

A Collection of Presentations from the BDEC2 Workshop in Kobe, Japan

February 19–21, 2019

Call for Short Presentations for the BDEC2 Workshop in Kobe, Japan	iv
Workflow environments in Fog-to-cloud infrastructures	1
Rosa M. Badia	
Considering a Clean Slate	7
Micah Beck	
Toward an Advanced Cyberinfrastructure Platform	12
François Bodin	
Multi-Hybrid Accelerated Computing GPU + FPGA = ?	18
Taisuke Boku	
Lossy Compression for a New, Shared, Advanced Cyberinfrastructure Platform (ACP)	28
Franck Cappello	
Industrial Materials Design – An Exemplar for BDEC	34
Alok Choudhary	
Converged Software Platform for Data Analytics and Extreme-Scale Computing	39
Carlos Costa	
Improving Workflow Management Systems	45
Ewa Deelman	
Edge Computing Applications	52
Nicola Ferrier	
Current and Future Plans with AI Bridging Cloud Infrastructure	58
Katsuki Fujisawa	
Convergence of Equation-Based Modeling and Data Analytics on HPC Resources	64
Kohei Fujita	
HEP Challenges for HPC	69
Maria Girone	
The Promise of Learning Everywhere and MLforHPC	76
Geoffrey Fox and Shantenu Jha	
Computational Media	81
Yoshinari Kameda	
Optimized Images via ManifestList	86
Christian Kniep	
Geospatial and Global Earth Mapping and Modelling Applications Requirements for the BDEC2 Workshop	103
William Kramer	
Benchmarking Huawei ARM Server Processor for HPC Workloads	109
James Lin	
Innovative Method for Integration of Simulation/Data/Learning in the Exascale/Post-Moore Era	115
Kengo Nakajima	
Farming the Environment into the Electrical Power Grid	121
Yiwei Qiu	
System Infrastructure for Elevating Edge to be a Peer of the Cloud	128
Kishore Ramachandran	
BDEC2 Kobe	133
Glenn Ricart	
Earth System Modelling: Requirements and Challenges	138
Kim Serradell	

Comments for BDEC Meeting	145
Dan Stanzione	
A Brief Report on the Utilization Status of Sunway Taihulight	154
Lin Gan	
Convergence of Computing and Storage Models in Big Data and Extreme Computing	160
Martin Swany	
Application Drivers	166
Christine Sweeney	
Building the Open Storage Network	171
Alex Szalay, Christine Kirkpatrick, Kenton McHenry, Alainna White, Steve Tuecke, Ian Foster, and Joe Mambretti	
Scientific Methods Transformation via AI/Deep Learning & Advanced Cyberinfrastructure Platforms (ACP): Fusion Energy Exemplar	178
William M. Tang	
In Situ Data Analytics for Next Generation Molecular Dynamics Workflows	184
Michela Taufer	
Patient Dossier: Implementing Medical Queries over Distributed Resources	192
Miguel Vazquez	

Acknowledgment

This workshop was supported in part by the National Science Foundation under Grant No. 1849625.

Call for Short (5 slides or less) Presentations for the BDEC2 Workshop in Kobe Japan

Last year's BDEC Report highlighted two momentous trends: The first is the transformation in scientific methods brought on by the ongoing revolution in machine learning. The second is the explosive proliferation of data generators spreading out across the "digital continuum," from cloud and HPC data centers to major new instruments, sensor networks and cyber-physical systems in the vast "data periphery." With these trends in view, the goal of the BDEC2 workshop series, and the collateral activities of its working groups, is to help develop a plan for a new, shared, advanced cyberinfrastructure platform (ACP) that can support future-looking applications in science and engineering. The previous Bloomington meeting began the process of ensuring that the BDEC2 platform design will be firmly rooted in the requirements that future applications are likely to exhibit.

The upcoming Kobe meeting will focus on the design of the ACP, so the profile of that design will need to reflect the nature of those application requirements. More specifically, we aim to develop a draft of a *reference architecture* for a future ACP, i.e., a "... design pattern that indicates how an abstract set of mechanisms and relationships realizes a predetermined set of requirements," and which thereby guides the realization of actual systems based on that pattern.

An initial draft the architecture document is available (<http://bit.ly/bdec2-pf-draft>), but it is still very much a work in process and will continue to be updated as meeting approaches. Of course, comments and suggestions are welcome. The primary objective of the Kobe meeting will be to catalyze a dialogue among participants in which that skeleton design becomes more well defined, addressing the common, basic mechanisms (i.e., services, protocols, software building blocks) that can be composed and used to meet a diverse set of application requirements.

Accordingly, we invite workshop participants to submit presentations of **no more than 5 slides** addressing either the base level components/services/software building blocks or the most basic application requirements/patterns that future ACP must or should support. Here are a few examples of such services and/or application patterns include the following:

1. Raw provisioning or containers of server cloud resources, so communities can pick their own stack,
2. The high volume distribution of files from a single network location to multiple destinations distributed throughout the network within a limited period of time but without a requirement of extremely low latency or skew.
3. A high quality Content Delivery Network (CDN), as in 2, but with the ability to process data anywhere along the branches of the tree.
4. The application of a partially ordered graph consisting of in-situ data transformation and movement operations to data stored at multiple locations distributed throughout the network.
5. Aggregation/merger of a set of data items from a distributed set of sources distributed throughout the network to a single network location with the application of compression of a global reduction function across the set.
6. An object store with "DropBox" like functionality for data publishing.
7. Automated resource management and arbitration.

The organizers will review and categorize the submissions and will select presentations for short talks based upon relevance of submissions, strategic vision and expertise. However, all presentations submitted will be published online with the other products of the workshop. These short presentations are due by February. 14 and should be sent to Terry Moore (tmoore@icl.utk.edu). If you have any questions, please contact BDEC2 workshop organizers: Jack Dongarra (dongarra@icl.utk.edu), Pete Beckman (beckman@mcs.anl.gov), Geoffrey Fox (gcf@indiana.edu), and Dan Reed (dan.reed@utah.edu).



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



Workflow environments in Fog-to-cloud infrastructures

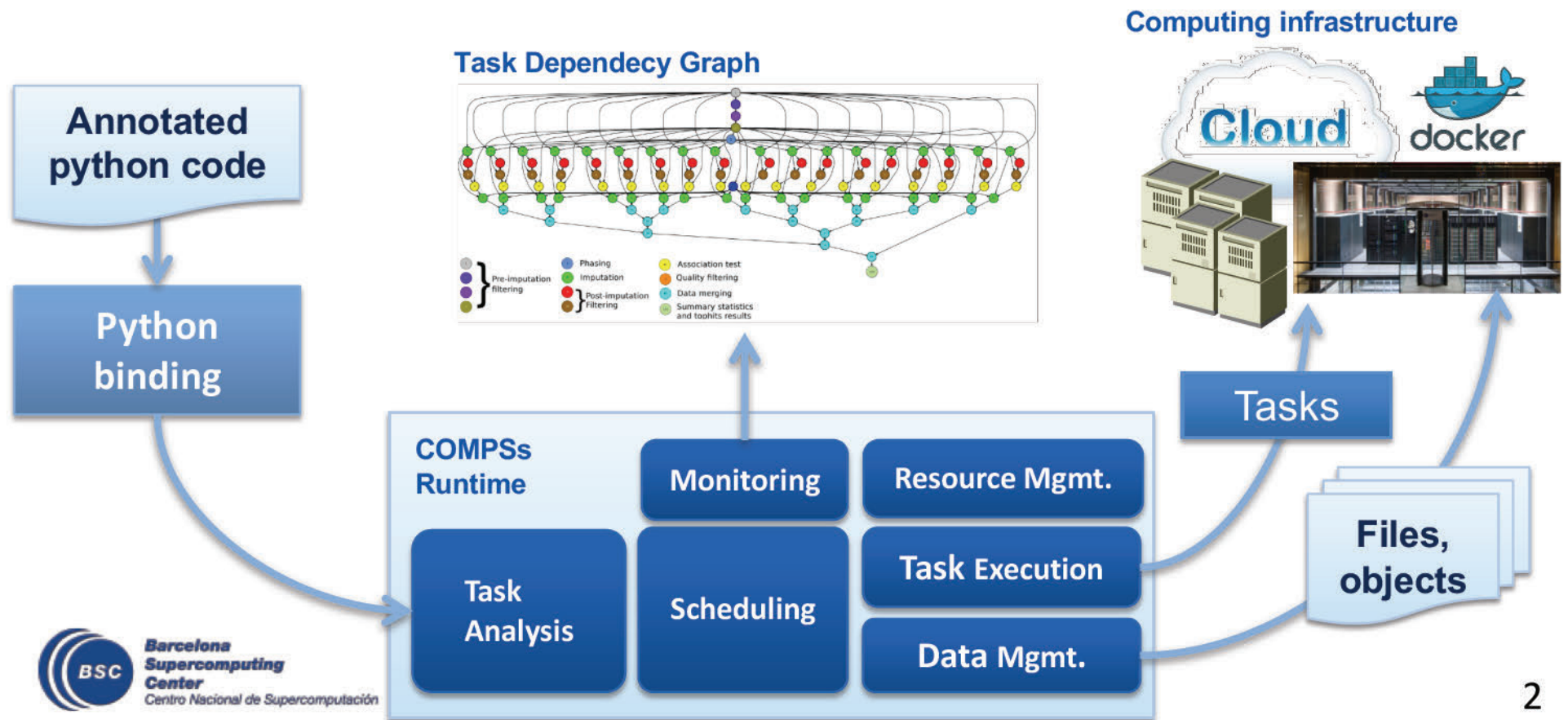
Rosa M Badia

19-21 Feb 2019

BDEC2, Kobe Japan

PyCOMPSs/COMPSs programming model

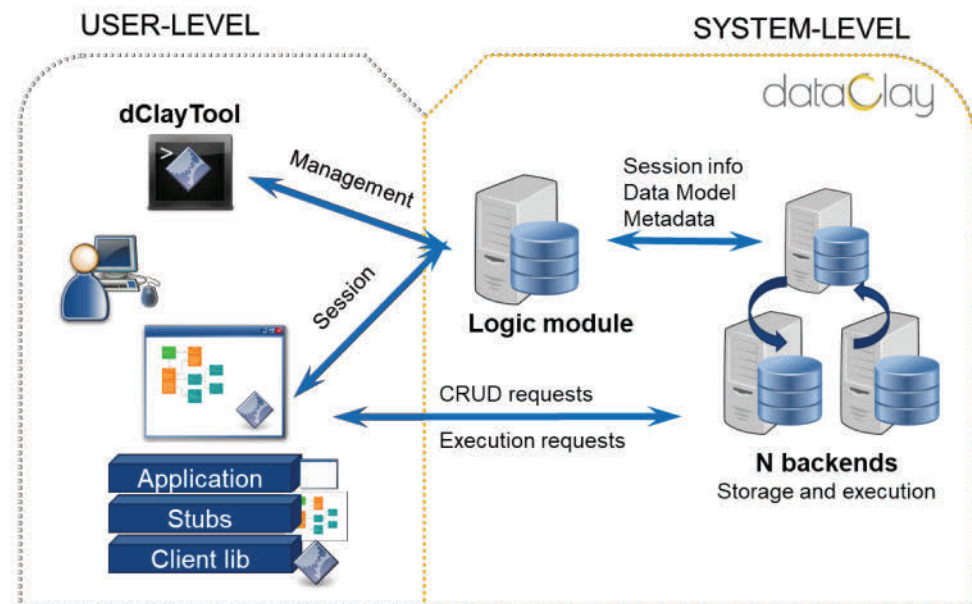
- Task-based programming model...
 - ... But also a workflow orchestrator system
- Tasks can be serial tasks, multi-threaded (OpenMP tasks), or parallel tasks (MPI, multi-node)
- Provides a programming environment for the convergence of HPC and HDA



dataClay platform

3

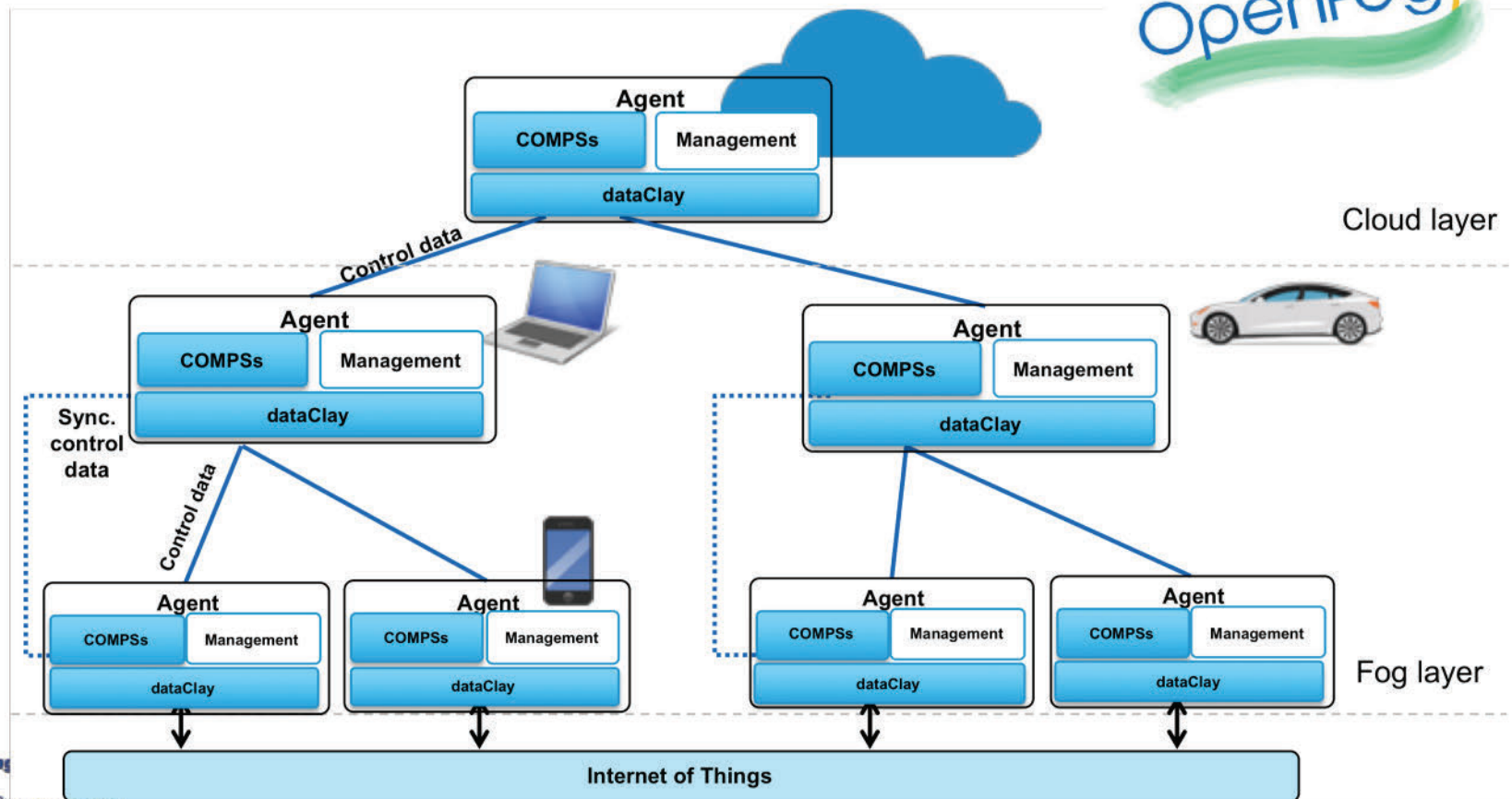
- Distributed storage platform based on objects (including methods)
 - Supports Java and Python applications
- A single data model to manage transparently:
 - Persistent and volatile data
 - Local and remote data
- Fully integrated with the OO programming model
 - Code is linked to data
 - Exploits data locality
- Integrated with PyCOMPSs/COMPSs through a storage interface



3

COMPSs in a fog-to-cloud architecture

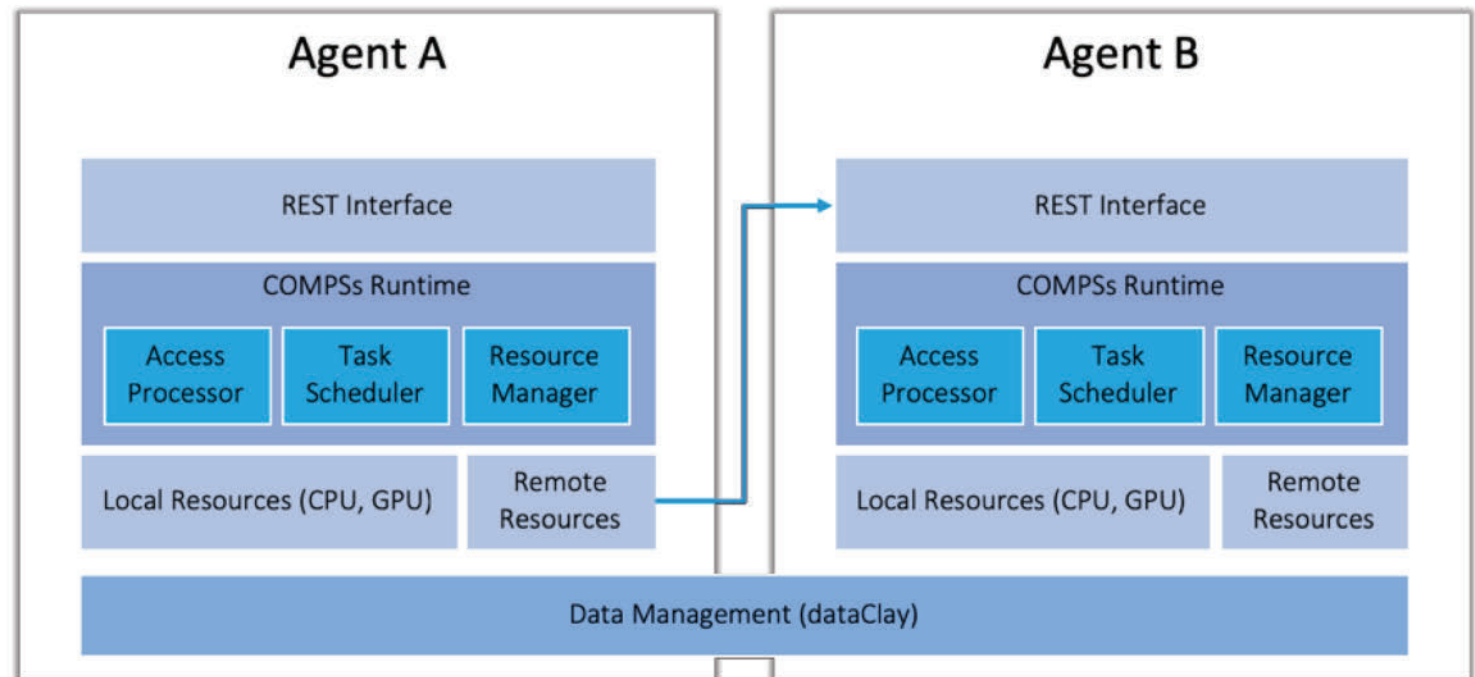
- Decentralized approach
- Lower layer: low processing, data generation
- Middle layer: fog devices, some processing, workflow orchestration fog-to-fog
- High layer: cloud, high processing, global control, fog-to-cloud



Agent-based approach



- Runtime deployed as a microservice in an agent
- Agents are independent, can act as master or worker in an application execution, agents interact between them
- Set of resources in the execution of an application is configurable
 - Can be local to an agent or remote in other agents
- dataClay provides data transparently regardless of its location
 - Including data recovery when fog nodes disappear
 - All data stored in dataClay
- Federation of dataClay instances
- COMPSs syntax unchanged



Objectives of research

- Programming interfaces:
 - Explore graphical or higher-level interfaces to describe the workflows
- How to better integrate the compute and data flows
 - Integrate metadata, enable data traceability
- Better integration with machine learning programming
 - Initial experimentation with TensorFlow and PyTorch
 - Development of dislib, a PyCOMPSs based machine learning distributed library
- Support for interactivity, steering
- Add more intelligence to the runtime
 - Using machine learning techniques
 - Taking into account performance aspects, resilience and energy efficiency
 - Modelling and metrics

Considering a Clean Slate

(or “Why Go Commando?”)

Micah Beck

University of Tennessee, Knoxville

Fundamental goal: Achieve portability across The Continuum

- How do we do it?
 - Enable extreme abstractness in specification (Not native code only)
 - Allow as many choices as possible within the common model (leases vs. indefinite duration)
 - Allow *very* weak assumptions (best effort)
 - Work at the lowest level possible (local)

What does a clean slate mean? Virtualization.

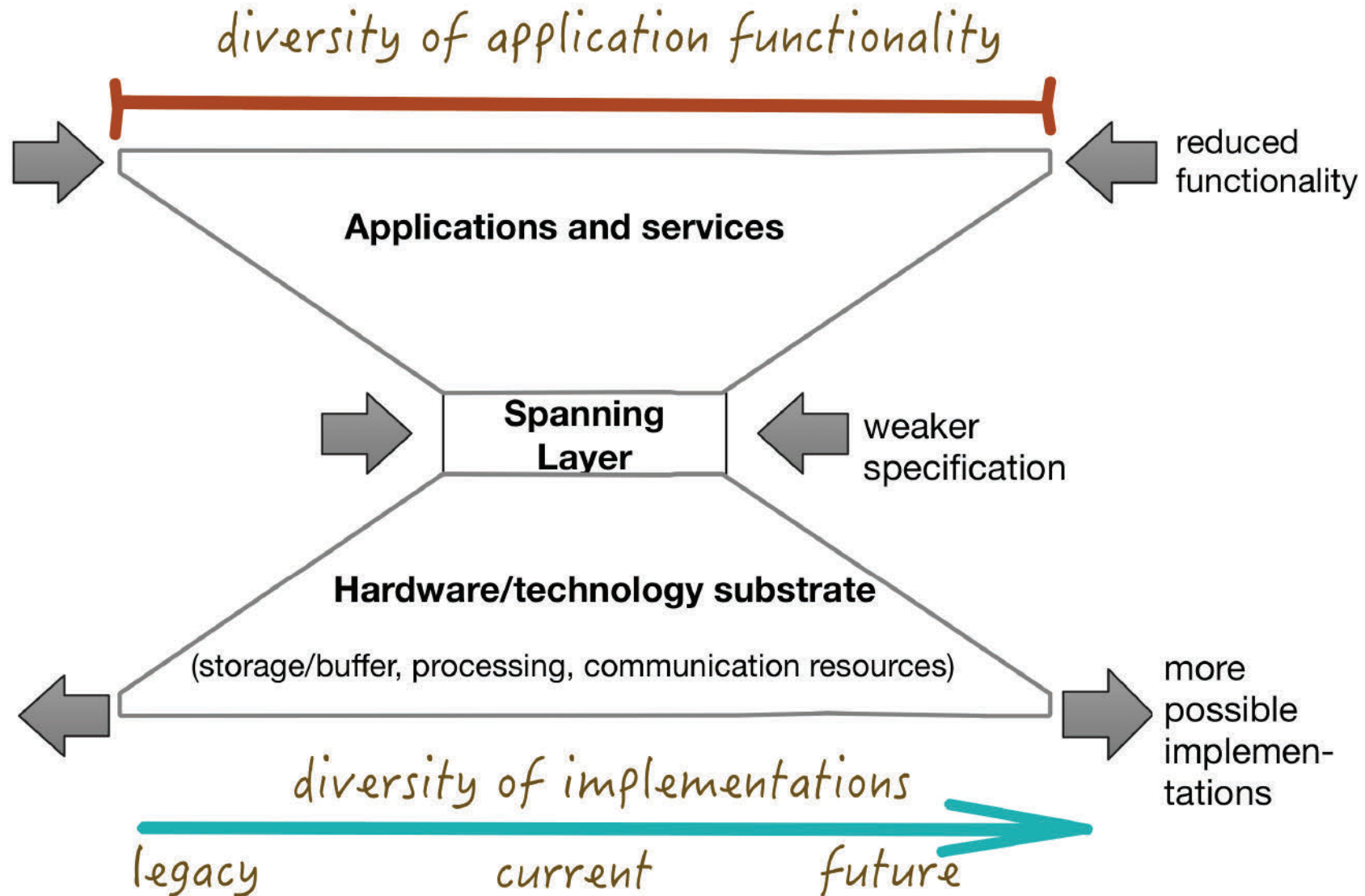
- Defining a service architecture that need not (but can) incorporate existing interfaces without modification. Customizable as needed.
- Not being bound by the conventions and rules of current shared infrastructure
- Freedom to design for maximum utility to scientific communities
- Examples of virtualization for commonality: POSIX, SRB, Java, PVM/MPI
- “Let’s not recreate what already exists” is ambiguous:
Which differences do we take into account in “already exists”?
It depends on what the meaning of “is the same as” is.
(with apologies to Bill Clinton)

Designing a common interface

- Model important functionality.
- Only include features needed by most of the community.
- Avoid vendor lock-in.
- Keep it simple, generic and limited!

