human-genome-assembly-under-9-minutes/
http://gauss.cs.ucsb.edu/~aydin/sc15_genome.pdf

“Using our HipMer technology enables for the first time assembly throughput to exceed the capability of all the world's sequencers”

HipMer: an Etreme-Scale DeNovo Genome Assembler
Georganas, Buluc, Chapman, Hofmeyr, Aluru, Egan, Oliker, Rokshar and Yelick

Magnus (CX30) Whitefly study at IVEC

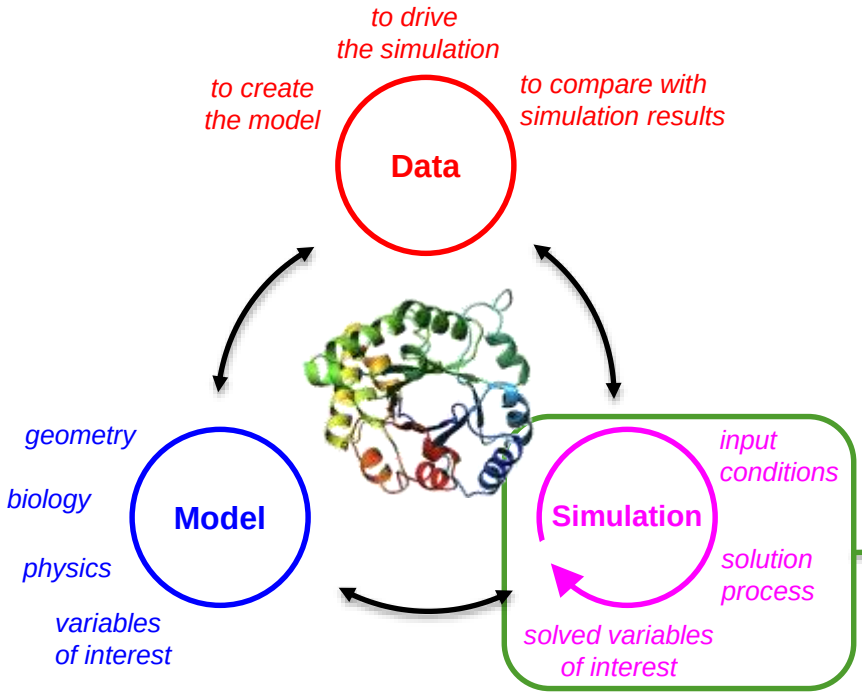
They are battling a species complex of at least 34 morphologically indistinguishable species.

- Dr. Laura Boykin
  - funded by the Gates Foundation and a TED Fellowship.
- Awarded time on “Magnus,” Her team is marrying genomics, supercomputing and evolutionary history to help African farmers develop management strategies and breed pest-resistant crops

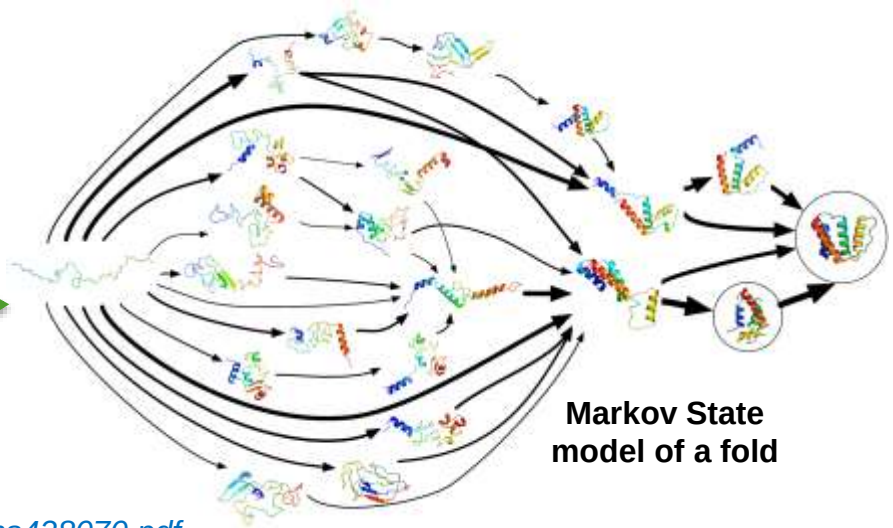
“It’s a massive problem. I’m one of 15 principal investigators working on a new project whose mission is to give farmers a cassava plant that’s resistant to the viruses and the whiteflies.”