- Put existing products (e.g., containers, AWS, etc.) under the hood (i.e., use as implementation tools)

The Hourglass Theorem



Toward an Advanced Cyberinfrastructure Platform

February, 2019

F. Bodin

EXDCI2 Scientific Director

<http://exdci.eu>

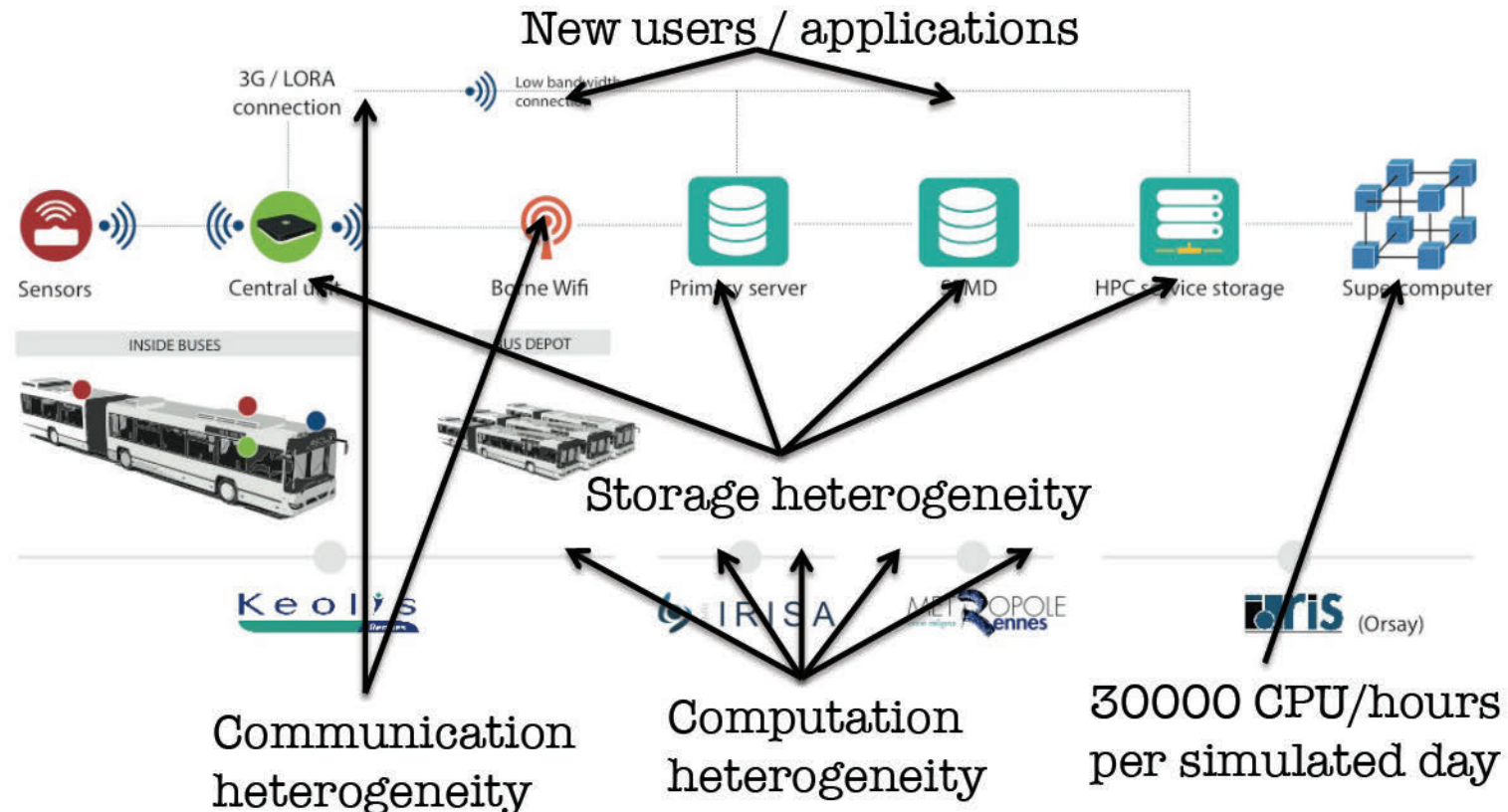
AQMO Coordinator

<http://aqmo.irisa.fr>

BDEC Related Background Activities

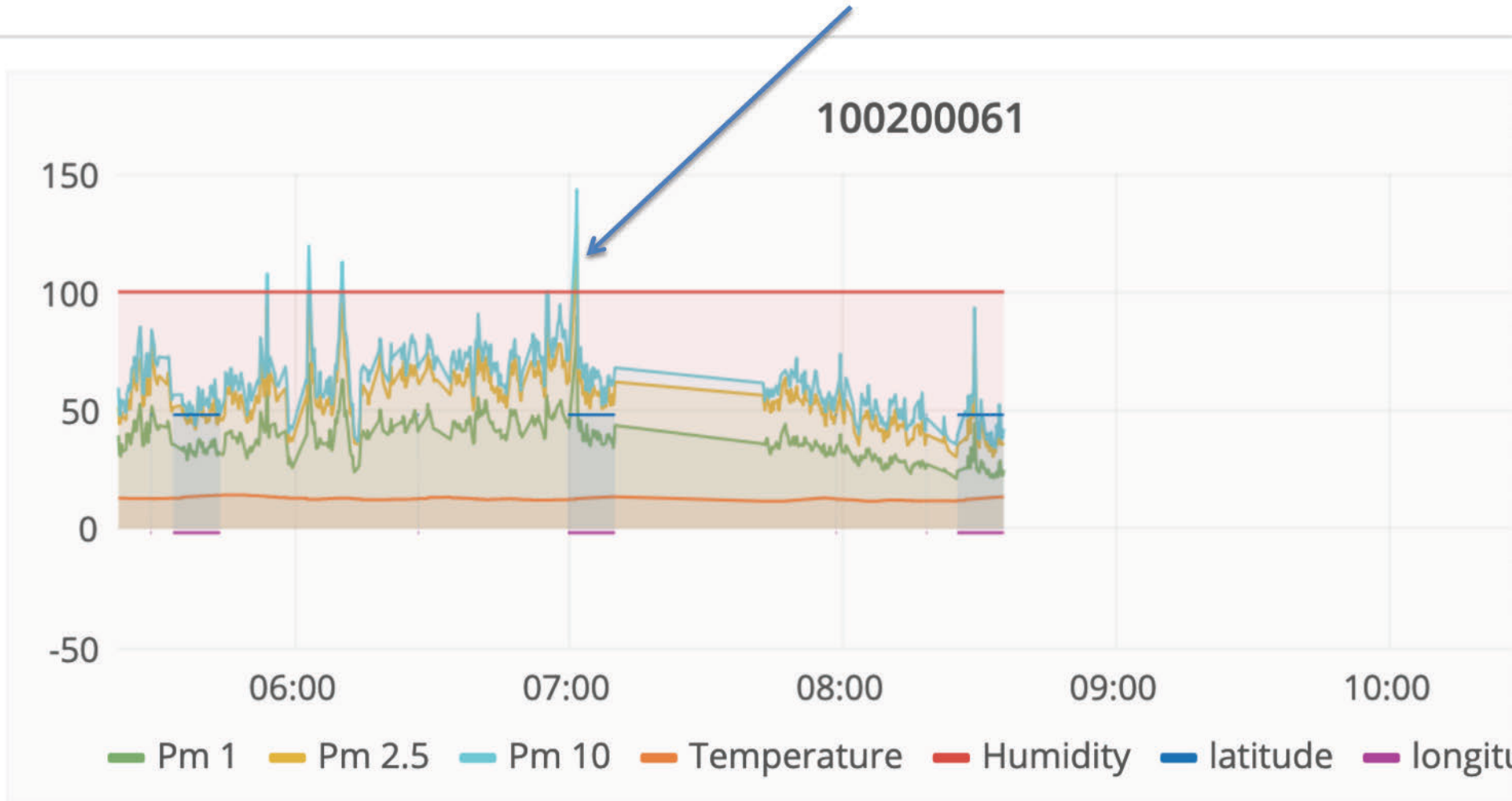


- EXDCI-2 Think Tank
- Air Quality & MObility CEF-Project



Sensor Data from Bus

Use of ML to get measurement context



AQMO Technical Background

- Massive heterogeneity in compute, network and storage resources
 - At the technical level
 - At the governance level multi-owner, multi-tenant
- Sensors, weather, topology and simulation data
 - Accumulation over time
 - Edge and HPC computing in the same workflow
- Multiple kinds of networks (LoRa, 4G, WLAN)
 - Some connectivity can be intermittent

What Have We Learn So Far?

- Federation of private and public infrastructures
 - Need to increase permeability of supercomputing systems to facilitate data injection
 - Currently too many one-to-one agreements
- Need a formulation of a vision for a future service-oriented architecture framework
 - For HPC compute and other compute services as well as storage and other data services
 - Basis for expressing applications as complex workflows
- Archiving data is under-estimated (*une patate chaude*)
 - 50K€ / petabytes / years (~ AWS)
- The main issue is political
 - Many infrastructures in silos cannot be federated without a strong political will (and funding)

How to Go Forward

- Requires a kind of *meta* governance with APIs for
 - Enrolling facilities
 - Monitoring & logging & debugging
 - Billing capacities
- Need convincing security & privacy solution
 - Needed for a future service-oriented architecture framework
- Common metadata and data management rules
 - What are the common meta-data ecosystem-wide?
 - One focus should be “when can we erase a data set?”
- High connectivity
 - To allow exploiting complementarities between infrastructures

Multi-Hybrid Accelerated Computing

GPU + FPGA = ?

Taisuke Boku

Deputy Director, HPC Group Leader
Center for Computational Sciences
University of Tsukuba

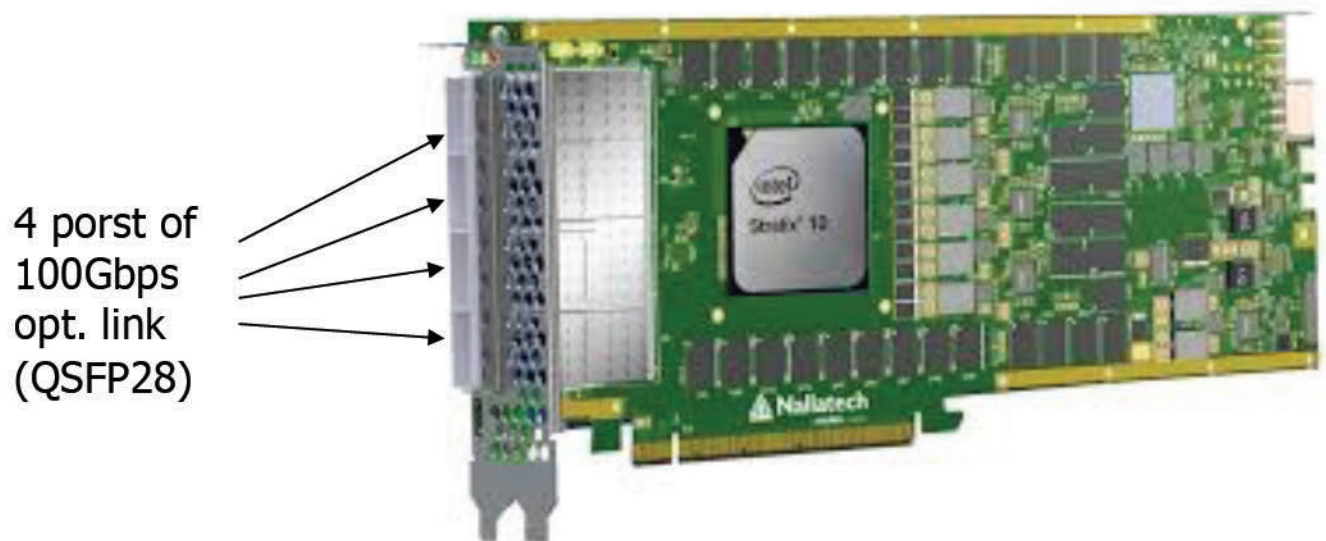
GPU + FPGA -> Compete or Collaborate ?

19

- GPU for
 - Up to 7P, 14P or 28P FLOPS for DP, SP or HP
 - Bulk computation with large degree of constant parallelism (big SIMD)
 - Non-frequent communication
 - Regular computation without exception
 - High bandwidth demand to memory
- FPGA for
 - 10PFLOPS for SP, but reconfigurable for any bit size
 - Low~medium spatial parallelism but high pipelined parallelism (gate level)
 - High performance direct network ~ 100Gbps x 4
 - Simultaneous computation on any branch, everything is pipelined
 - Medium bandwidth on memory -> HBM2 soon
- Multi-physics, partially serial (to be bottleneck) apps require both

Large scale logic elements and high speed ext. link on FPGA

20

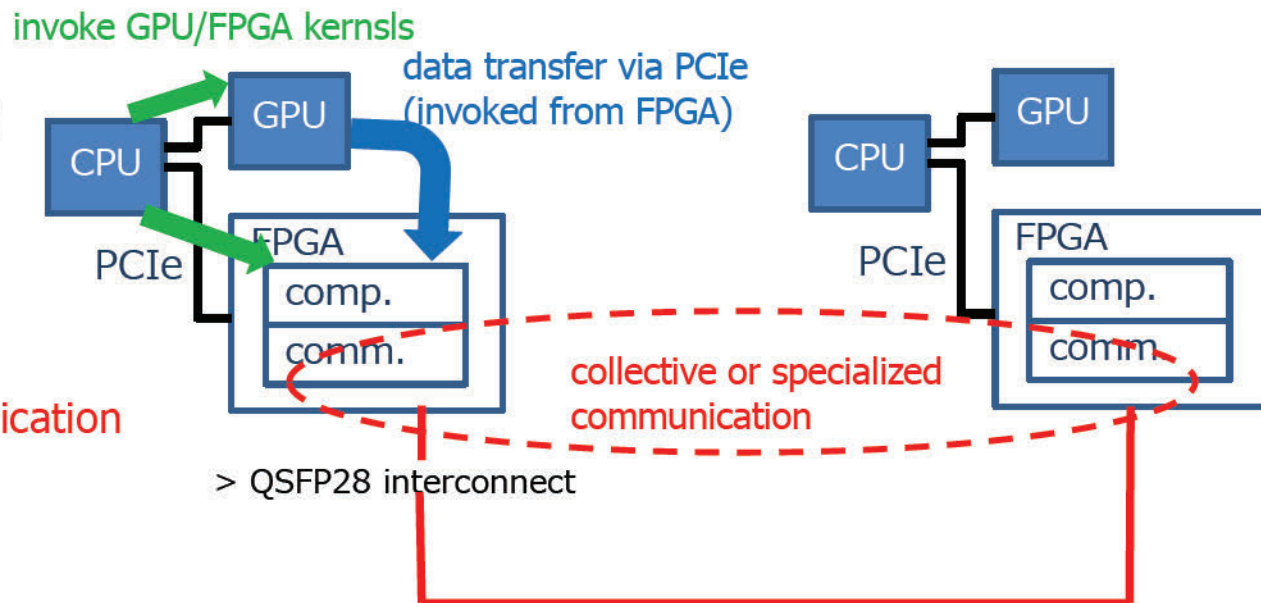


4 ports of
100Gbps
opt. link
(QSFP28)

Nallatech 520N with Intel Stratix10 (H-Tile)
with 2M LEs, 240Mb memory + 16 Gbyte DDR memory

AiS: Accelerator in Switch

- FPGA can work both for computation and communication in unified manner
- GPU/CPU can request application-specific communication to FPGA



OpenCL-ready modules to support apps.

CoE: Channel over Ethernet

OpenCL-ready FPGA direct link driver

sender code on FPGA1

```

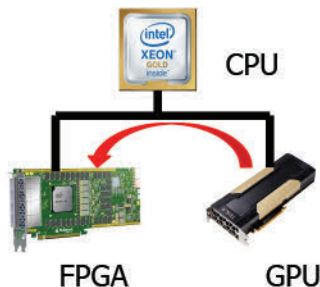
_kernel void sender(__global float* restrict x, int n) {
    for (int i = 0; i < n; i++) {
        float v = x[i];
        write_channel_intel(network_out, v);
    }
}
    
```

receiver code on FPGA2

```

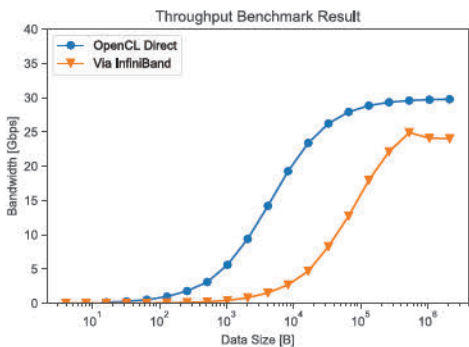
_kernel void receiver(__global float* restrict x, int n) {
    for (int i = 0; i < n; i++) {
        float v = read_channel_intel(network_in);
        x[i] = v;
    }
}
    
```

OpenCL-ready GPU-FPGA DMA on PCIe

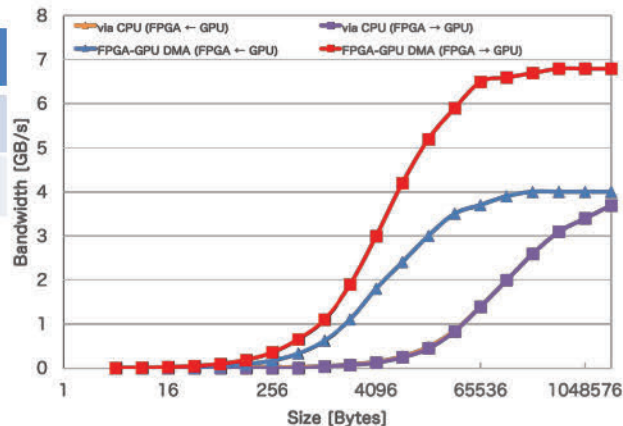


```

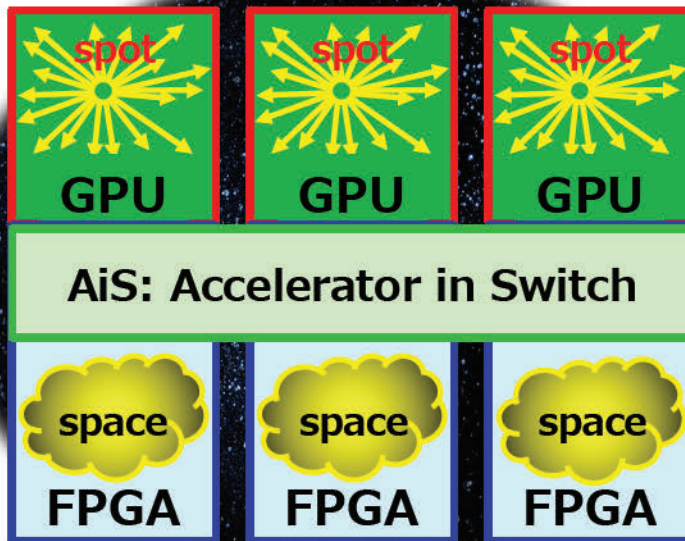
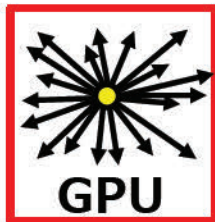
_kernel void fpga_dma(__global float *restrict fpga_mem,
                    const ulong gpu_memadr,
                    const uint id_and_len)
{
    cldesc_t desc;
    // DMA transfer GPU -> FPGA
    desc.src = gpu_memadr;
    desc.dst = (ulong>(&fpga_mem[0]));
    desc.id_and_len = id_and_len;
    write_channel_intel(fpga_dma, desc);
    ulong status = read_channel_intel(dma_stat);
}
    
```



direction	via CPU	FPGA-GPU DMA
GPU→FPGA	17	1.44
FPGA→GPU	20	0.60



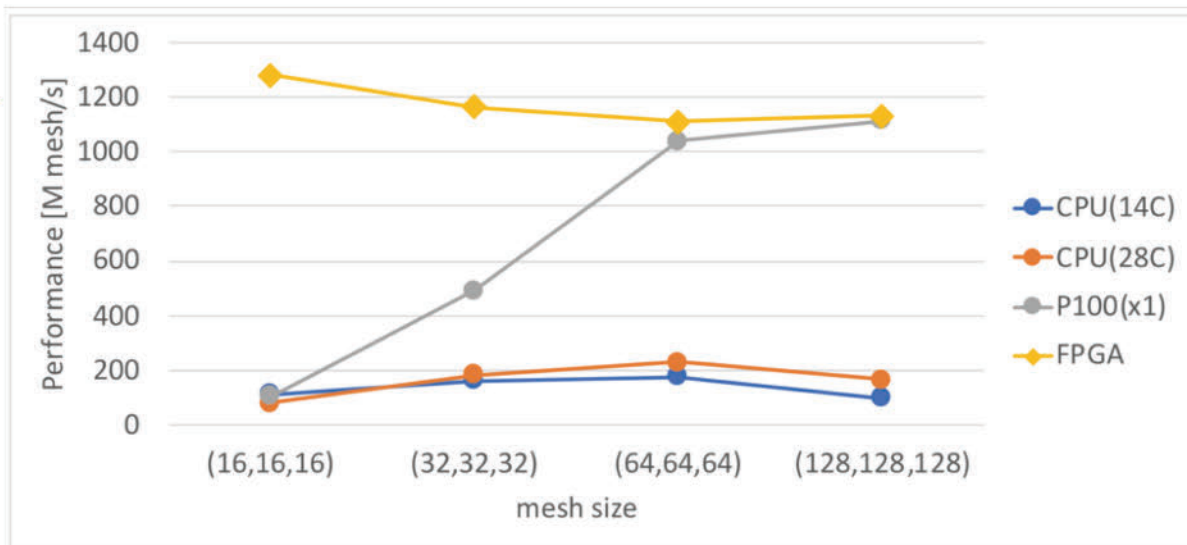
Radiation from spot
light source (ARGOT
method)



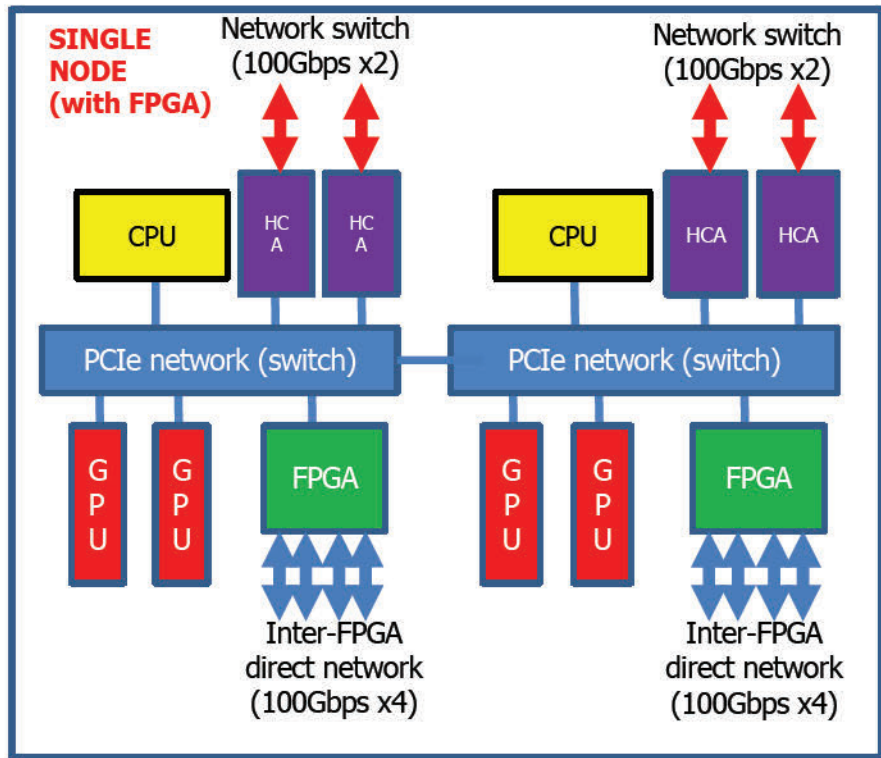
Radiation from
spatially distributed
light source (ART
method)



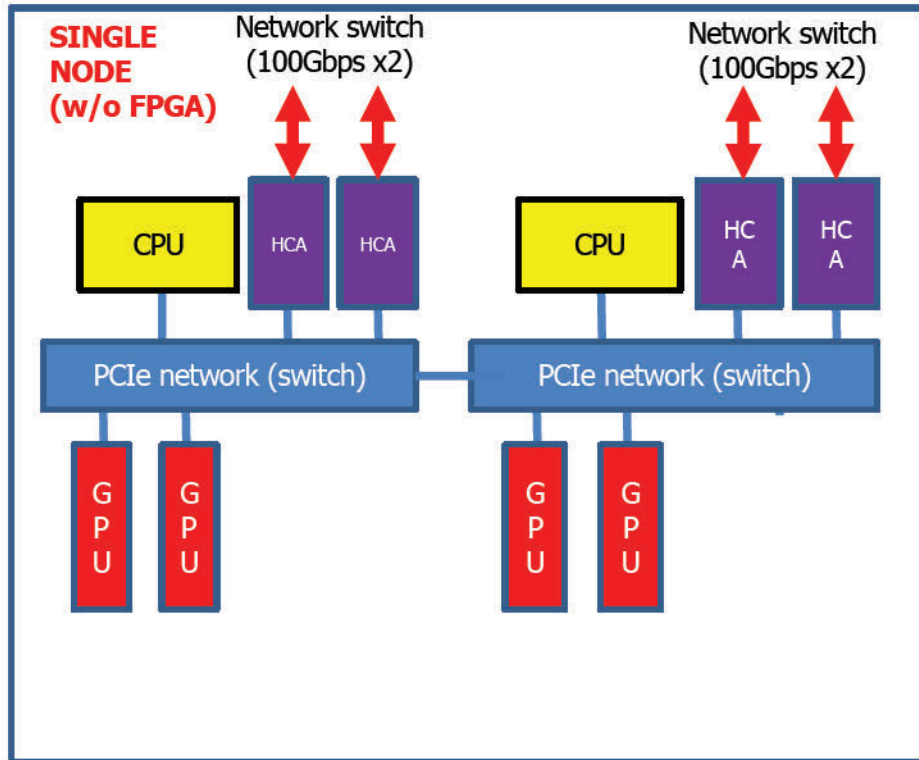
Mesh calculation speed in ART method (early universe simulation)



Block diagram of Albireo and Deneb node



Albireo (w/ FPGA)

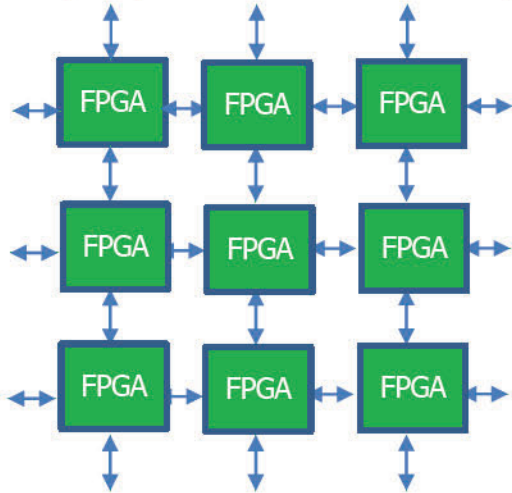


Deneb (w/o FPGA)

Two types of interconnection network

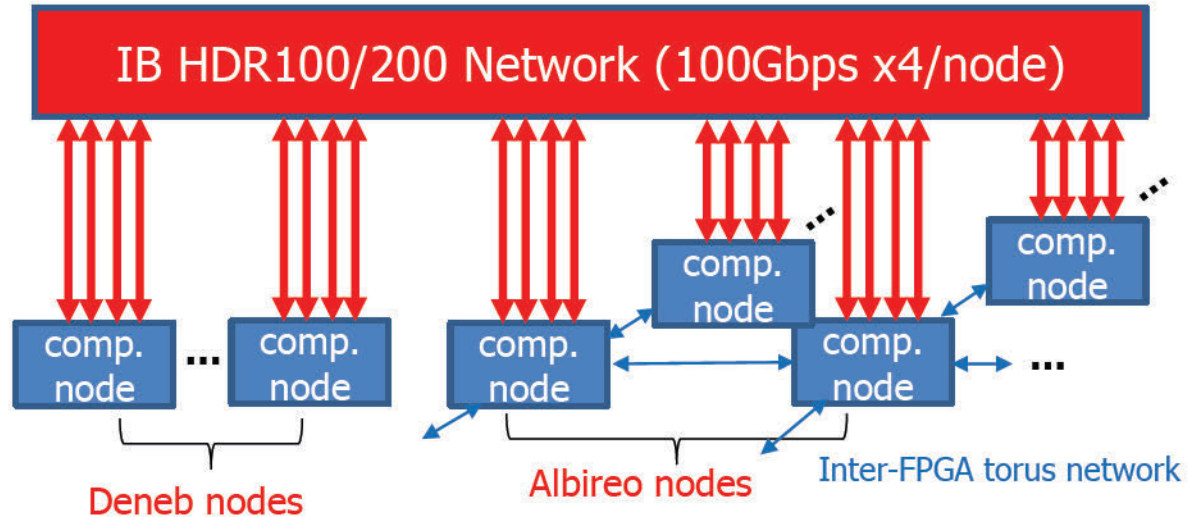
26

Inter-FPGA direct network
(only for Albireo nodes)



64 of FPGAs on Albireo nodes (2 FPGAs/node) are connected by 8x8 2D torus network without switch

InfiniBand HDR100/200 network for parallel processing communication and shared file system access from all nodes

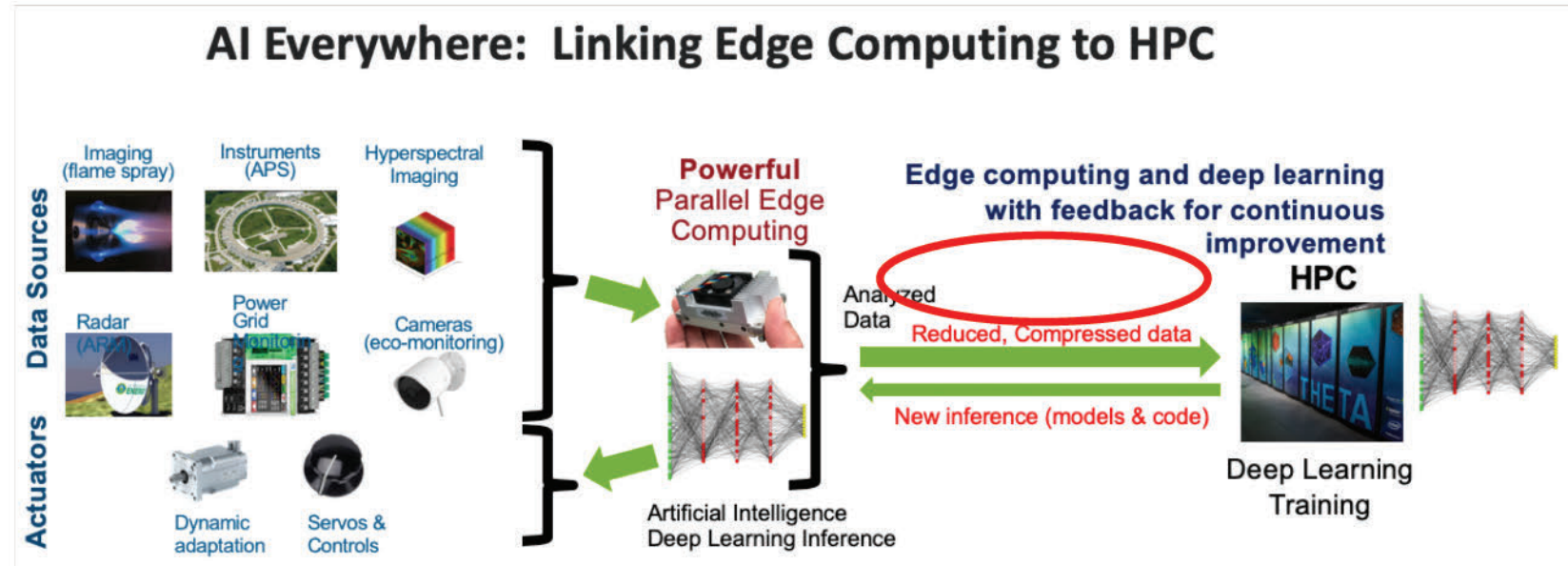


For all computation nodes (Albireo and Deneb) are connected by full-bisection Fat Tree network with 4 channels of InfiniBand HDR100 (combined to HDR200 switch) for parallel processing communication such as MPI, and also used to access to Lustre shared file system.

- Multi-physics simulation
 - Astrophysics
 - Climate (high density LES in City level)
 - Particle physics (low latency collective comm.)
 - ...
- AI
 - GPU (of course) + FPGA
 - ...
- Collaboration
 - OpenACC for GPU and FPGA: with ORNL (J. Vetter)
 - FPGA for HPC: with ANL (F. Cappello)
 - FPGA direct-link and router: R-CCS (K. Sano)
 - GPU/FPGA cloud simulation: AIST/U-Tokyo (T. Kudoh)

Lossy compression for a new, shared, advanced cyberinfrastructure platform (ACP)

*Pete's diagram
of ACP:*



+ many ECP Applications including EXAFEL: $\sim x10$ compression with minimal distortion

We will need to deal with lossy scientific data:

Good news: we start to have a good understanding of how to perform lossy compression

The other good news for CS researchers: a lot of research is needed.

Compression is at the core of Edge/Cloud computing

Between edge and Cloud:

- Push and Pull modes

PULL mode (mostly)

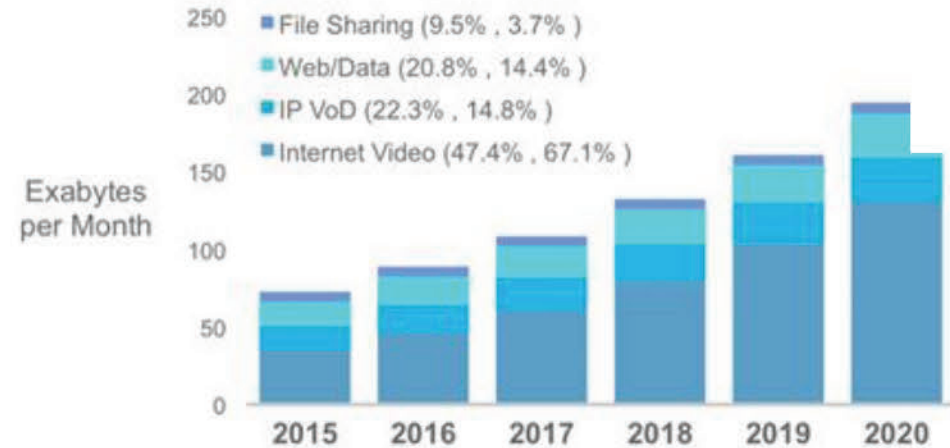
CISCO statistics:

- Annual global IP traffic: 2.3 ZB (10^{21}) in 2020
 - IP video traffic will be 82 percent of all IP traffic
- >>80% of the IP traffic will be lossy compressed!

PUSH mode (mostly)

Facebook data:

- Stores more than 240 billion photos (all of them is compressed)
- Users uploading 350 million new photos every single day
- Data center deploys 7 petabytes of storage every month.



Lossy compression for music, video, photos, images

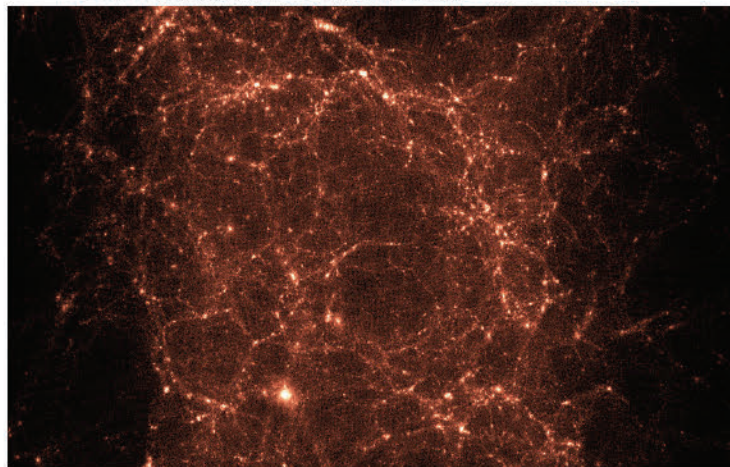
- is very well understood
- We know how to compress (image, sound properties)
- We know very well Human perception → error tolerance
- Error tolerance valid for all of us

Not as simple in scientific domains

- Each dataset, simulation, instruments, sensor network is different
- We need different lossy compression algorithms
- Not trivial at all to understand the effect of distortion from lossy compression
- We need compression error assessment methods, tools and metrics to understand the nature of compression error (Trust)
- All applications (that we know) ask to respect user set point wise error bounds

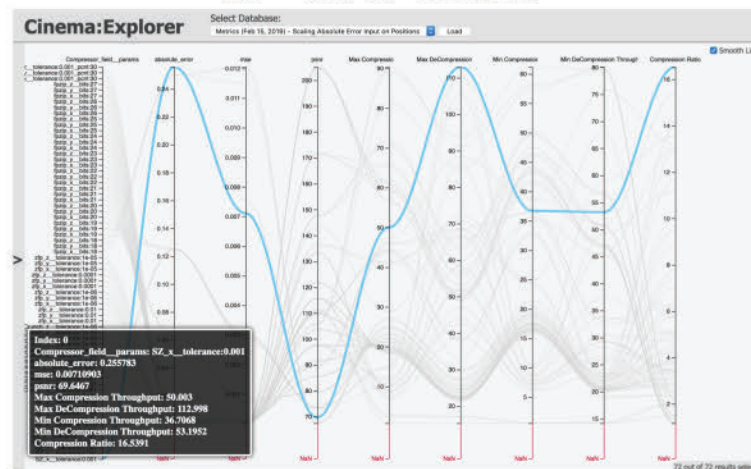


Cosmology simulation
Trillions of Particles



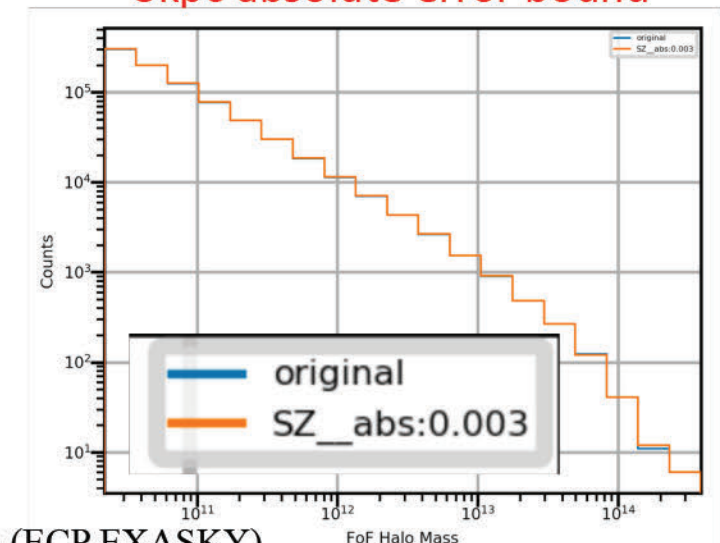
Figures from HACC team

SZ 2.0: CR ~10 (~3bits/value) at 10^{-3} error bound



Figures from Cbench (ECP EXASKY)

Results validation
3kpc absolute error bound



FoF Halo Mass

Example of compression workflow

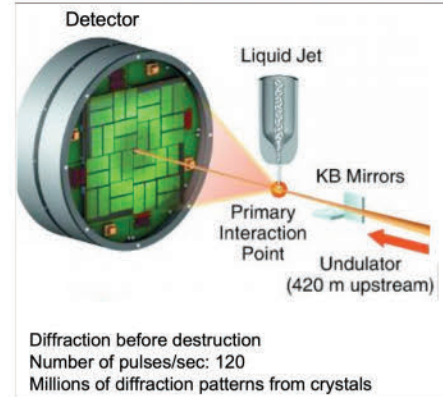
ECP EXAFEL: Context of LCLS-II

- Instrument produces 2D images:
- 1 LCLS-II area detector: 250GB/s
- With today technology: ~x1000s disks.
- Data is unsigned integers (RAW, Calibrated), in HDF5 format
- Goal: **CR of 10 with error bound**

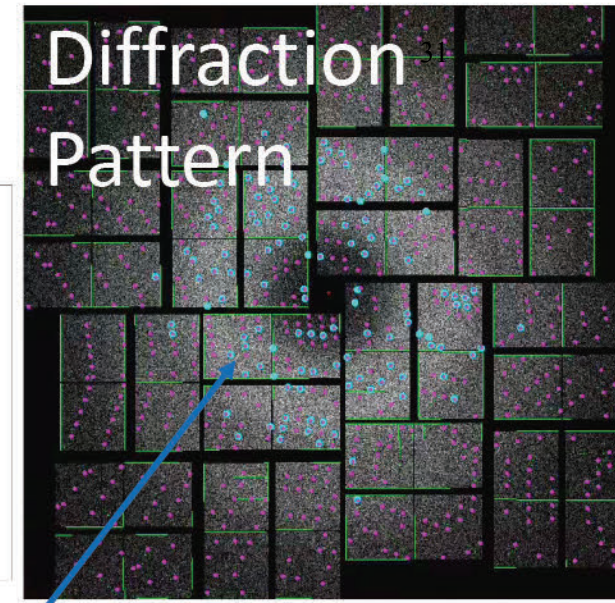
→ True Co-Design (algorithm, hardware)



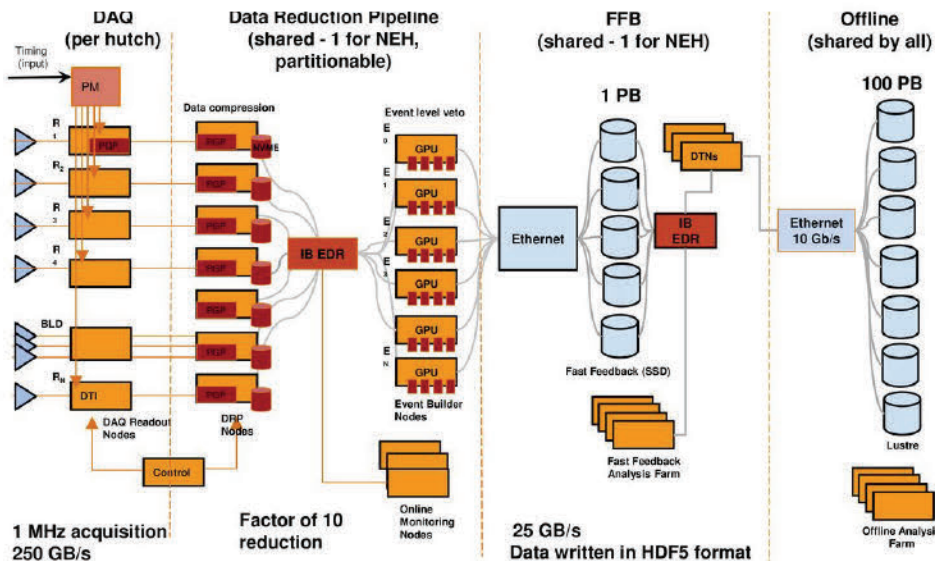
T.O. Raubenheimer for the LCLS-II Collaboration, SLAC, Menlo Park, CA 94025, USA, 2015, JACoW Publishing



Diffraction before destruction
Number of pulses/sec: 120
Millions of diffraction patterns from crystals



This "ring" is the water drop



JPEG not good: does not respect point wise error bound

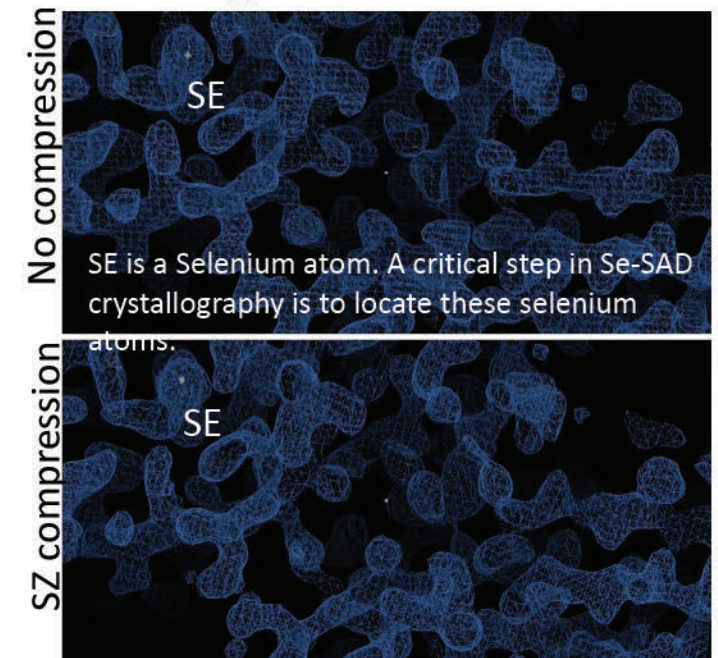
Compression with SZ 2.0:

- Ratio: 4
- Speed 120 MB/s/core

Needed:

- Ratio: 10
- Speed 500 MB/s/core

Quality/error assessment



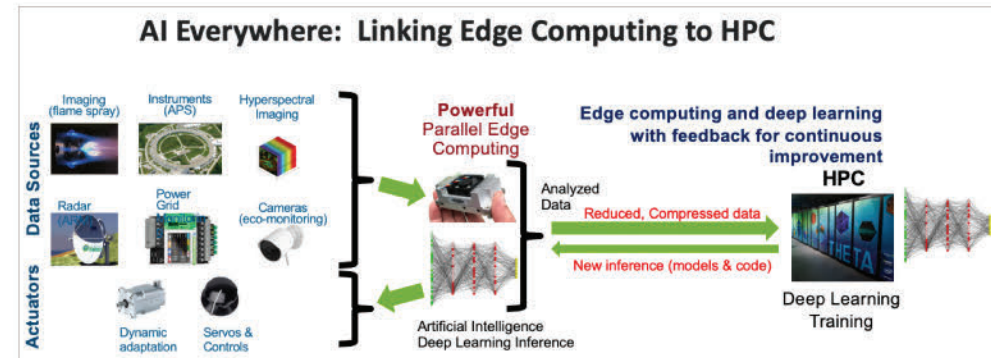
Why does it matter for advanced cyberinfrastructure platform? ³²

Use cases from 6 ECP applications and more (HACC, LCLS2, GAMESS, NWCHEM, EXAALT, Urban):

- Reduction of storage footprint
- Reduction of I/O, communication time/energy
- Faster execution state preservation/copy
- Reduction of memory footprint (run larger problems)
- Computation acceleration though re-computation avoiding
- Reducing streaming intensity (instruments)

Research questions:

- Lossy compression when AI in the loop?
- AI for lossy compression?
- Lossy compression algorithms for different apps
- Faster lossy compression/decompression algorithms
- Lossy compression assessment tools
- Multilevel/progressive decompression: e.g. for feature search
- More...



Scientific Data Reduction Benchmarks (SDRBench)

Scope and objectives

- A community repository providing reference scientific datasets, compressors (lossless and lossy), and error analysis tools
- Significance: Improve the methodology in the domain by providing reference information for scientific data compressor users and developers

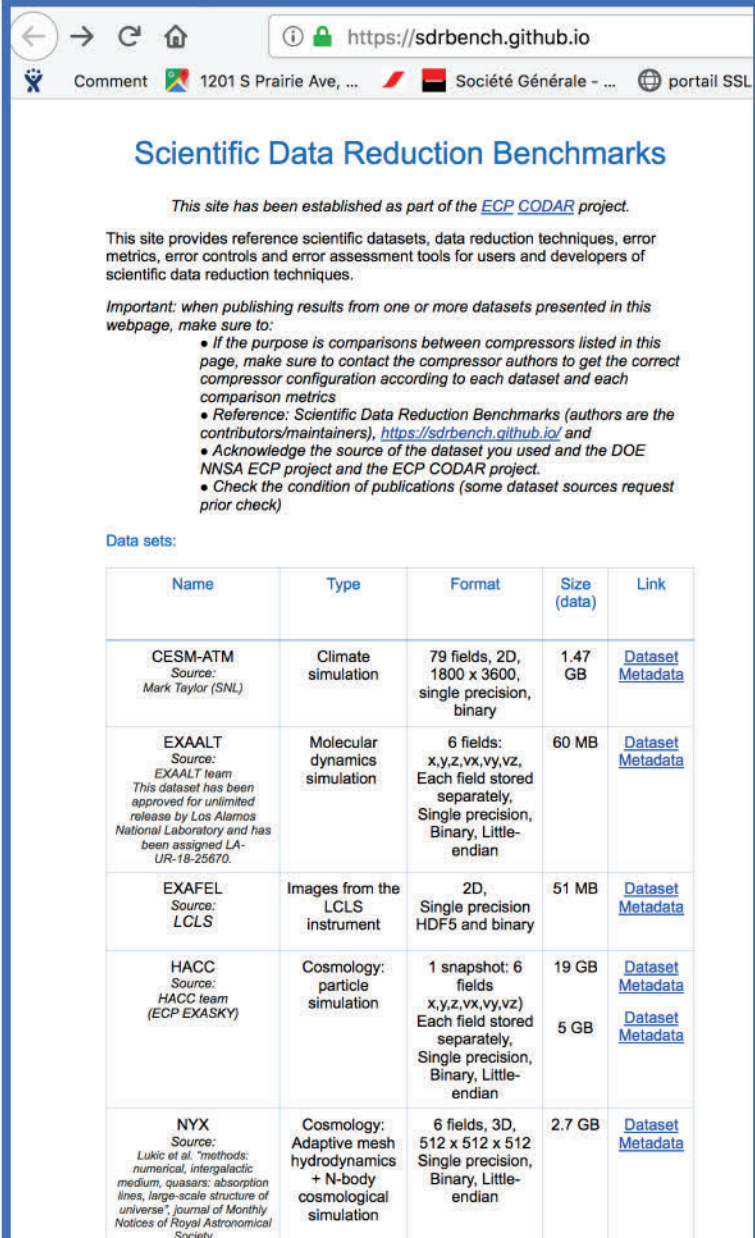
Impact

Opened on July 1st 2018.
reference source of information
main lossy and lossless compressors

Google: SDRBench

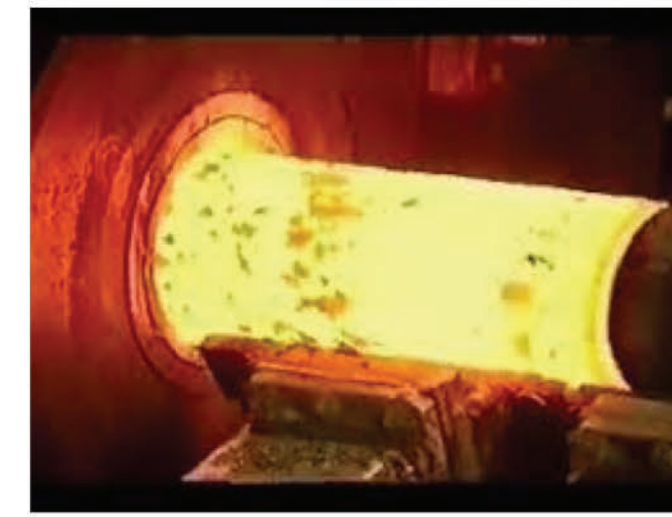
Project accomplishment

- Collection of representative datasets from ECP and other applications via direct communication with application developers and users
- Storage of the datasets on the Petrel server at Argonne with Terabytes of storage capacity
- Fast access to the datasets using Globus and GridFTP
- Access open to public



The screenshot shows the website <https://sdrbench.github.io>. The page title is "Scientific Data Reduction Benchmarks". Below the title, there is a paragraph stating: "This site has been established as part of the [ECP CODAR](#) project." Another paragraph explains: "This site provides reference scientific datasets, data reduction techniques, error metrics, error controls and error assessment tools for users and developers of scientific data reduction techniques." An important note follows: "Important: when publishing results from one or more datasets presented in this webpage, make sure to:" followed by a bulleted list of instructions for users. Below this is a section titled "Data sets:" which contains a table with columns for Name, Type, Format, Size (data), and Link. The table lists five datasets: CESM-ATM, EXAALT, EXAFEL, HACC, and NYX, each with detailed source information, simulation type, format, and size.