Dr. Laura Boykin

Protein Folding – Mixed Simulation and Analytics



- Focuses on *how* protein folding happens
- Model possible paths to folded end-state
- Temporal resolution matters, but drives data size

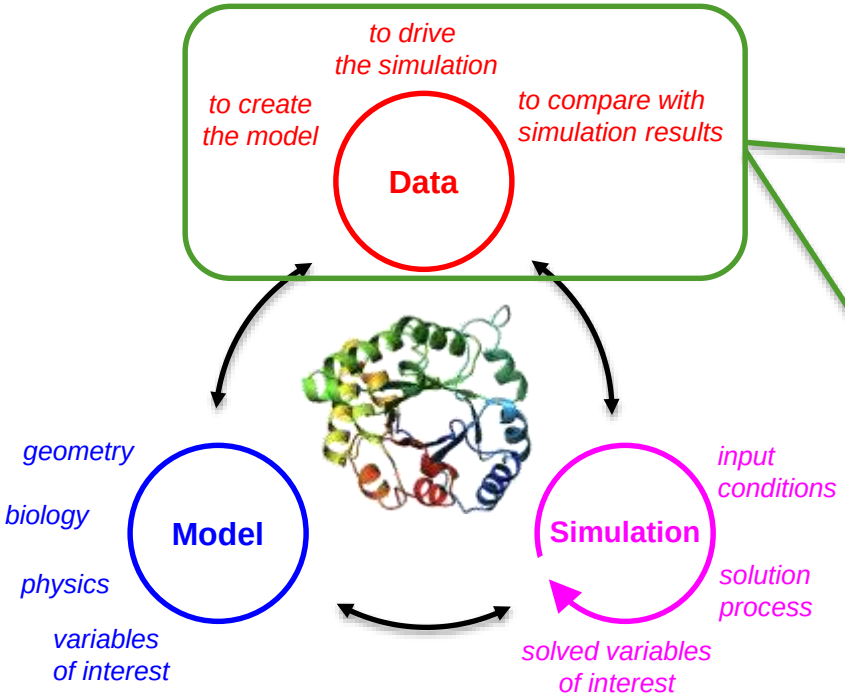


Markov State model of a fold

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3673555/pdf/nihms428070.pdf
Curr Opin Struct Biol. 2013 February ; 23(1): 58–65. doi:10.1016/j.sbi.2012.11.002.

Copyright 2015 Cray Inc.

Protein Folding – Mixed Simulation and Analytics



Enabling in-situ data analysis for large protein-folding trajectory datasets

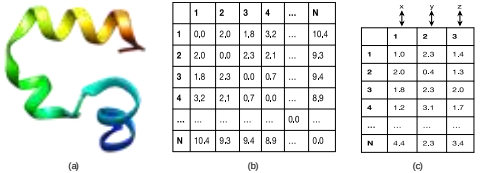


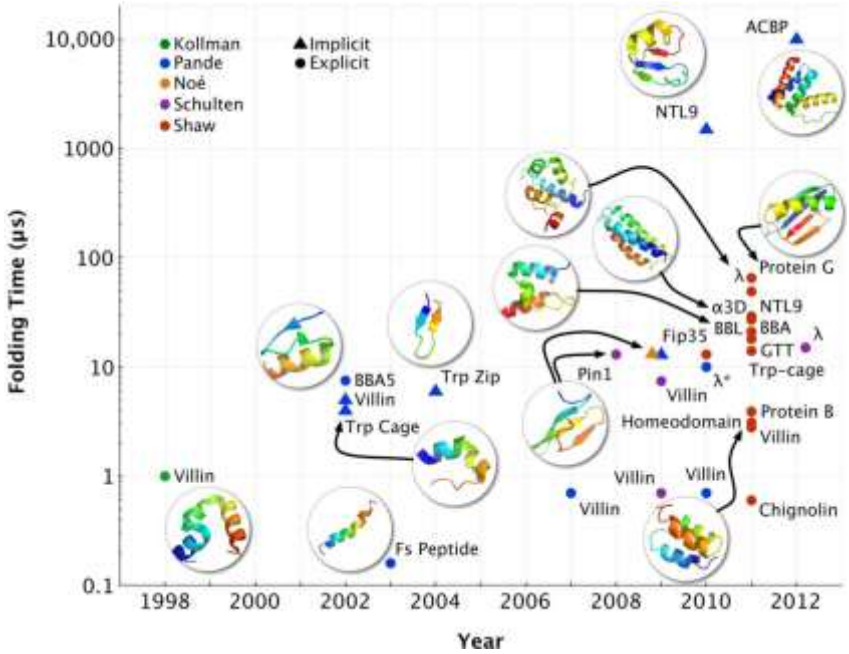
Figure 1: One conformation of the villin HP-35 protein (a); part of its distance matrix using only its backbone atoms in the conformation (b); and three eigenvectors and the associated eigenvalues capturing and synthesizing the conformation geometry (c).

Abstract—This paper presents a one-pass, distributed method that enables in-situ data analysis for large protein-folding trajectory datasets by executing sufficiently fast, avoiding moving trajectory data, and limiting the memory usage.