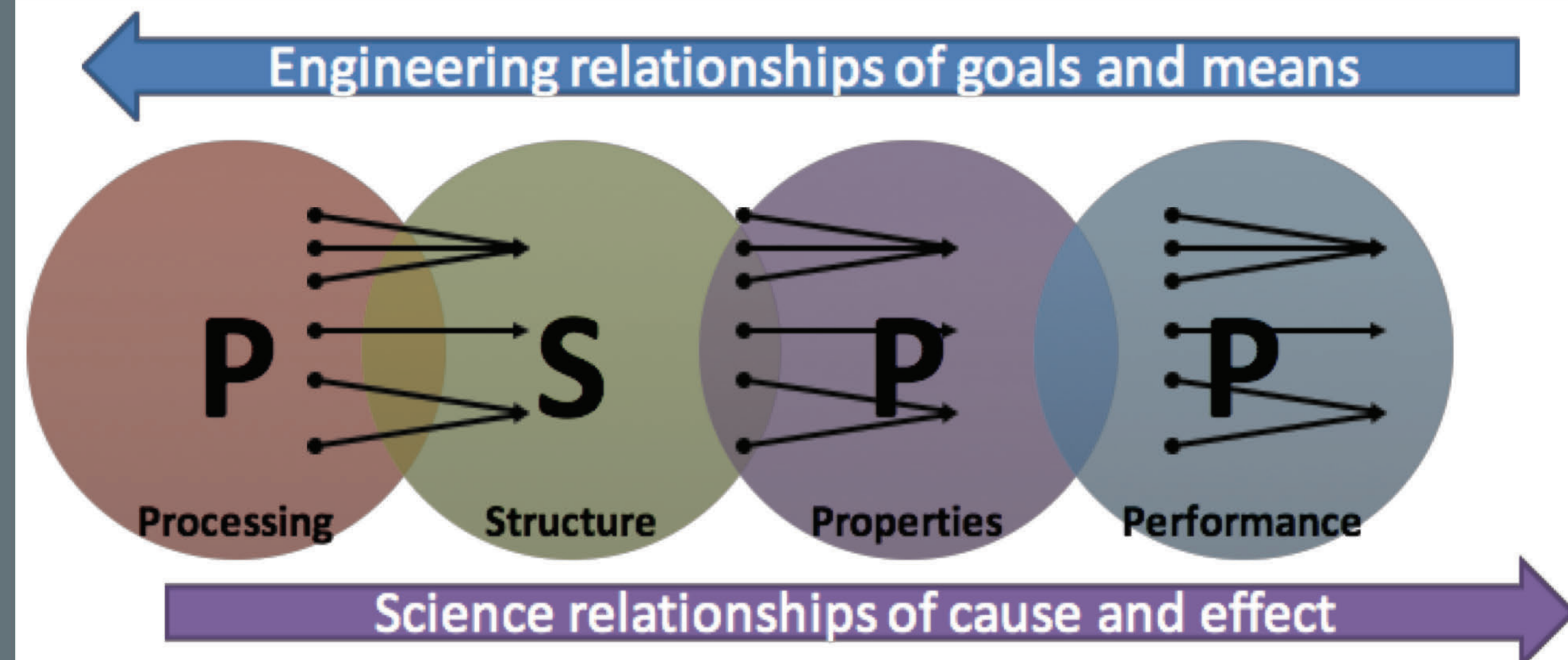
Name	Type	Format	Size (data)	Link
CESM-ATM Source: Mark Taylor (SNL)	Climate simulation	79 fields, 2D, 1800 x 3600, single precision, binary	1.47 GB	Dataset Metadata
EXAALT Source: EXAALT team This dataset has been approved for unlimited release by Los Alamos National Laboratory and has been assigned LA-UR-18-25670.	Molecular dynamics simulation	6 fields: x,y,z,vx,vy,vz, Each field stored separately, Single precision, Binary, Little-endian	60 MB	Dataset Metadata
EXAFEL Source: LCLS	Images from the LCLS instrument	2D, Single precision HDF5 and binary	51 MB	Dataset Metadata
HACC Source: HACC team (ECP EXASKY)	Cosmology: particle simulation	1 snapshot: 6 fields x,y,z,vx,vy,vz Each field stored separately, Single precision, Binary, Little-endian	19 GB 5 GB	Dataset Metadata Dataset Metadata
NYX Source: Lukic et al. "methods: numerical, intergalactic medium, quasars: absorption lines, large-scale structure of universe", Journal of Monthly Notices of Royal Astronomical Society	Cosmology: Adaptive mesh hydrodynamics + N-body cosmological simulation	6 fields, 3D, 512 x 512 x 512 Single precision, Binary, Little-endian	2.7 GB	Dataset Metadata



Scanning Electron Microscope

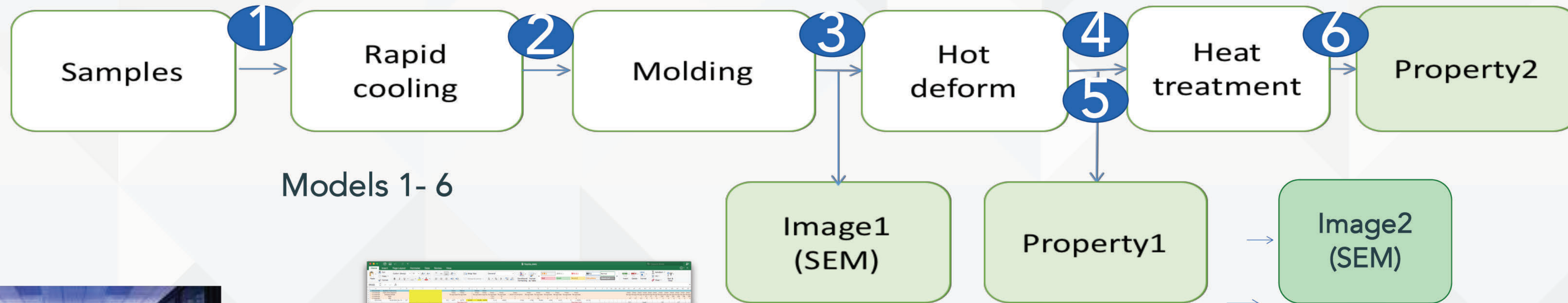
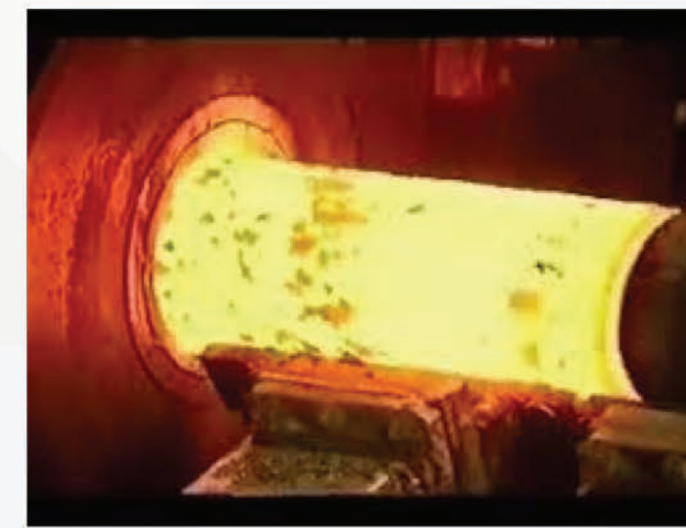
Industrial Materials Design – An Exemplar for BDEC

Dr. Alok Choudhary
Henry and Isabel Dever Professor
EECS and Kellogg School of Management
Northwestern University
choudhar@eecs.northwestern.edu

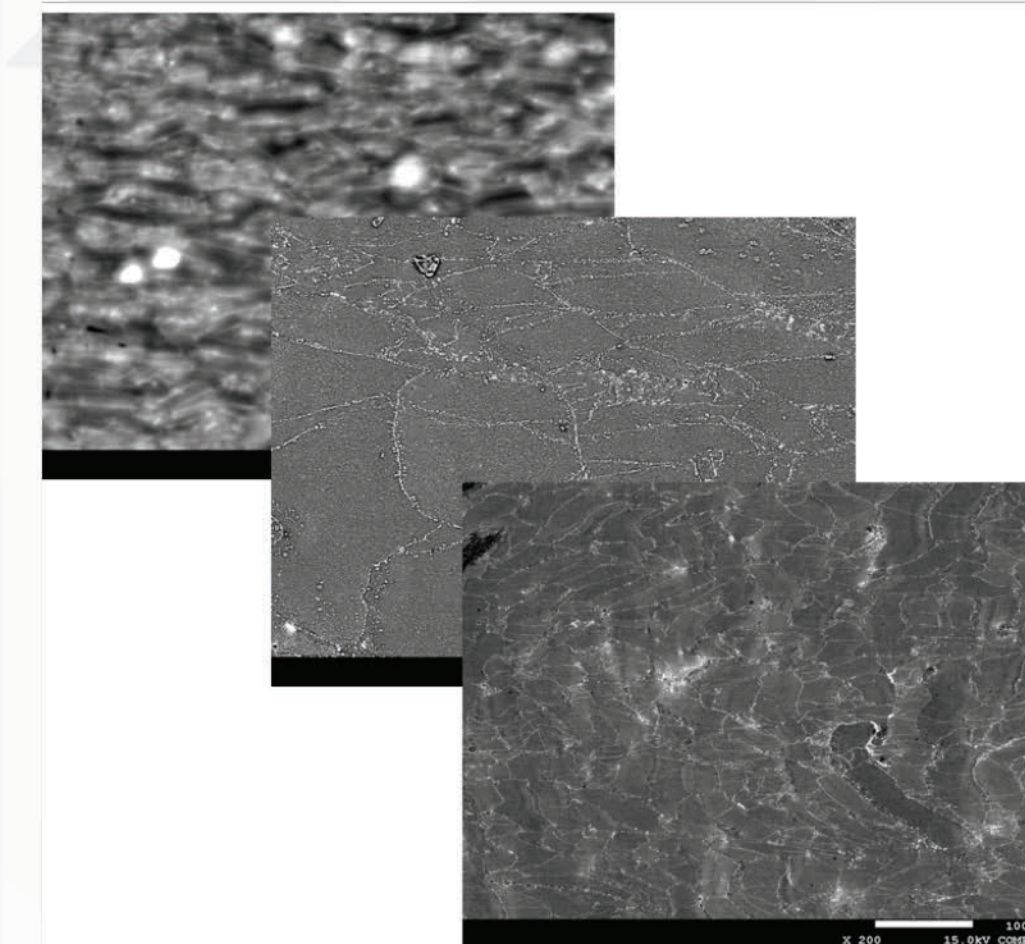
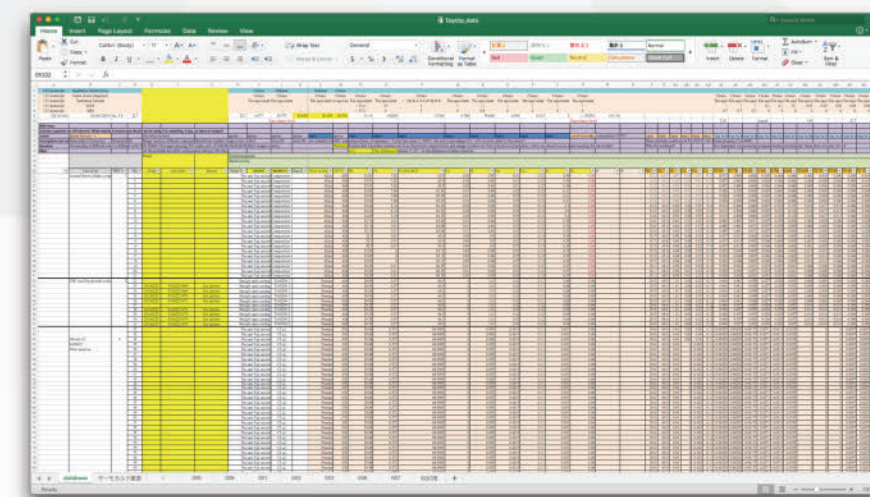


A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science, APL Materials, 4, 053208 (2016), <https://doi.org/10.1063/1.4941204>

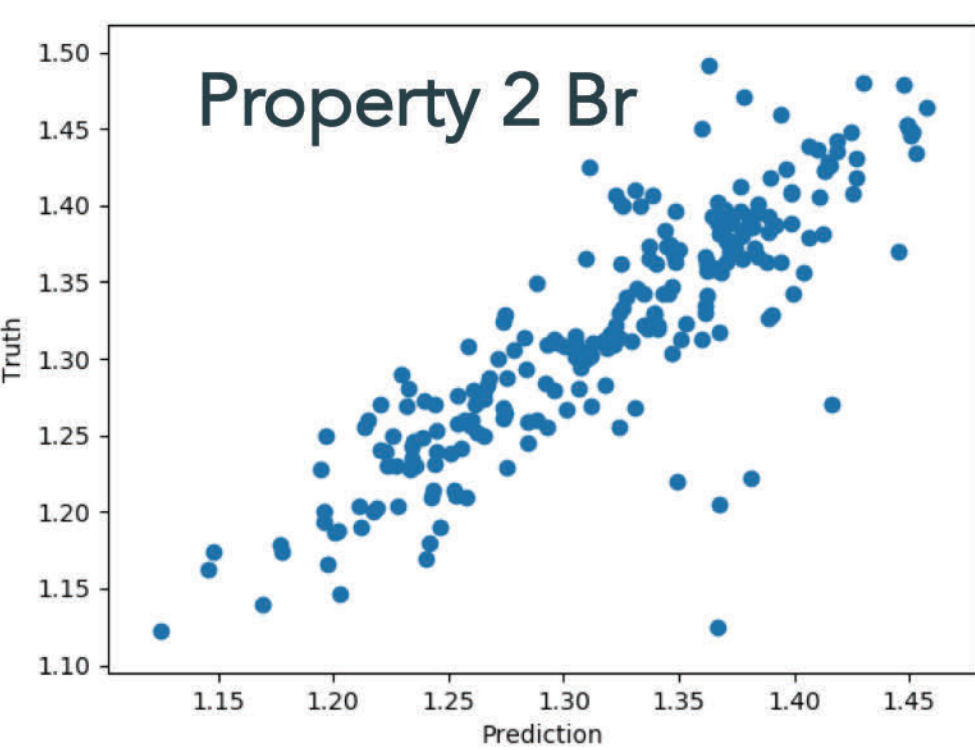
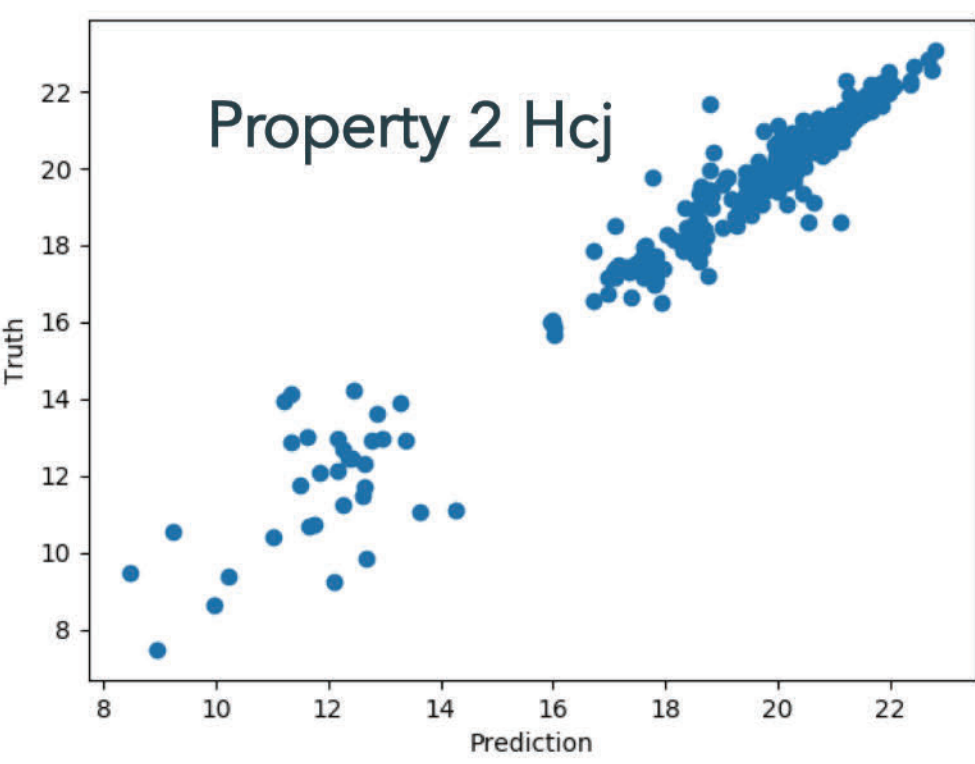
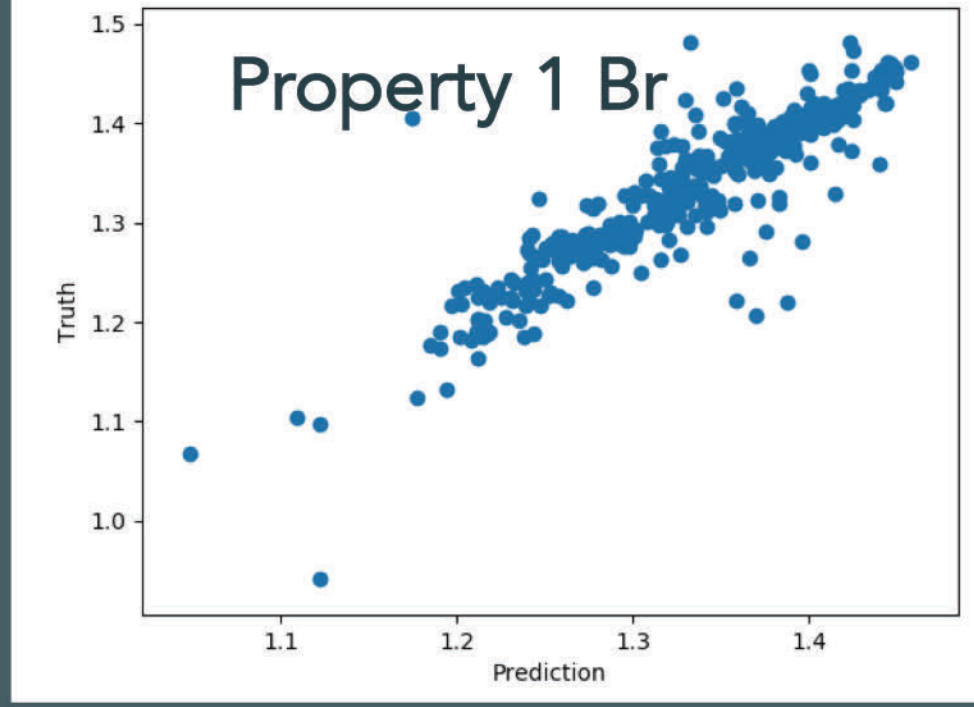
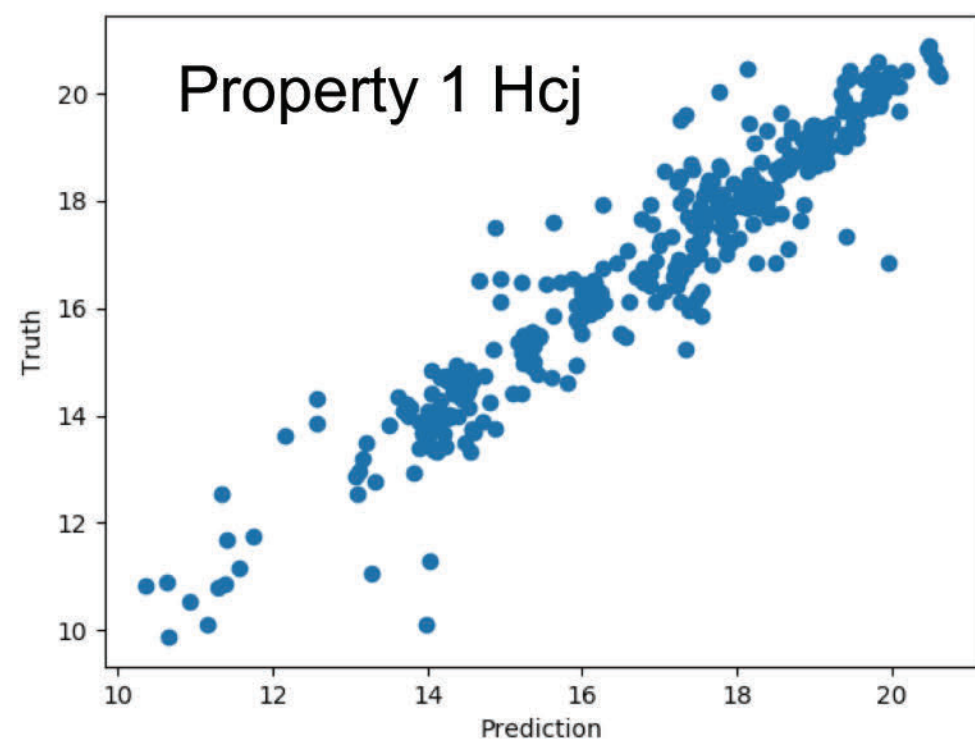
Problem definition



Models 1- 6



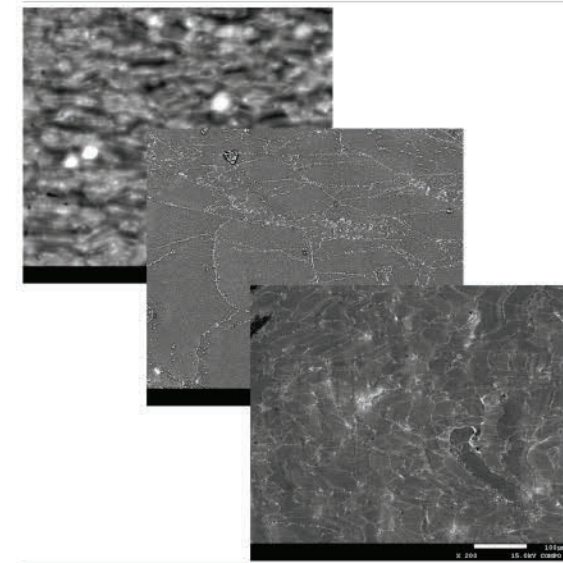
➤ Prediction of properties of samples under certain processing conditions



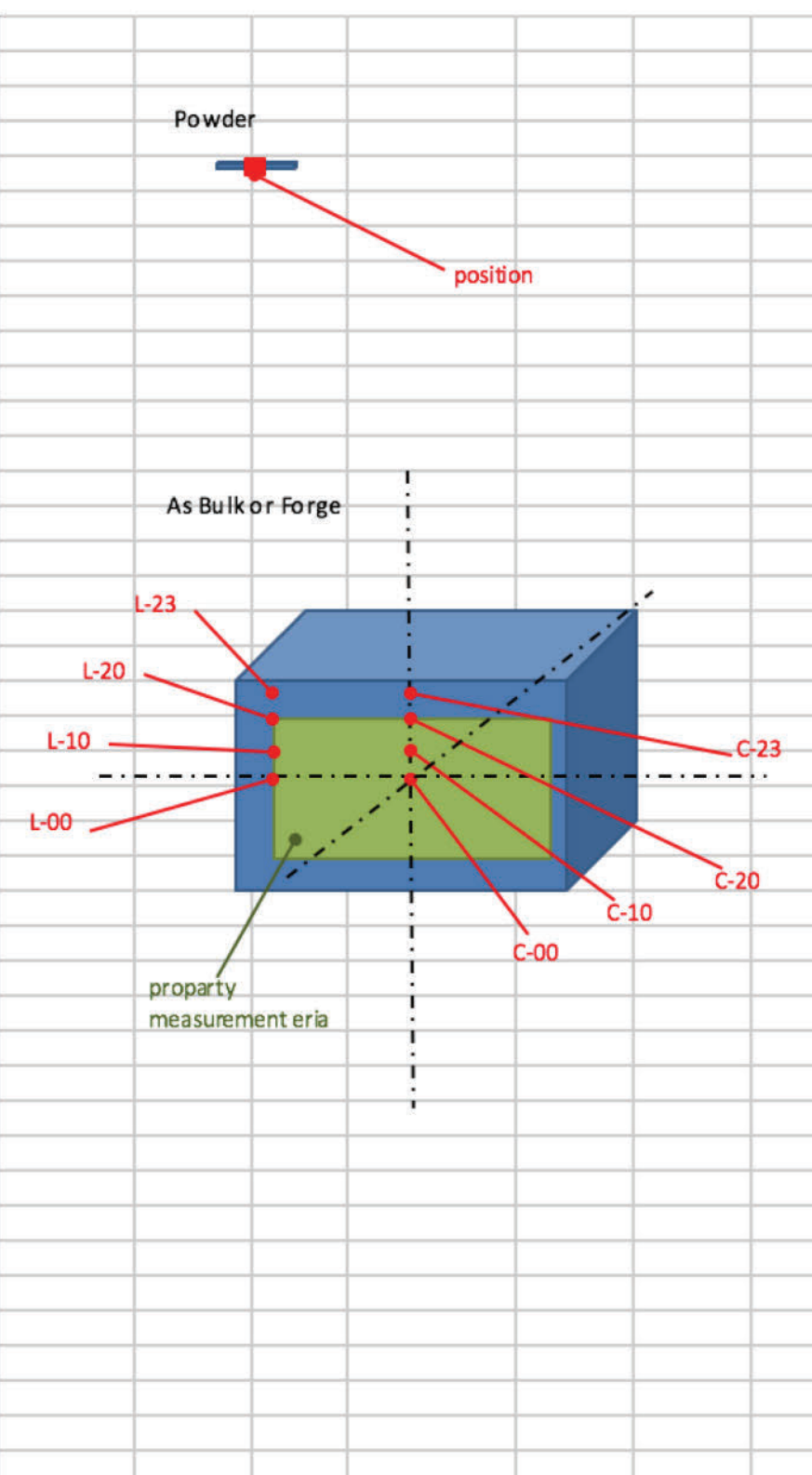
Some Level of Prediction with Process (numerical) data

- Two image modes: COMPO, SEI
- Two targets: powder, as bulk or forge
- eight positions: C00, C10, C20, C23, L00, L10, L20, L23
- magnifications: x200, x1000 and x30000

Image data

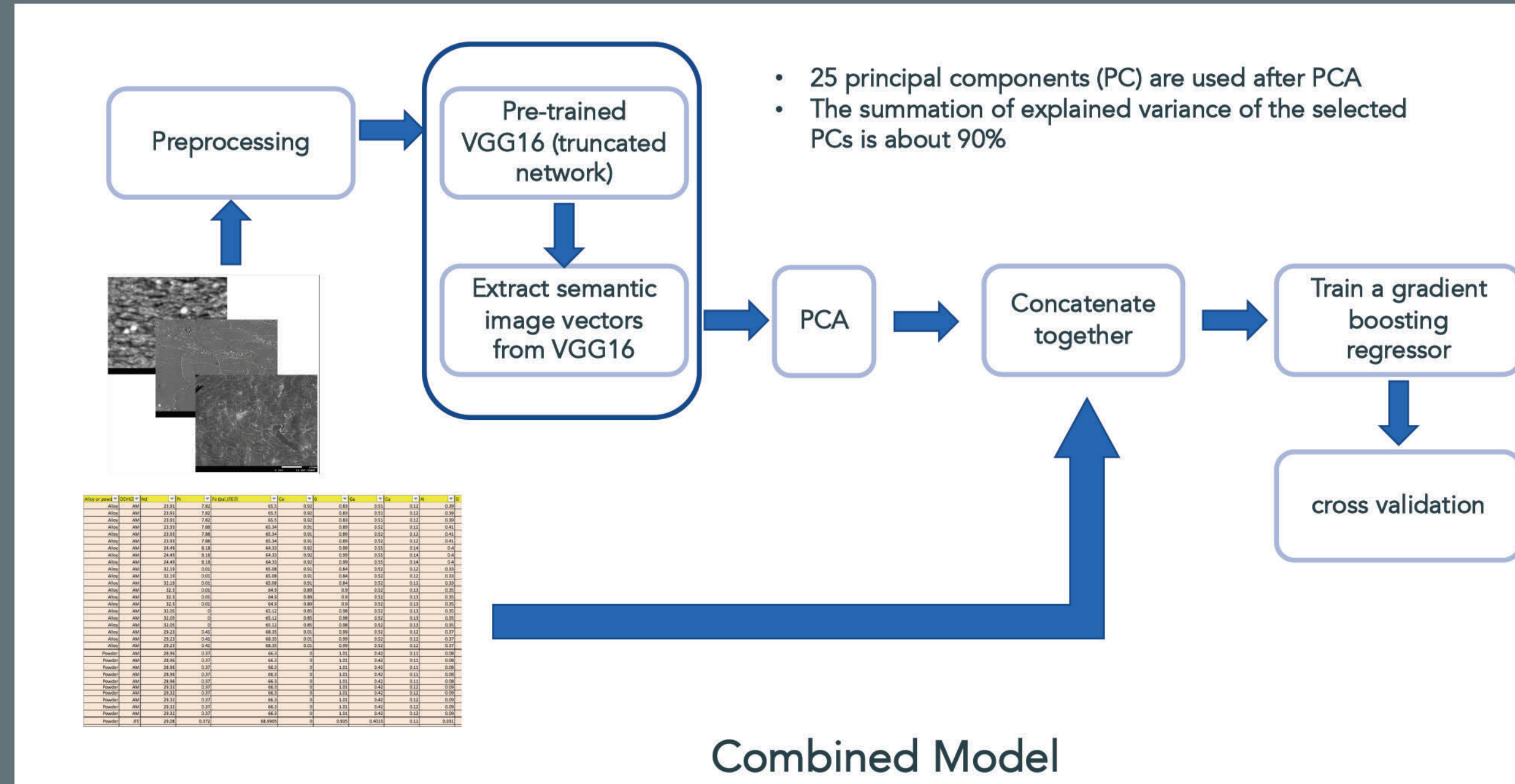
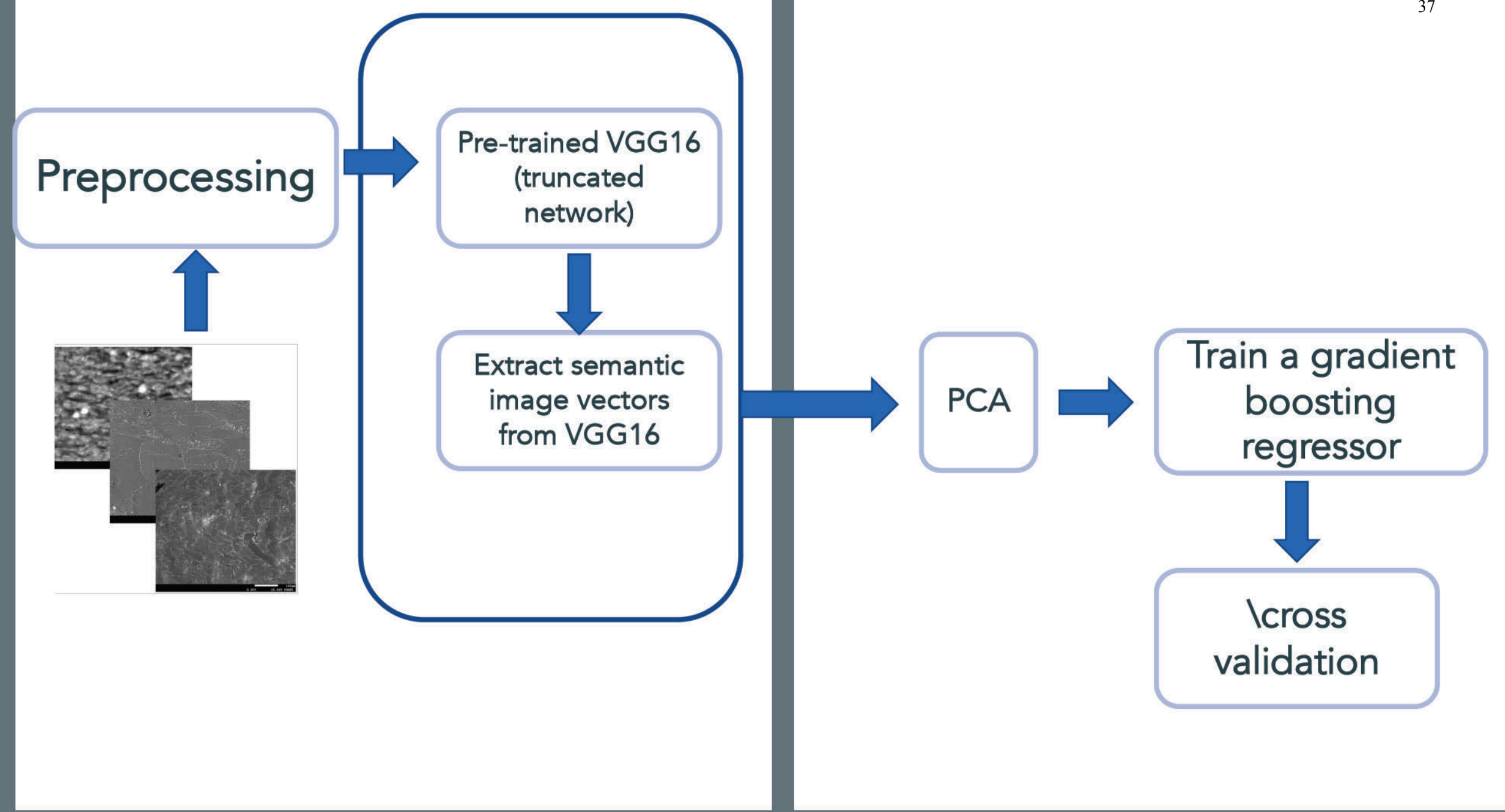
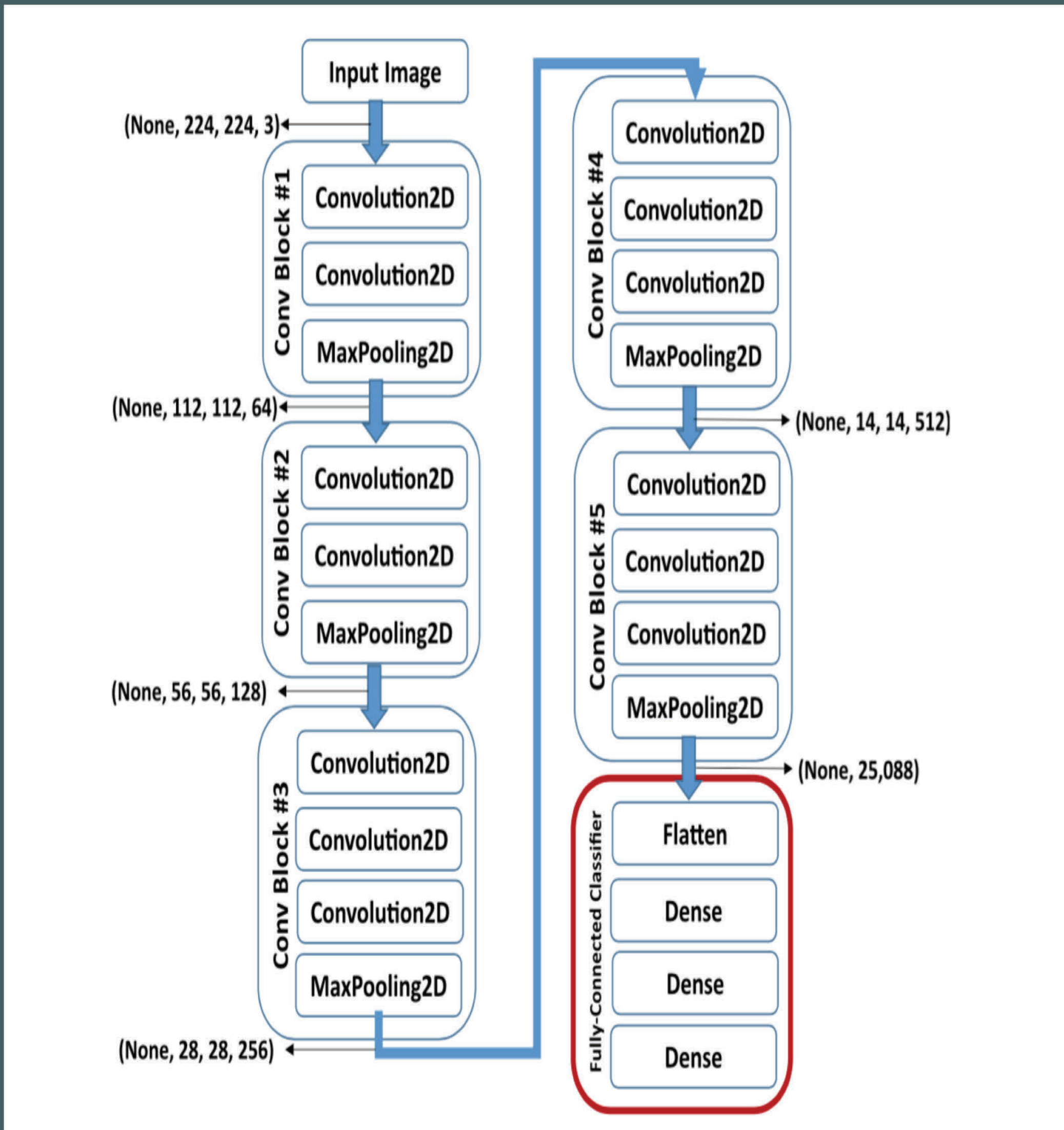


folder	file name	machine	mode	target	position (fig.)	magnification
T160223	T160223-068	SEM	SE	powder N1	cross section/ part of length	×2000
	T160223-069		COMPO		cross section/ part of length	×2000
	T160223-070	SEM	SE	powder N2	cross section/ part of length	×2000
	T160223-071		COMPO		cross section/ part of length	×2000
	T160223-072	SEM	SE	powder N3	cross section/ part of length	×2000
	T160223-073		COMPO		cross section/ part of length	×2000
	T160223-074	SEM	SE	powder N4	cross section/ part of length	×2000
T160223-075	COMPO		cross section/ part of length		×2000	
B150831	B150831-08-C-00-x30k-01	SEM	COMPO	As forge	cross section/ C-00	×30000
	B150831-08-C-00-x30k-02		SE		cross section/ C-00	×30000
	B150831-08-C-00-x30k-03	SEM	COMPO	As forge	cross section/ C-00	×30000
	B150831-08-C-00-x30k-04		SE		cross section/ C-00	×30000
	B150831-08-C-00-x200-01	SEM	COMPO	As forge	cross section/ C-00	×200
	B150831-08-C-00-x200-02		SE		cross section/ C-00	×200
	B150831-08-C-00-x1000-01	SEM	COMPO	As forge	cross section/ C-00	×1000
	B150831-08-C-00-x1000-02		SE		cross section/ C-00	×1000
	B150831-08-C-10-x30k-01	SEM	COMPO	As forge	cross section/ C-10	×30000
	B150831-08-C-10-x30k-02		SE		cross section/ C-10	×30000
	B150831-08-C-10-x30k-03	SEM	COMPO	As forge	cross section/ C-10	×30000
	B150831-08-C-10-x30k-04		SE		cross section/ C-10	×30000
	B150831-08-C-10-x200-01	SEM	COMPO	As forge	cross section/ C-10	×200
	B150831-08-C-10-x200-02		SE		cross section/ C-10	×200
	B150831-08-C-10-x1000-01	SEM	COMPO	As forge	cross section/ C-10	×1000
	B150831-08-C-10-x1000-02		SE		cross section/ C-10	×1000
	B150831-08-C-20-x30k-01	SEM	COMPO	As forge	cross section/ C-20	×30000
	B150831-08-C-20-x30k-02		SE		cross section/ C-20	×30000
	B150831-08-C-20-x30k-03	SEM	COMPO	As forge	cross section/ C-20	×30000
	B150831-08-C-20-x30k-04		SE		cross section/ C-20	×30000
	B150831-08-C-20-x200-01	SEM	COMPO	As forge	cross section/ C-20	×200
	B150831-08-C-20-x200-02		SE		cross section/ C-20	×200
	B150831-08-C-20-x1000-01	SEM	COMPO	As forge	cross section/ C-20	×1000
	B150831-08-C-20-x1000-02		SE		cross section/ C-20	×1000
	B150831-08-L-00-x200-01	SEM	COMPO	As forge	cross section/ L-00	×200
	B150831-08-L-00-x200-02		SE		cross section/ L-00	×200
	B150831-08-L-10-x200-01	SEM	COMPO	As forge	cross section/ L-10	×200
	B150831-08-L-10-x200-02		SE		cross section/ L-10	×200
	B150831-08-L-20-x200-01	SEM	COMPO	As forge	cross section/ L-20	×200
	B150831-08-L-20-x200-02		SE		cross section/ L-20	×200
B150831-08-L-23-x200-01	SEM	COMPO	As forge	cross section/ L-23	×200	
B150831-08-L-23-x200-02		SE		cross section/ L-23	×200	



or later same with above

Deep (transfer) learning



Gopalakrishnan, Kasthurirangan, et al. "Deep Convolutional Neural Networks with transfer learning for computer vision-based data-driven pavement distress detection." *Construction and Building Materials* 157 (2017): 322-330.

Observations and Impact

➤ Workflow

- Complex
- Many teams
- Each needs expertise, resources and access
- Involves Experiments, simulations, Instruments and ML

➤ Cost Implications (savings)

- E.g., 2 out of 8 image orientations have predictive value => significant reduction in (1) SEMs (2) people time, (3) sample materials => More explorations faster
- Fewer and relevant experiments
- Avoid higher-end processing steps for not-so-promising candidates
- Only most important locations/magnifications for SEM
- *Millions of \$\$\$ or Billions of Yen savings*

➤ Discovery and Design acceleration

- Identify and explore the most promising materials
- Discover the high performing materials faster

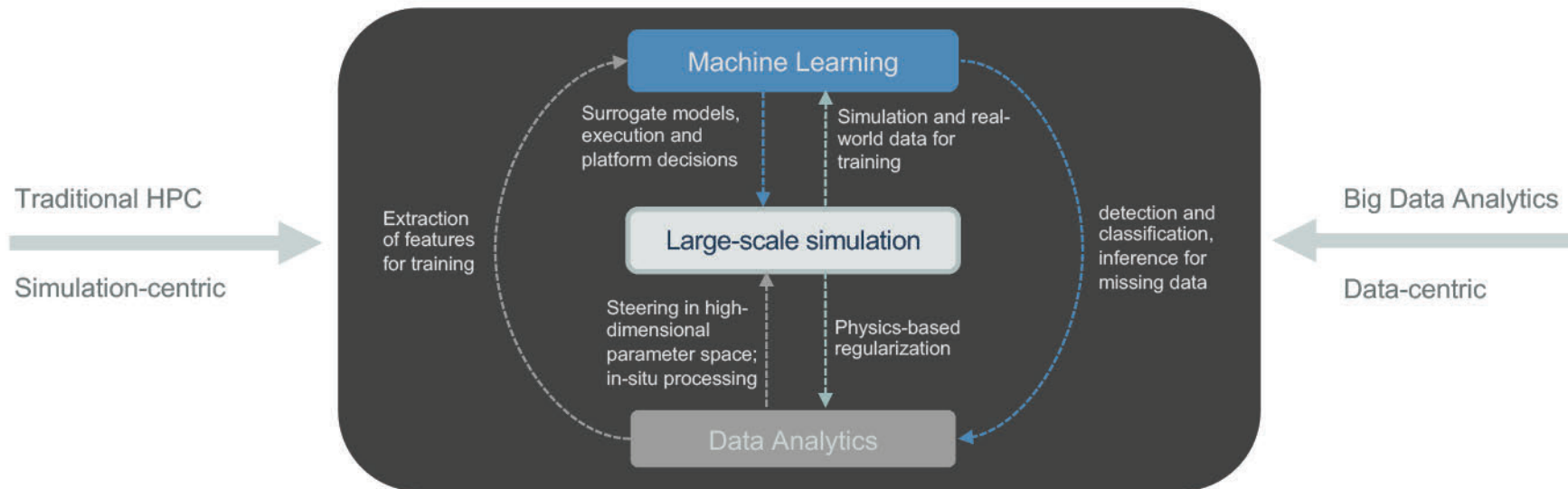
Converged Software Platform for Data Analytics and Extreme-Scale Computing

Carlos Costa

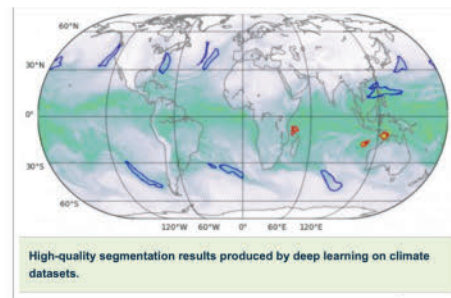
Data-Centric Solutions (DCS)
IBM T.J. Watson Research Center, NY USA

chcost@us.ibm.com

Emerging Intelligent Discovery Workflows



- Emerging workflows impose a paradigm shift from simulation-centric to **data-centric discovery**
- In-situ analytics and machine learning for simulation steering and generation of surrogate models becoming key pieces to enable **next-gen workflows**
 - e.g., 3 out of 6 2018 Gordon Bell finalists had some sort of simulation+ML/DL+analytics hybrid workflow
- Challenge in deploying and integrating disjoint software platforms and enabling efficient **data flow**

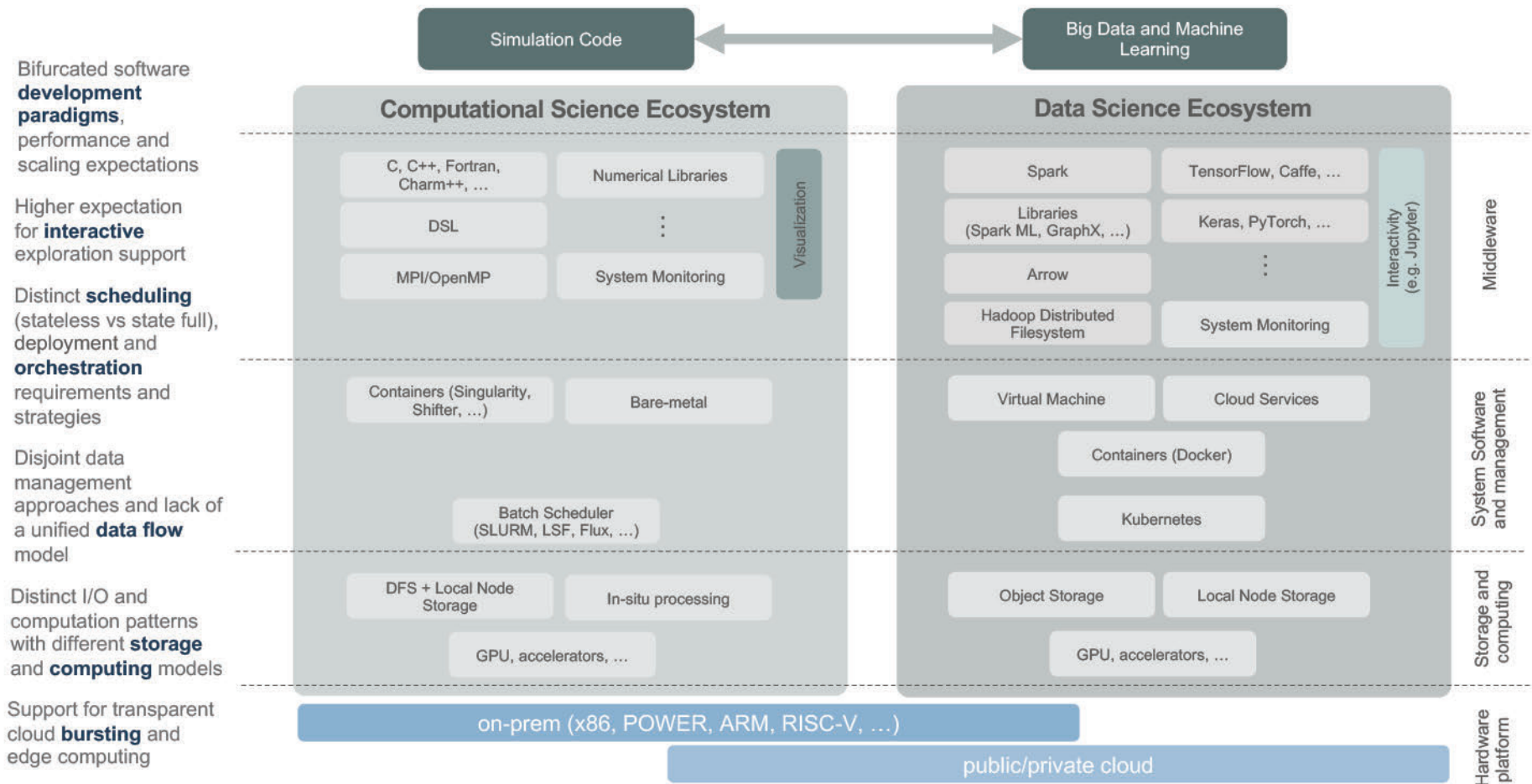


Employing Deep Learning Methods to Understand Weather Patterns (LBL)
2018 Gordon Bell Prize

<https://bit.ly/2X42Vur>



Challenges in a Converged Software Platform



Bifurcated software **development paradigms**, performance and scaling expectations

Higher expectation for **interactive** exploration support

Distinct **scheduling** (stateless vs state full), deployment and **orchestration** requirements and strategies