In-situ analysis of folds – Dimensionality reduction using PCA & MDS

Automated classification in protein databases

Copyright 2015 Cray Inc.

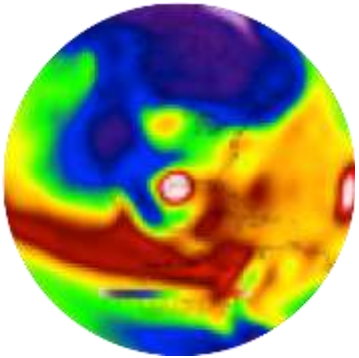


Thomas J. Lane, Diwakar Shukla, Kyle A. Beauchamp, and Vijay S. Pande

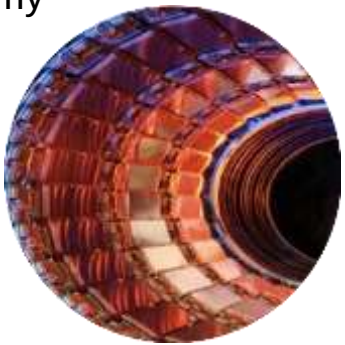
DOE Facilities are Facing a Data Deluge

Astronomy

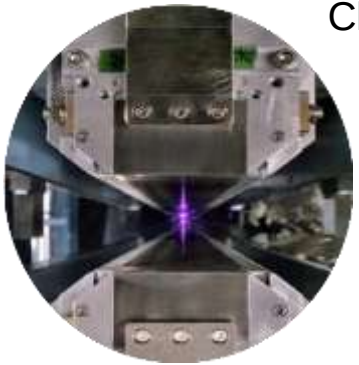
Genomics



Climate

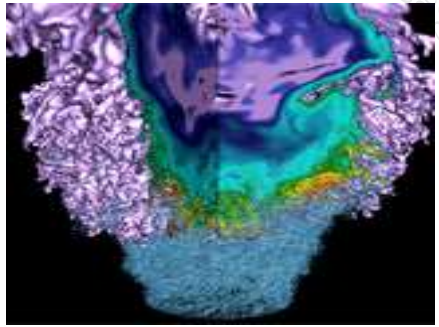


Physics

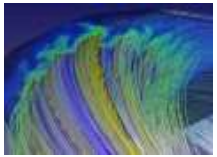


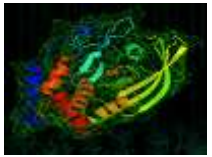
Light Sources

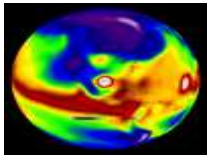
Cray/AMPLab/LBL collaboration

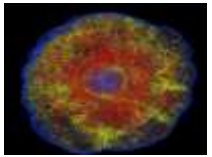


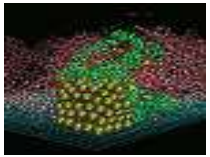
NERSC











U.S. DEPARTMENT OF ENERGY | Office of Science

Top Data Analytics Problems for NERSC users

1. **Generative Model for the Visible Universe**
   - Joint Inference across telescope images
2. **Pattern Detection for Climate**
   - Scalable Deep Learning on Cori
3. **Stephen Hawking Device**
   - Machine Learning for speech prosthetic
4. **Google Maps for BioImaging**
   - Semantic databases
5. **Genome Assembly for Wheat**
   - Graph Analytics

Data Characteristics

Science Domain	Data Source	Data Characteristics	Data Volume	Analysis Challenge
Cosmology	Multiple Telescopes	Noisy, multi-band, artifacts	O(100) TB	Data Fusion, Inference
HEP	Anti-Neutrino detectors	Noisy, artifacts, spatio-temporal	O(10) TB	Pattern/Anomaly Detection
BioImaging	*Mass-spec instruments*	*Noisy, artifacts, multi-modal*	*O(10) TB*	*Dim. Reduction Clustering, Pattern Detection*
Genomics	Sequencers	Missing data, errors	O(1-10) TB	Clustering, Pattern Detection
Neuroscience	Neural Recorders	Spatio-temporal, high dimensional, noisy	O(1) TB	Dim. Reduction, Pattern Detection
Climate	Satellites, Simulation o/p	Multi-variate, spatio-temporal	O(10) TB	Pattern/Anomaly Detection

"Current HPC machines are optimized for scientific simulations, arithmetic-intensive workloads, and regular computation, whereas data-analytics applications requires hardware optimized for bandwidth, throughput, concurrency and all-to-all communication."

Tumeo & Feo (IEEE Computer, Aug)

“The age of insight will require hardware and software infrastructures and ecosystems optimized for data analytic modeling & data-intensive simulations, including regular and irregular data access, high bandwidth, throughput, concurrency, all-to-all communication and scalability.”