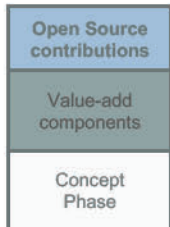
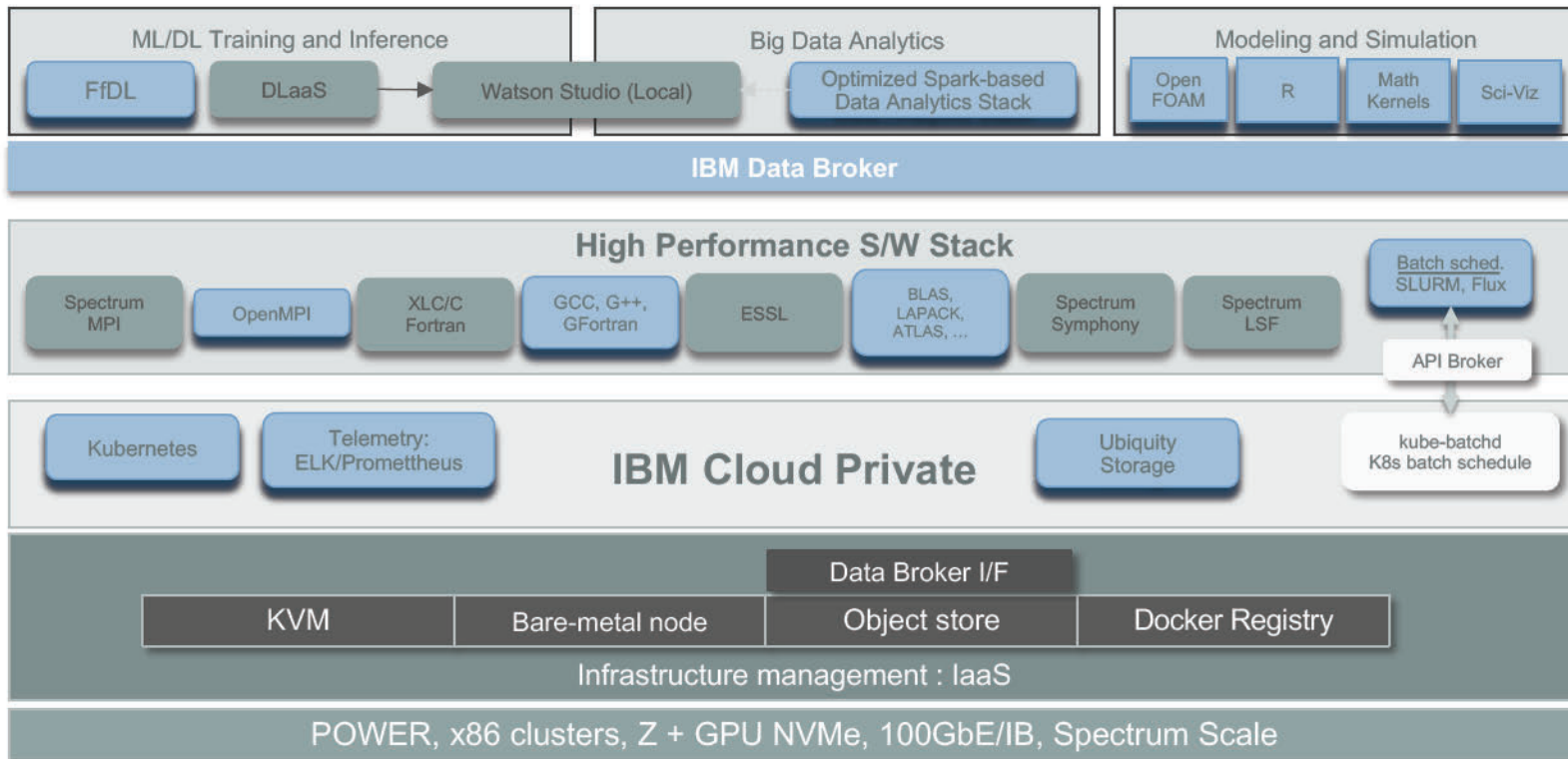
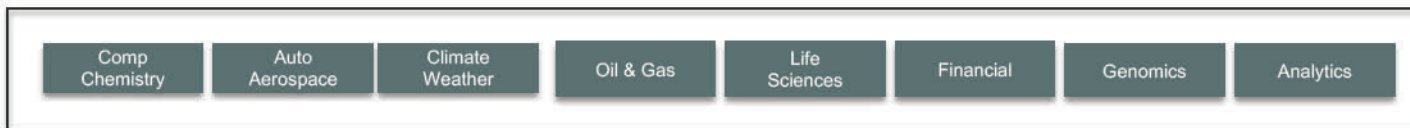
Disjoint data management approaches and lack of a unified **data flow** model

Distinct I/O and computation patterns with different **storage** and **computing** models

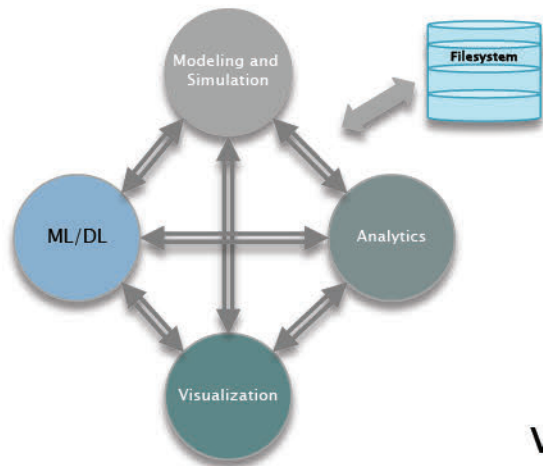
Support for transparent cloud **bursting** and edge computing

Our approach for a Converged Software Platform

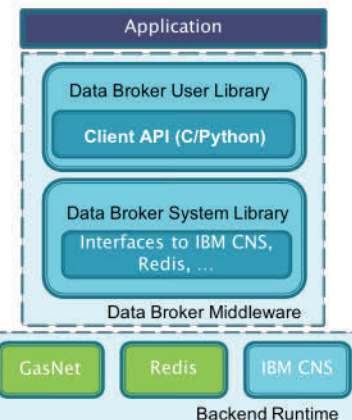
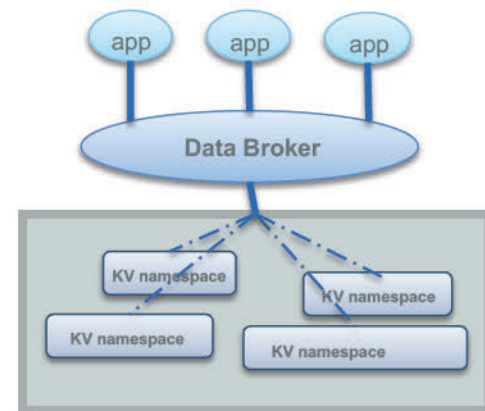
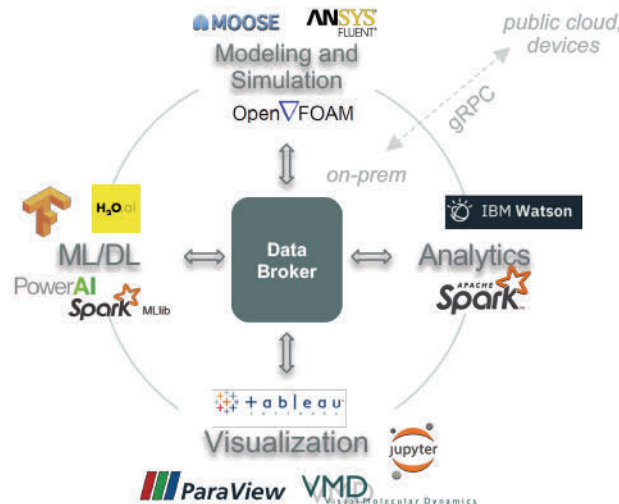
Workflows



Unified Data flow with Data Broker



VS.



Files I/O

- Longer latency
- Less granularity

Sockets

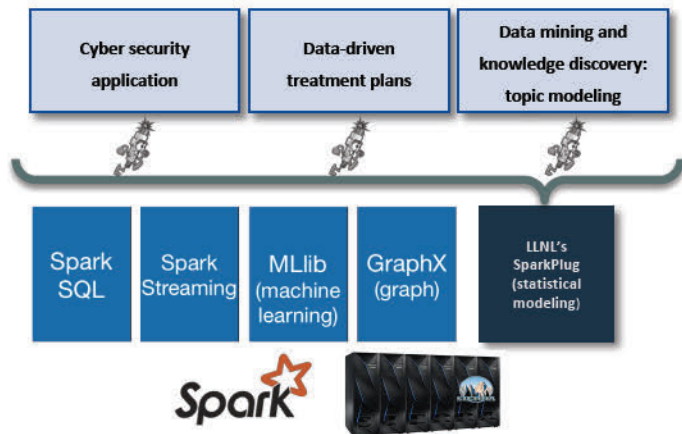
- Longer latency
- Multiple sockets per application
- Discovery for new apps is complicated

Data Broker

- Shared storage framework for data and message exchange
- Simple API to access persistent or volatile storage through distributed tuple-based **global namespaces**
- Data Broker can be accelerated via H/W support
- Discovery of apps via Data Broker

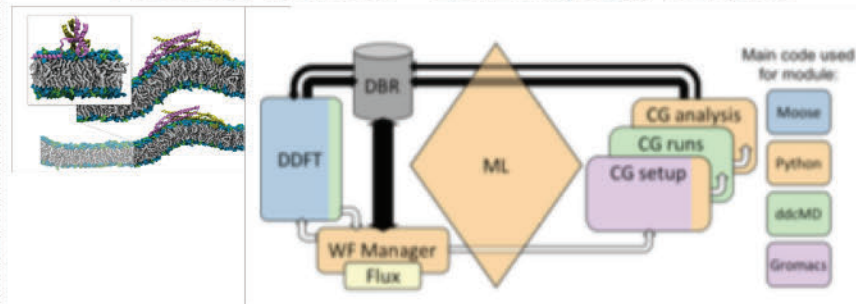


Knowledge Discovery - LLNL's SparkPlug



- Spark-based toolbox for big data machine learning at scale
 - Distributions, estimators, combinator, graphical model templates, samplers
- Allows complex models to utilize application specific understanding
- Demonstrated extreme-scale topic modeling (Latent Dirichlet Allocation) on **LLNL's Sierra**
- Significantly better scaling and performance with optimized software stack

Precision Medicine - LLNL's Splash Workflow



- Multiscale framework to simulate the bending of Lipid Cell membrane
- Impacts how molecules enter the cell and key for designing targeted drugs for RAS-initiated cancer
- The WorkFlow (WF) Manager connects two scales: Dynamic Density Functional Theory (DDFT) and coarse grain (CG)
- DDFT simulation are decomposed into patches, and the WF Manager feeds them to the machine learning (ML) infrastructure, which maintains a priority queue of candidate patches
- WF Manager picks top candidates and uses the Flux resource manager to start new CG simulations
- Data transfer and messaging is handled through the DataBroker (DBR), a fast, system-wide key-value store
- Runs natively on LLNL's Sierra
- *Containerization* effort to run on IBM ICp

Improving Workflow Management Systems

Ewa Deelman, Ph.D.

University of Southern California,
Information Sciences Institute

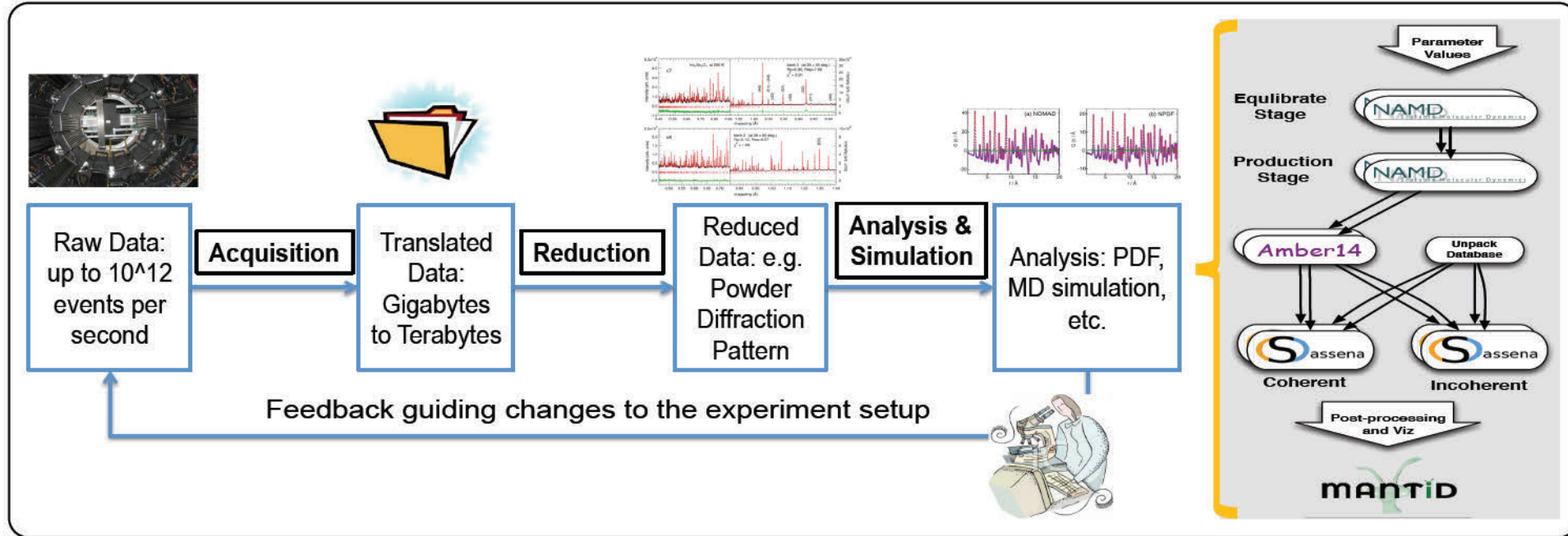
Big Data and Exascale Computing 2

February 2019, Kobe



Scientist @ Instrument

- Need for real time feedback to manage experiment



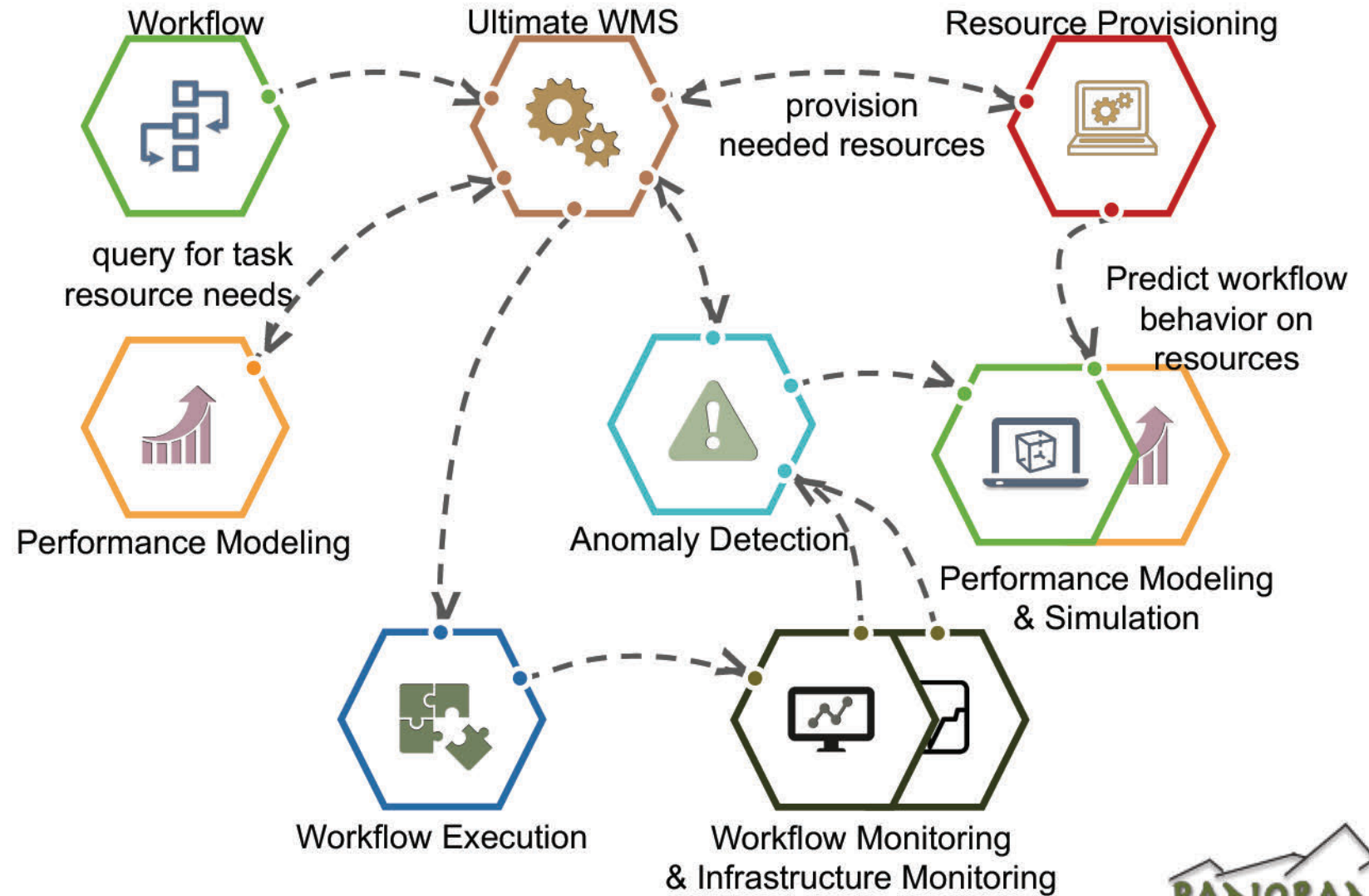
A Pegasus workflow was developed that showed that *nanodiamonds* can enhance the dynamics of tRNA

It compared SNS neutron scattering data with MD simulations by calculating the epsilon that best matches experimental data

Ran on a Cray XE6 at NERSC using 400,000 CPU hours, and generated 3TB of data.

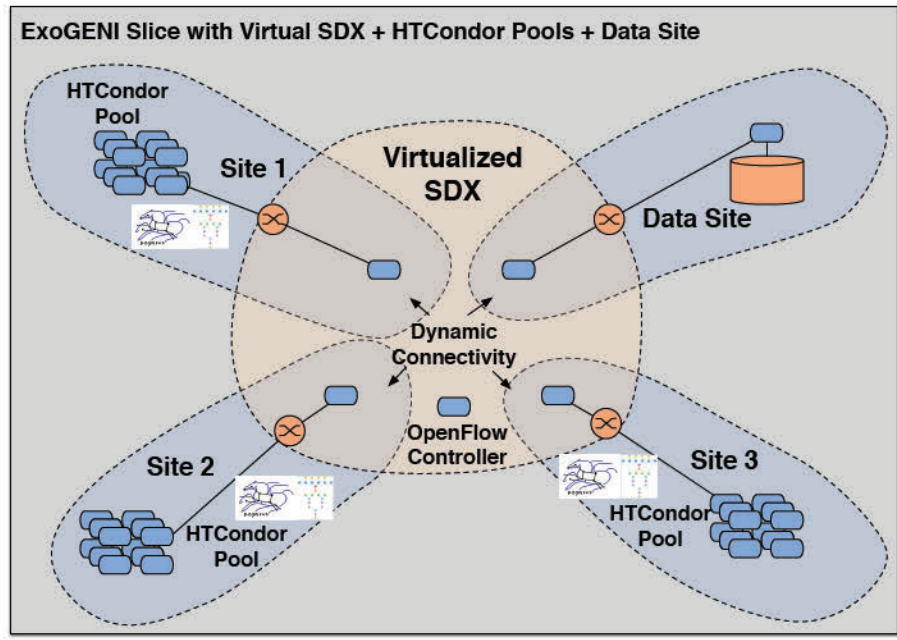


Furthering Workflow Automation



Network Provisioning

- We predicted completion times of current and future workflow tasks and start times for data staging under different scenarios to provision appropriate compute & network resources
- We developed mechanisms to arbitrate and prioritize data flows from competing workflows by leveraging advanced network provisioning technologies like a virtual Software Defined Exchange (SDX)



Software Defined Exchanges (SDX)

Meeting point of networks to exchange traffic, securely and with QoS, using SDN protocols

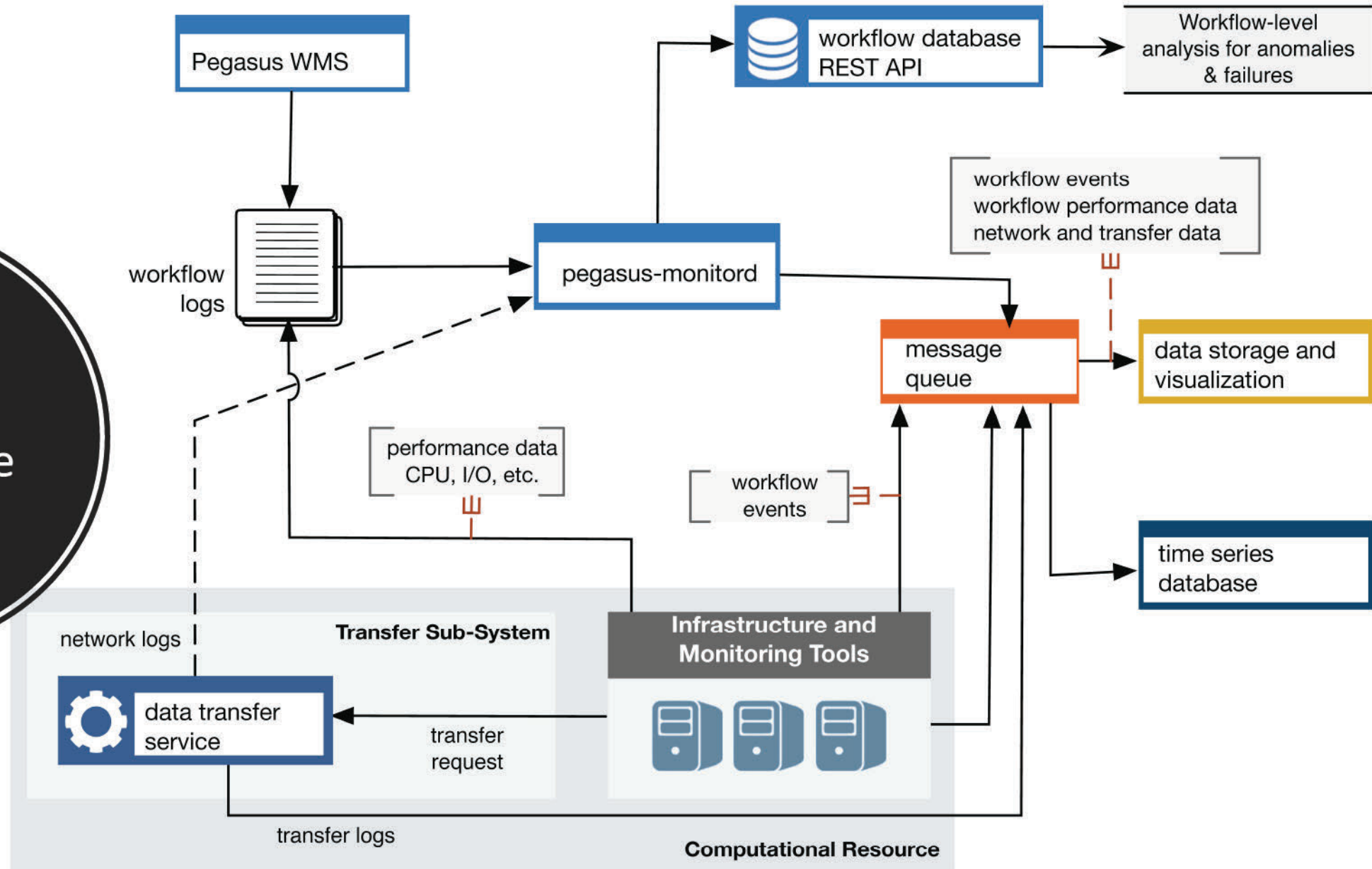
Virtual SDX

Virtual overlay acting as SDX without persistent physical location

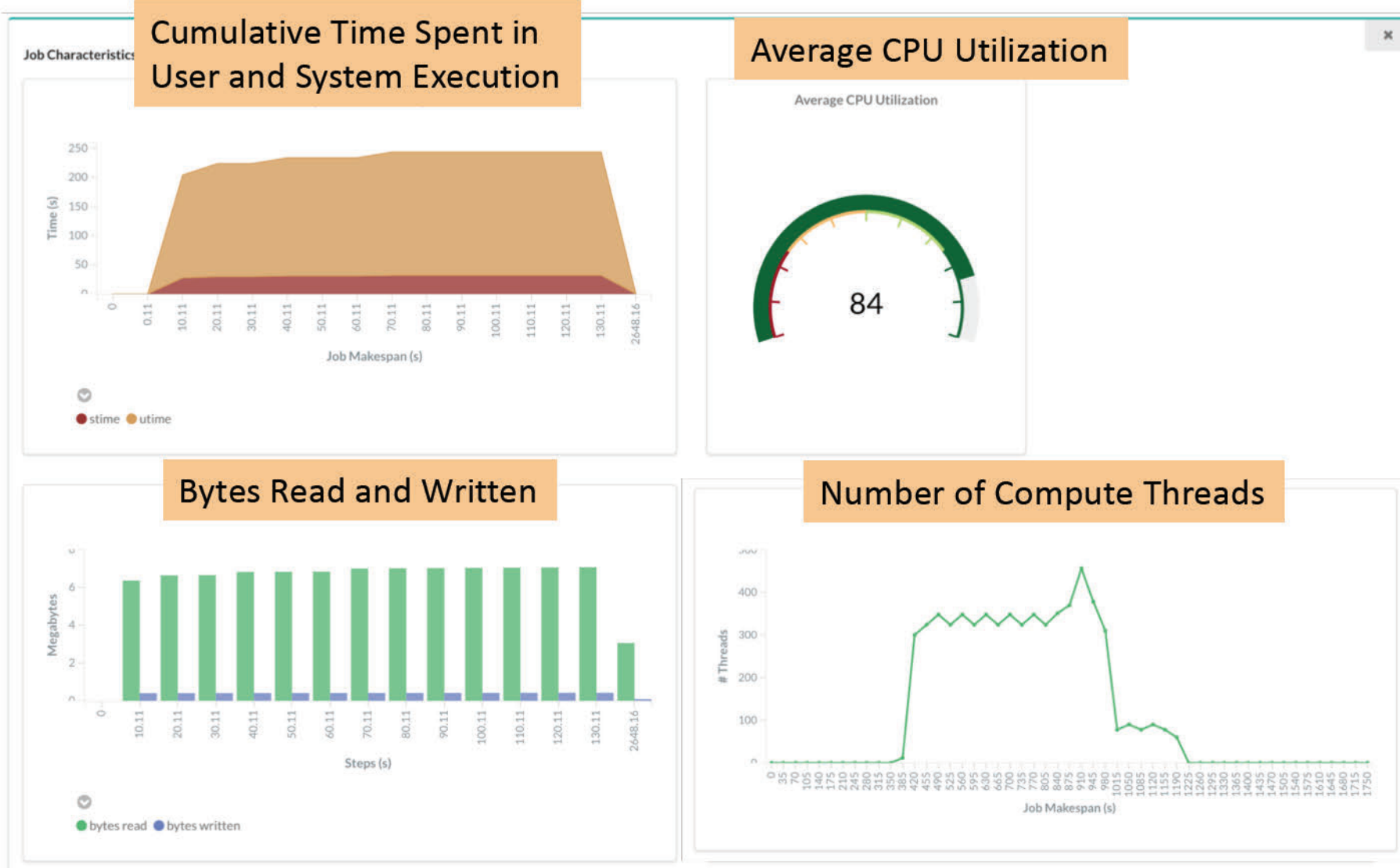
ExoGENI virtual SDX can modify compute, network, storage to support changing demands of SDX



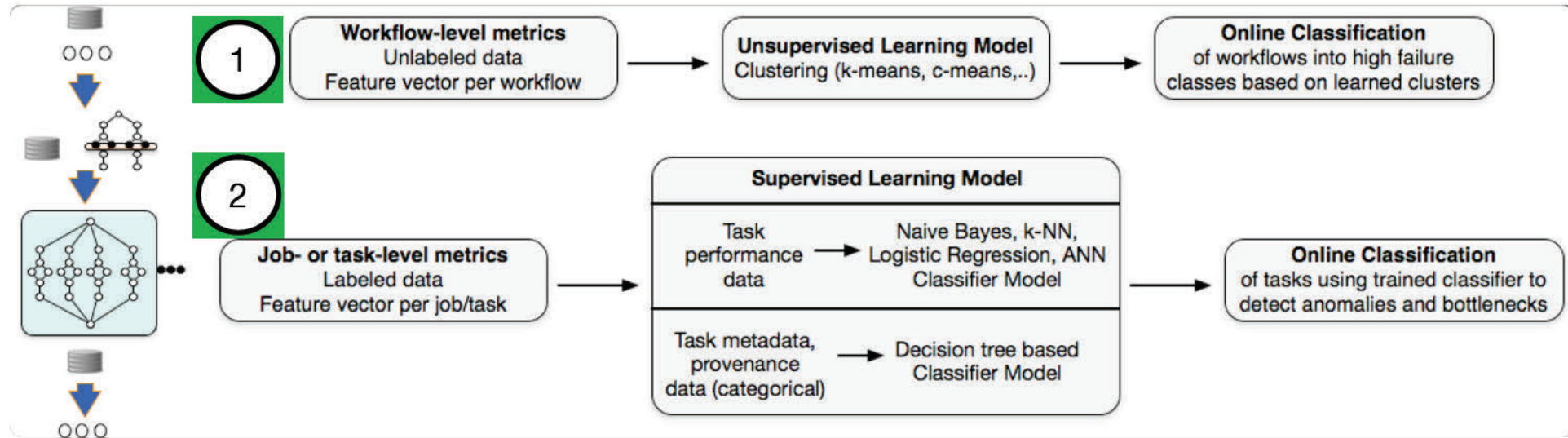
Data Collection: Architecture Overview



Visualization: Time Series Data of Workflow Performance



Workflow Performance Analysis: Anomaly Detection



- Multivariate techniques, particularly Machine Learning (ML) algorithms provide the appropriate theoretical foundation.
- Apply ML algorithms in a top-down approach.
- **1** Use workflow-level performance analysis to predict overall behavior of running workflow by clustering statistically similar workflows. **2** Job/task-level analysis is triggered to detect faults and bottlenecks using task-level metrics.



EDGE COMPUTING APPLICATIONS

NICOLA FERRIER

Senior Computer Scientist
Mathematics and Computer Science Division
Argonne National Laboratory, USA

Exploring the needs of three “edge computing” applications for a new, shared, advanced cyberinfrastructure platform

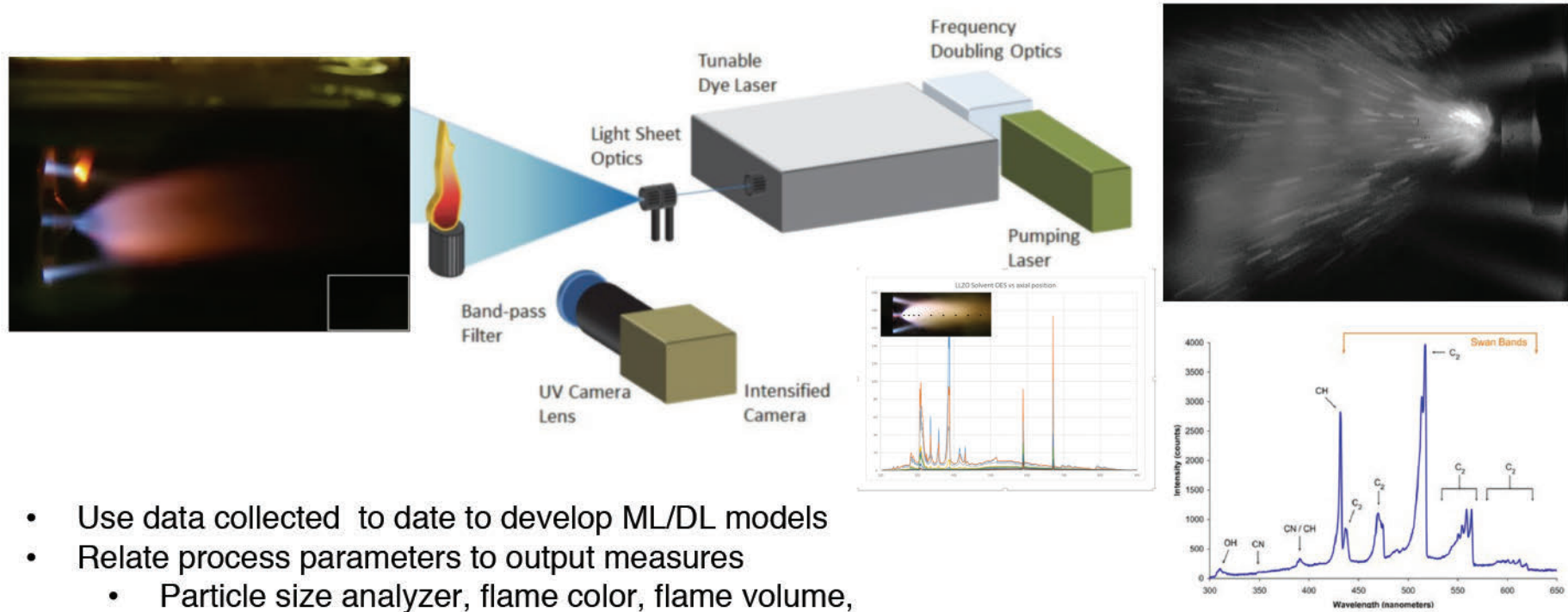
- Manufacturing
- Scientific image data pipelines
- Urban sensing platforms



Edge Computing for Manufacturing:

53

Example manufacturing process: Flame Spray Pyrolysis

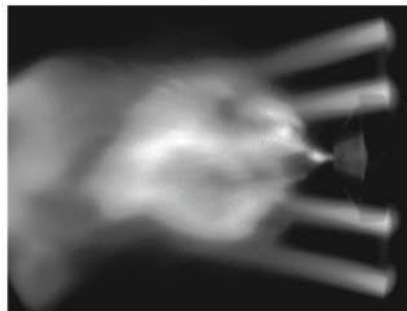
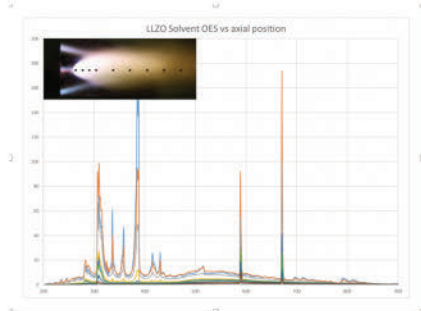


- Use data collected to date to develop ML/DL models
- Relate process parameters to output measures
 - Particle size analyzer, flame color, flame volume, optical emission spectrometer, Laser PLIF
- Optimize process

~20 parameters:

- Composition
- Gas flow rates
- Temperature
- Nozzle geometry
- ...

Process control/feedback
active learning

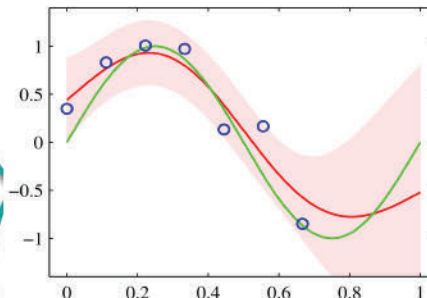


Collect data

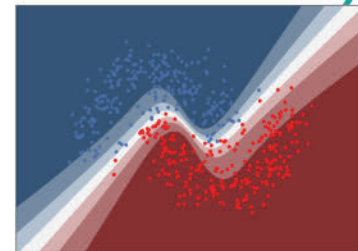
Characterize product, e.g.
particle size distributions

HPC or Cloud

Develop machine learning
surrogate model(s)



Thermo-chemical
Models



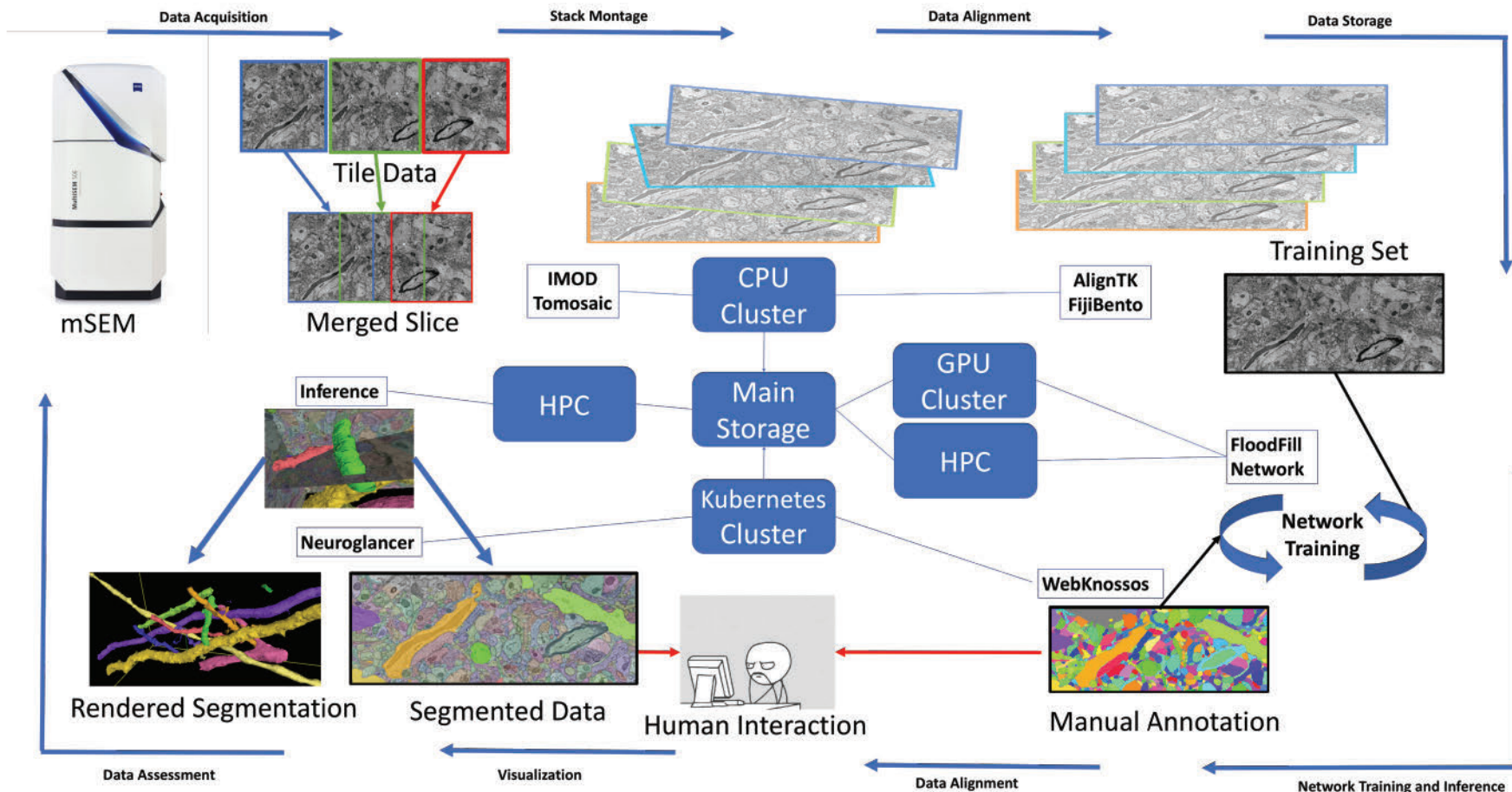
Bayesian Neural
Network

54



Processing pipeline for EM brain data

55



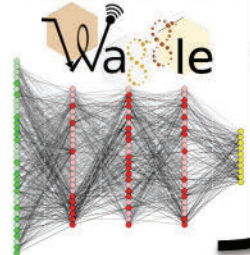
WAGGLE PLATFORM VISION ⁵⁶



Sensors



Powerful Parallel Edge Computing



Semantic Output

Edge computing and deep learning with feedback for continuous improvement

HPC/Cloud



Reduced, Compressed data

New inference (program code)

Deep Learning Training

Actuators



Artificial Intelligence
Deep Learning Inference

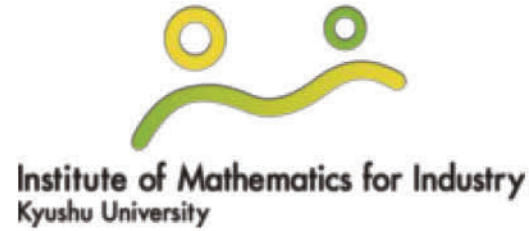
Edge Computing

Applications

- Manufacturing
- Scientific [image] pipelines
- Smart cities
 - Transportation
 - Health
 - Social implication
- Smart homes/buildings

Some Platform Requirements

- Automated resource management
 - Flexibility
 - Joint optimization of communication and computation
- Often low latency needed
- Process and clock synchronization
- Edge computing updates:
 - Data required for training ML
 - New models/inference for novel data, sensors or device placement



Current and future plans with ABCI - AI Bridging Cloud Infrastructure -

Katsuki Fujisawa

Director, AIST-Tokyo Tech Real World Big-Data Computation
Open Innovation Laboratory (RWBC-OIL)

Professor, Institute of Mathematics for Industry, Kyushu University

February 20-21, 2019

BDEC2 Kobe

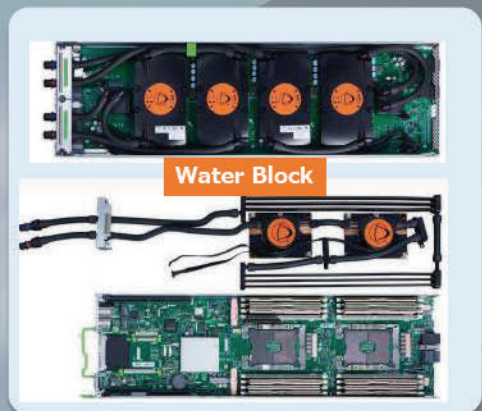
ABCI: The World's First Large-Scale Open AI Infrastructure



Computing Node



Univ. Tokyo / Kashiwa II Campus



Water Block



ABCI AI Bridging Cloud Infrastructure

- World Top-Level compute and data process capability
- **Open, Public, and Dedicated** infrastructure for AI & Big Data Algorithms, Software, and Applications
- **Open Innovation Platform** to accelerate joint academic-industry R&D for AI

Peak Performance:

550 PFlops (FP16)

37 PFlops (FP64)

Effective Performance:

19.88 PFlops (#7 in TOP500)

14.423 GFlops/W (#4 in GREEN500)

Power Usage: < 2.3 MW

Average PUE: < 1.1 (Estimated)



ABC I High-Performance Computing System

0.550 EFlops(FP16), 37.2 PFlops(FP64)
19.88 PFlops(Peak), Ranked #5 Top500 June 2018
Ranked #7 Top500 Nov. 2018

Chips
(GPU, CPU)



Tesla V100



Xeon Skylake-SP



PRIMERGY CX2570 M4

Compute Node
(4GPUs, 2CPUs)

Node Chassis
(2 Compute Nodes)



PRIMERGY CX400 M4

Rack
(17 Chassis)



System
(32 Racks)

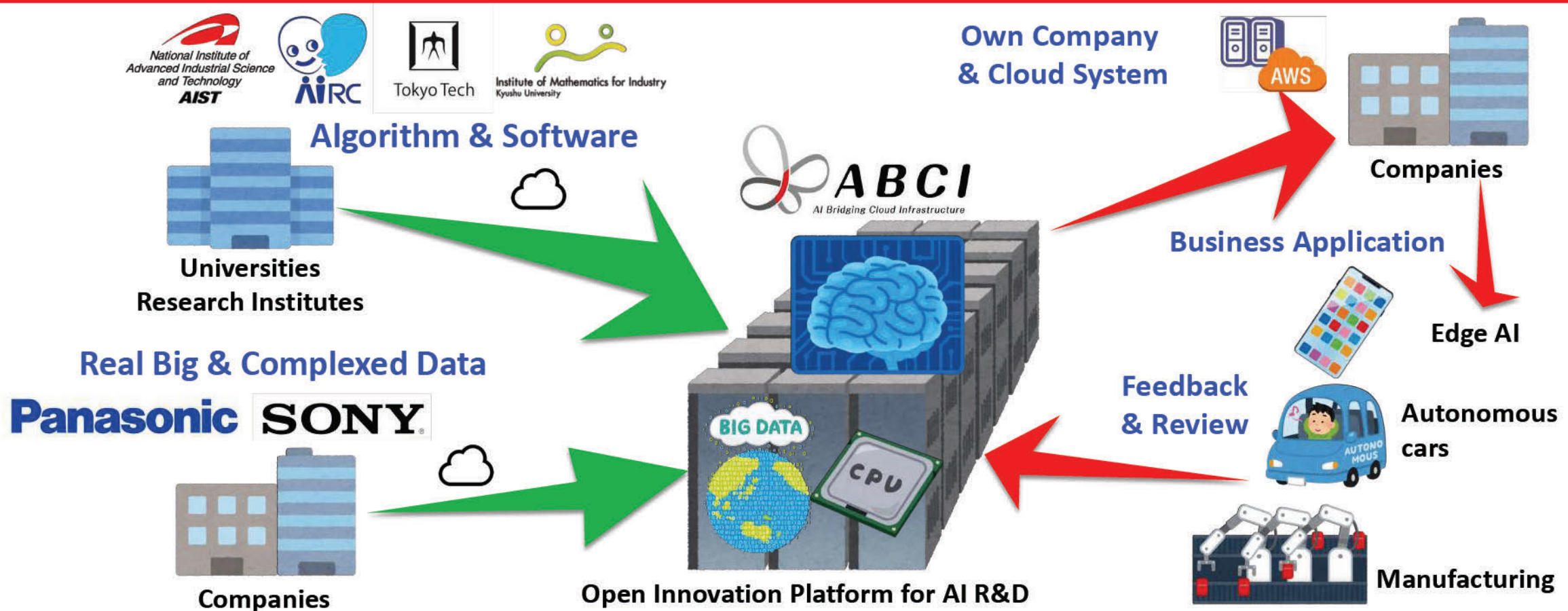


1088 Compute Nodes
4352 GPUs

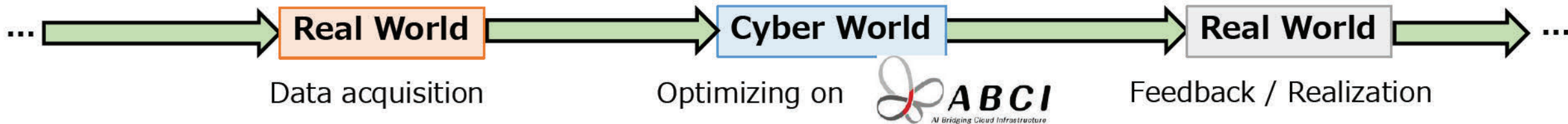
GPU:	CPU:	Node Chassis	Rack	System	
7.8 TFlops(FP64)	1.53 TFlops(FP64)	34.2 TFlops(FP64)	68.5 PFlops(FP64)	1.16 PFlops(FP64)	37.2 PFlops(FP64)
125 TFlops(FP16)	3.07 TFlops(FP32)	506 TFlops(FP16)	1.01 PFlops(FP16)	17.2 PFlops(FP16)	0.55 EFlops(FP16)
		~3.72 TB/s MEM BW		~131TB/s MEM BW	~4.19 PB/s MEM BW
NVIDIA Tesla V100 (16GB SMX2)	384 GiB MEM 200 Gbps NW BW 1.6TB NVMe SSD		Full Bisection BW within Rack 70kW Max	1/3 of Oversubscription BW 2.3MW	

Seamless and Simultaneous Parallel Development based on ABCI: AI Bridging Cloud Infrastructure

- The world's first large-scale Open AI Infrastructure
 - **Open, Public, and Dedicated** infrastructure for AI & Big Data Algorithms, Software, and Applications
 - **Open Innovation Platform** to accelerate joint academic-industry R&D for AI, *international collaborations are also welcome*



CPS(Cyber Physical System) and Industrial Applications



Information

Access log

YAHOO! JAPAN

Clustering

User set : U^i Website : a Cluster : c User set : $U^{i,c}$ Capacity : $C^{i,c}$ Who? Who?

Evaluate website's performance

Colombia opens World Cup with damaging loss to Japan after early red card

People / Objects

Camera Sensor

Graph Analysis Multiple Object Tracking

Panasonic

Visualize People Flow Control People Flow

Transportation

TOYOTA

SUMITOMO ELECTRIC

Modeling & Optimize

Support command CONTROL Segment

G_Sensor(X)

G_Sensor(Y)

Higuchi Method

Fuel Efficient Driving

fuel

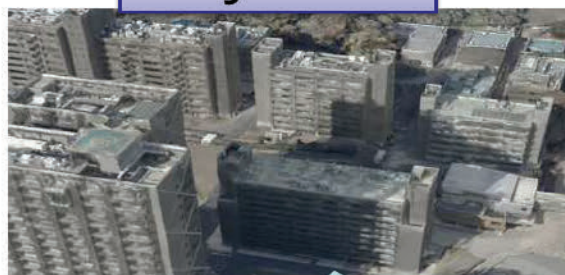
4D Geospatial Information System + CPS Mobility Optimization Engine

Real objects and events

Dynamic 3D map

Physical

Cyber



Data collection + Map renewal

Positioning + Optimized action based on simulation and big data analysis

4D (Position + time)
Geospatial Information

Wearable Devices VR + AR

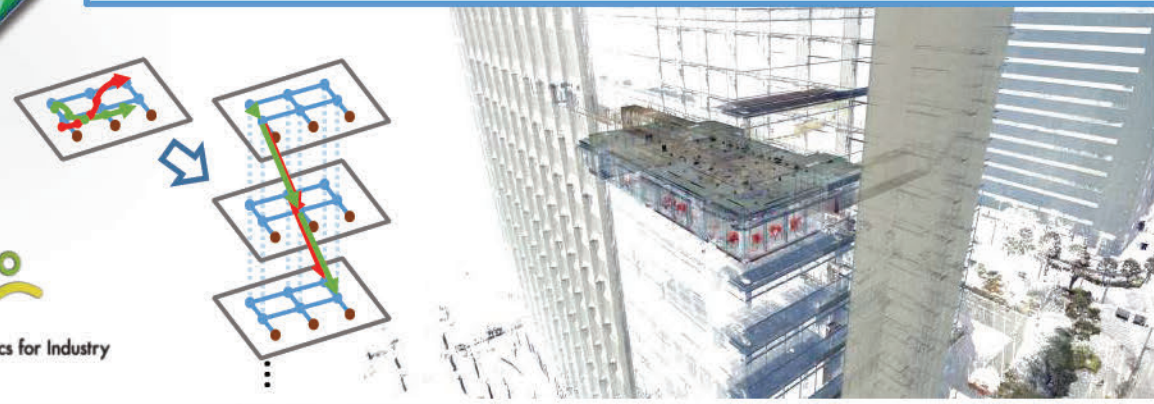
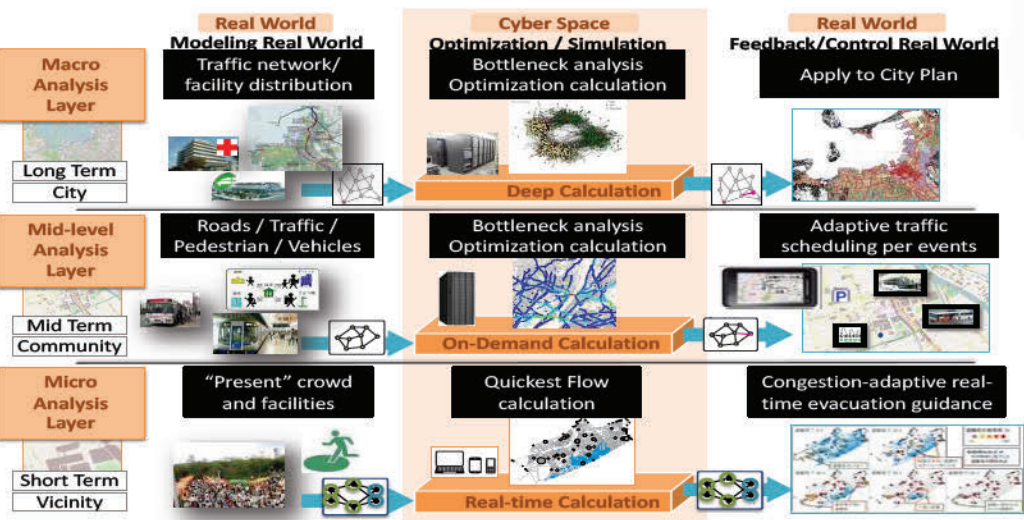
Data
(Imagery + Point)

Knowledge
4D map =
position(3D) + time

4D-GIS

CPS Mobility Optimization Engine

New Generation Personal Navigation System



Current System : Smartphone + Google Maps ⇒
New Generation Personal Navigation System
⇒ 4D Display + CPS Mobility (life, amusement, security) +
Wearable Devices (AR + VR)



OPEN AI Infrastructure

Convergence of equation based modeling and data analytics on HPC resources: example of accelerating finite-element earthquake simulation with data analytics

Earthquake Research Institute, The University of Tokyo

Kohei Fujita

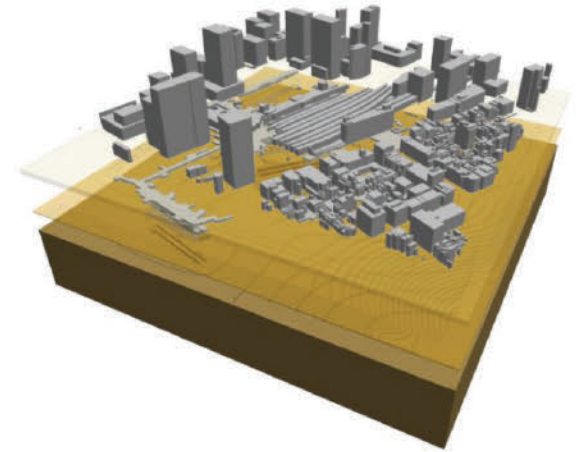
Convergence of equation based modeling and data analytics on HPC resources

- Equation based modeling and data analytics have different characteristics
 - Equation based modeling: Highly precise, but costly
 - Data analytics: Fast inferencing, but accuracy not as high
- More efficient computation is expected by using both methods to complement each other on HPC resources
 - We have been conducting research on this topic focused on earthquake simulation problems
 - In SC17 best poster [1], we used data generated by equation based modeling for data analytics training
 - In SC18 Gordon Bell Prize Finalist paper [2], we trained artificial neural network to accelerate equation based modeling

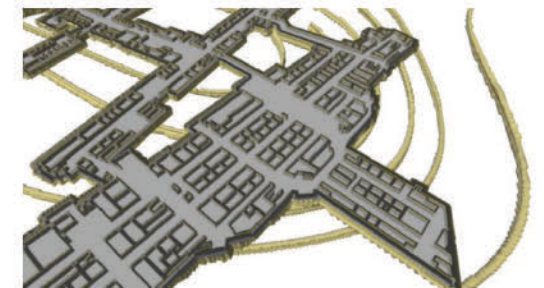
[1] Tsuyoshi Ichimura, et al., AI with Super-Computed Data for Monte Carlo Earthquake Hazard Classification, SC17 Research Poster
[2] Tsuyoshi Ichimura, et al., A Fast Scalable Implicit Solver for Nonlinear Time-Evolution Earthquake City Problem on Low-Ordered Unstructured Finite Elements with Artificial Intelligence and Transprecision Computing, Proceedings of SC18

Example of using data analytics to accelerate equation based modeling

- Target: Solve $Ax = f$ in unstructured implicit finite-element urban earthquake simulation
- Difficulty in using data analytics in solver
 - Data analytics results are not always accurate
 - We need to design solver algorithm that enables robust and cost effective use of data analytics
 - We also need to consider uniformity of computation for scalability on HPC resources
- Use information of underlying governing equation
 - Governing equation's characteristics with discretization conditions should include information about the difficulty of convergence in solver
 - Extract parts with bad convergence using artificial neural network and extensively solve extracted part in preconditioner
 - Same solution obtained with less compute cost



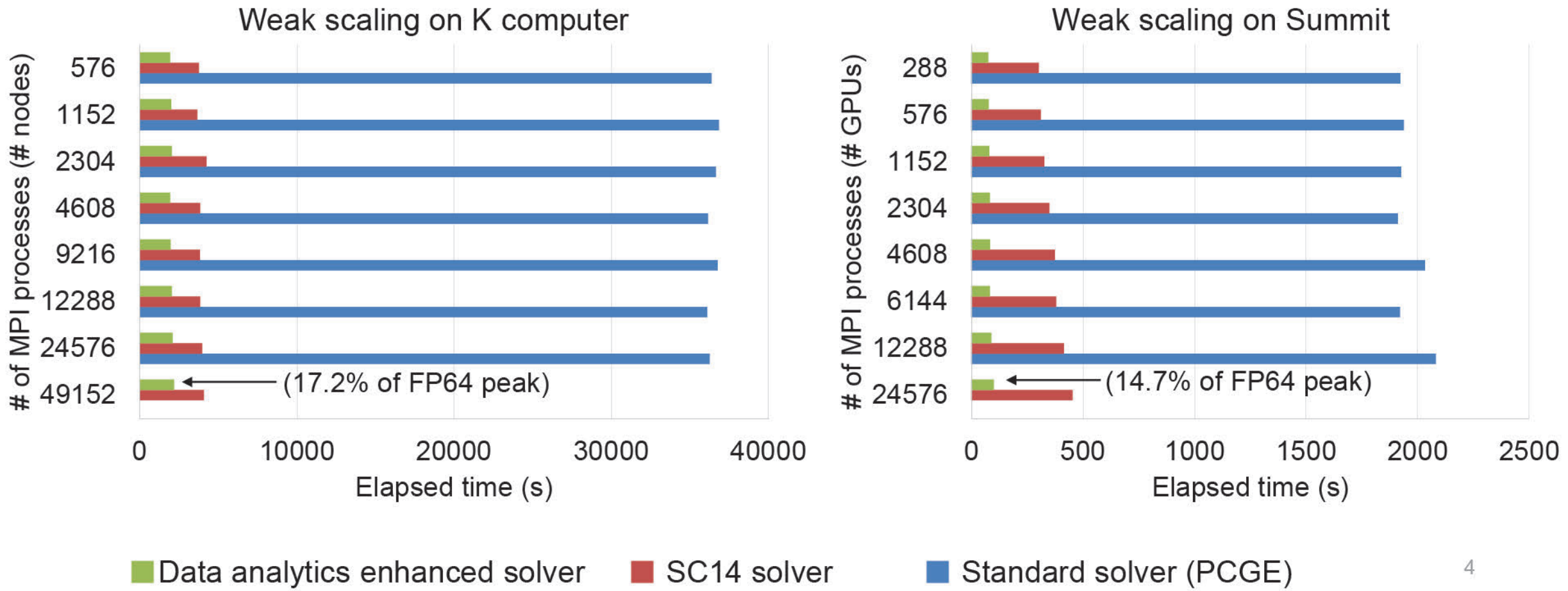
Whole city model



Extracted part by AI
(about 1/10 of model)

Performance of data analytics enhanced solver on K computer/Summit

- FLOP count of data analytics enhanced solver decreased by 5.56-times from PCGE (standard solver; Conjugate Gradient solver with block Jacobi preconditioning), and 1.32 times from a non data analytics enhanced SC14 solver
- Fast and scalable on both CPU based K computer and GPU accelerated Summit



Summary and future implications

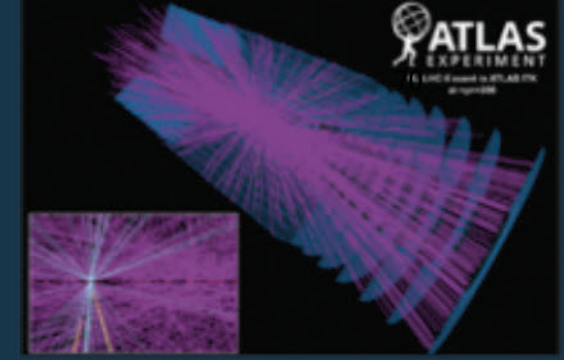
- Equation based modeling can be accelerated using data analytics by careful design of algorithms
 - We accelerated earthquake simulation by designing a scalable solver algorithm that can robustly incorporate data analytics
 - Idea of accelerating simulations with data analytics expected to be generalizable for other types of equation based modeling

HEP Challenges for HPC

Maria Girone

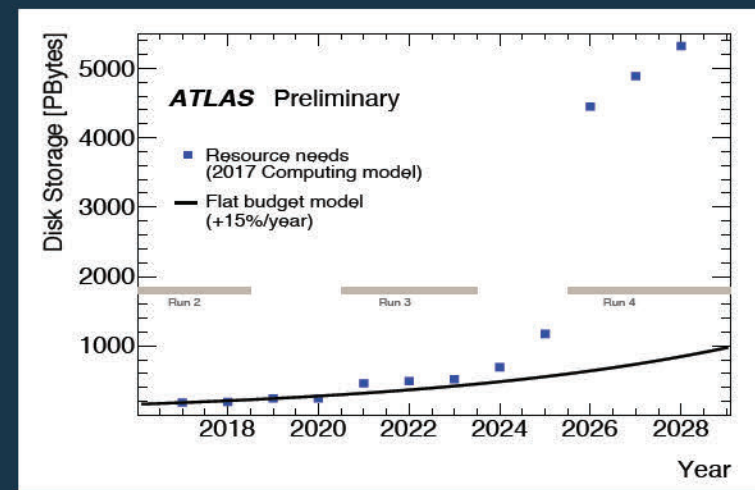
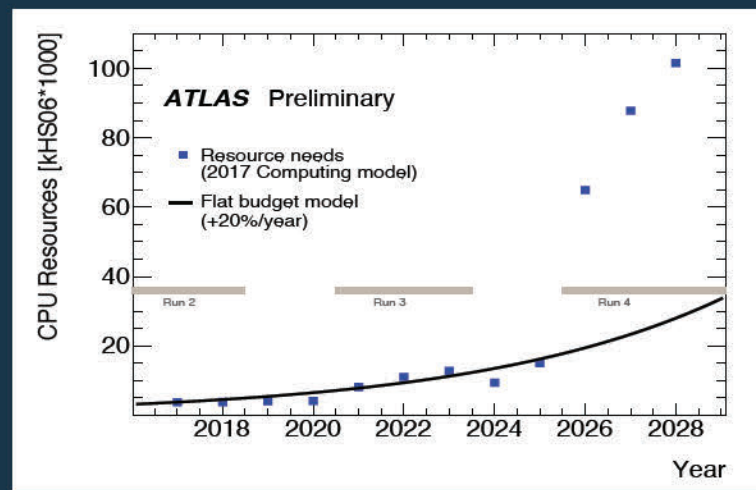
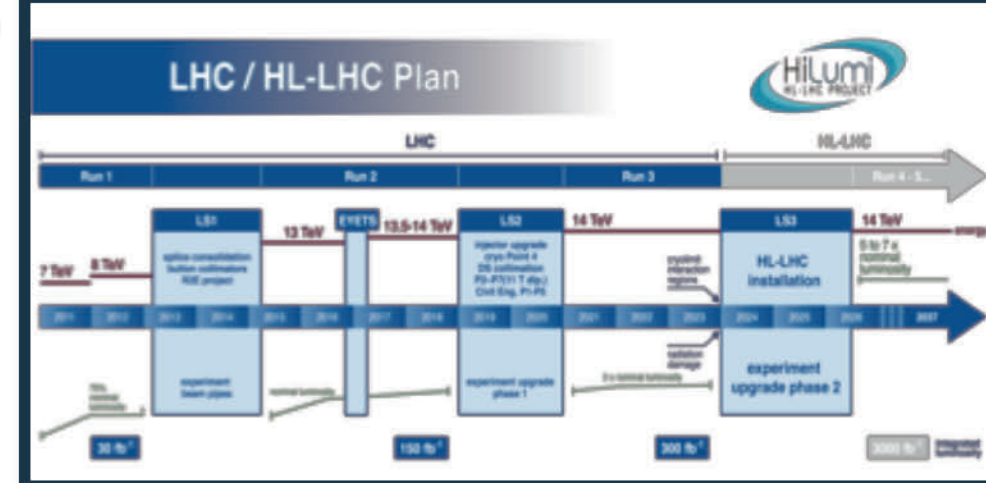
CERN openlab CTO and CERN Coordinator for HPC Europe

The LHC Upgrade Program



Launched in 2010, the LHC programme is only one third complete

- The HL-LHC upgraded accelerator will unleash an order of magnitude more events of much higher complexity
- Major detector upgrades planned for ALICE and LHCb for Run3 and ATLAS and CMS for Run4
- Data selection rates go from 1kHz to 10kHz
- Major computing model and software revisions ongoing for all experiments



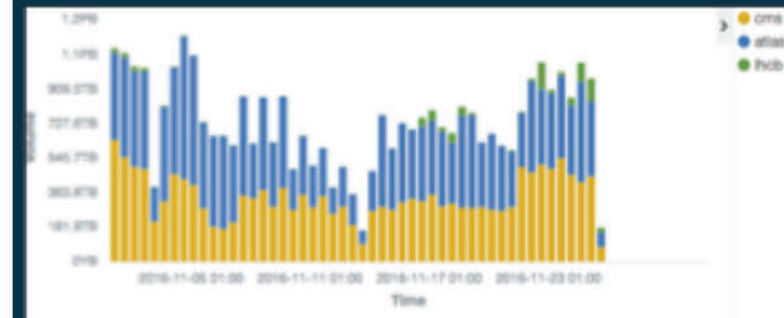
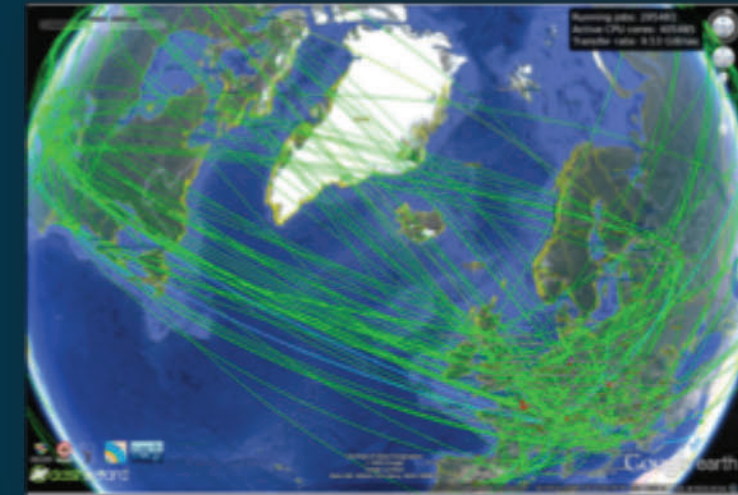
Even with technology improvements there is a huge resource gap

- Exploring new types of computing resources like HPC centres and accelerated architectures

LHC Data Challenges

71

- HEP applications involve the input and output of large data volumes
 - **Reconstruction applications** read raw data to find and write physics objects. They take a constant stream of input data
 - **Simulation applications** are processing intensive and produce large output data
 - The challenge is **delivering data to and from the processing resources**
- The LHC community has developed data management systems to replicate data between sites and **to stream data to running processes**
 - Currently moves **80PB per month** between production and processing and archiving centers
 - Data is accessed by central production and thousands of analysis users
- At HL-LHC the data will reach the **exa-scale** both in total volume and in movement
 - We are looking to improve solutions for Data Organization Management and Access (DOMA)



Maria Girone
CERN openlab CTO

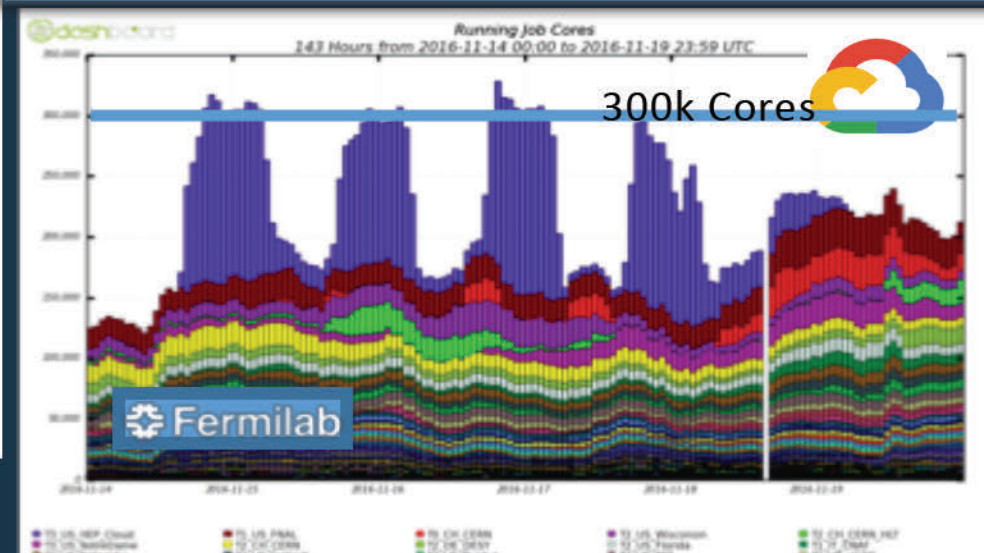
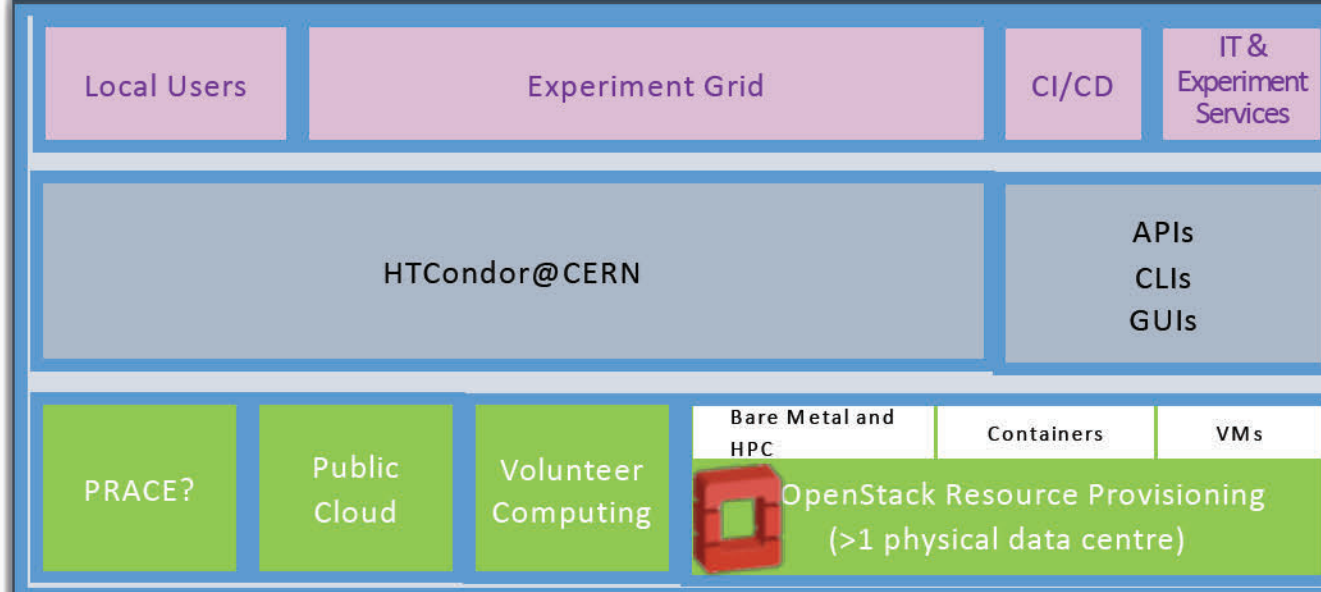
Handshaking with HPC centers

- LHC workflows have peculiar aspects with respect to more standard “HPC” workflows
 - **Architecture** (difficult to overcome x86_64 as primary arch)
 - Code is millions of lines written by hundreds of people over many years
 - **Data intensive**: from low I/O to very high I/O
 - Need for **remote data accesses** (possibly mediated by **edge caches**)
 - Need for **local virtualization** (docker, shifter, singularity, real VMs, ...)
 - Need to **access remote services** (possibly mediated by **edge services**)
 - Workflows are **highly parallelized**, involving 10s of thousands of independent running processes
 - **CVMFS** preferred software distribution solution (but can be worked around with containers)
 - Workflow Management systems currently **absolutely require outgoing network** from the worker nodes due to the tight coupling with central workflow management
 - Often more than one site is processing in parallel

Provisioning Model

CERN was an early adopter and continued leader in virtualization through OpenStack

- Allows for flexible and dynamic deployment
- Moving to containers for even more flexibility
- Expertise from LHC experiments in virtualized environments was critical for cloud deployments and opportunistic HPC access
- Common interfaces between HPC centers will be needed for authorization, access, discovery, and submission
 - Our adoption of HPC sites have each been so far effort intensive custom deployments
 - We would like to develop demonstrators to show a lower cost of adoption at new HPC sites



Heterogenous architectures

Achieving sustainable HEP computing for the HL-LHC and the upgrade program requires change

- CPU evolution is not able to cope with the increasing demand of performance
- Depending on the application, GPUs can provide better performance and energy efficiency

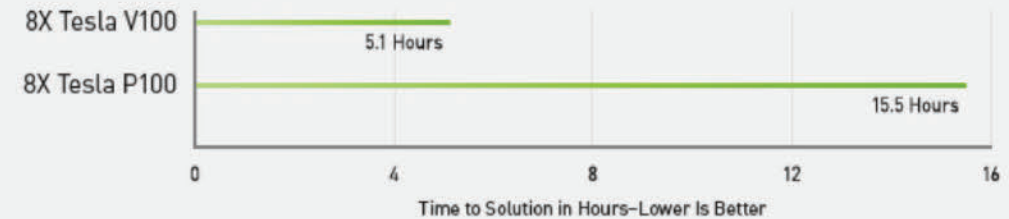
The HPC centers are huge processing resources and are often early adopters of accelerated hardware

- We have an active R&D program to exploit this technology

The next few years are a good opportunity to embrace a paradigm shift towards modern heterogeneous computer architectures and software techniques:

- **Heterogeneous** Computing
- **Machine Learning** (a very active area of development in HEP)

Deep Learning Training in Less Than a Workday



47X Higher Throughput Than CPU Server on Deep Learning Inference



Outlook

HEP is facing a huge resource gap for HL-LHC and would like to be able to utilize the HPC centers

- Exa-scale data needs exa-scale computing

HEP applications have unique challenges in terms of data access and data production

- We would like to work with the HPC centers to address processing challenges of data intensive sciences

We are undertaking a major initiative in software and computing R&D to realize the physics potential of the new program

- There are many activities that would benefit from CERN and the HPC community working together

The Promise of Learning Everywhere and ML⁷⁶forHPC

Geoffrey Fox and Shantenu Jha contribution to Kobe BDEC meeting

- HPC to enhance ML is important
- Arguably more important is the question: Can ML enhance the effective performance of HPC simulations ?
- We argue that ML can enhance HPC simulations by 10^6 if not greater!
 - Enhancement not measured by Flops or usual performance measures
 - But science done using same amount of computing for given accuracy
- Many challenges must be overcome
 - Right hardware, right software system (platform)
 - Application architecture and formulation.
- For details see team papers
 - http://dsc.soic.indiana.edu/publications/Learning_Everywhere.pdf
 - http://dsc.soic.indiana.edu/publications/Learning_Everywhere_Summary.pdf

MLforHPC and HPCforML

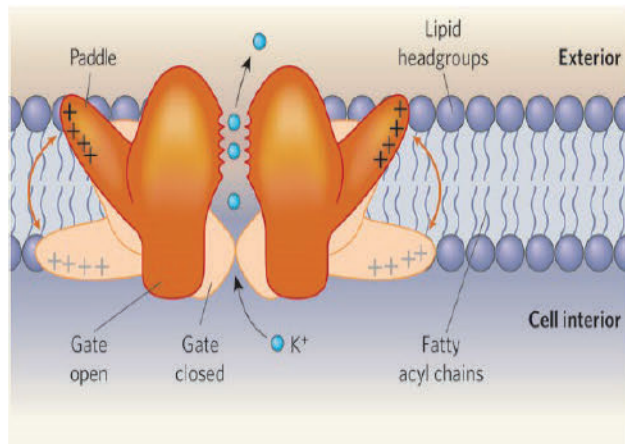
77

We distinguish between different interfaces for ML/DL and HPC.

- **HPCforML:** Using HPC to execute and enhance ML performance, or using HPC simulations to train ML algorithms (theory guided machine learning), which are then used to understand experimental data or simulations.
 - **HPCrunsML:** Using HPC to execute ML with high performance
 - **SimulationTrainedML:** Using HPC simulations to train ML algorithms, which are then used to understand experimental data or simulations.
- **MLforHPC:** Using ML to enhance HPC applications and systems
 - **MLautotuning:** Using ML to configure (autotune) ML or HPC simulations.
 - **MLafterHPC:** ML analyzing results of HPC, e.g., trajectory analysis in biomolecular simulations
 - **MLaroundHPC:** Using ML to learn from simulations and produce learned surrogates for the simulations. The same ML wrapper can also learn configurations as well as results
 - **MLControl:** Using simulations (with HPC) in control of experiments and in objective driven computational campaigns, where simulation surrogates allow real-time predictions.

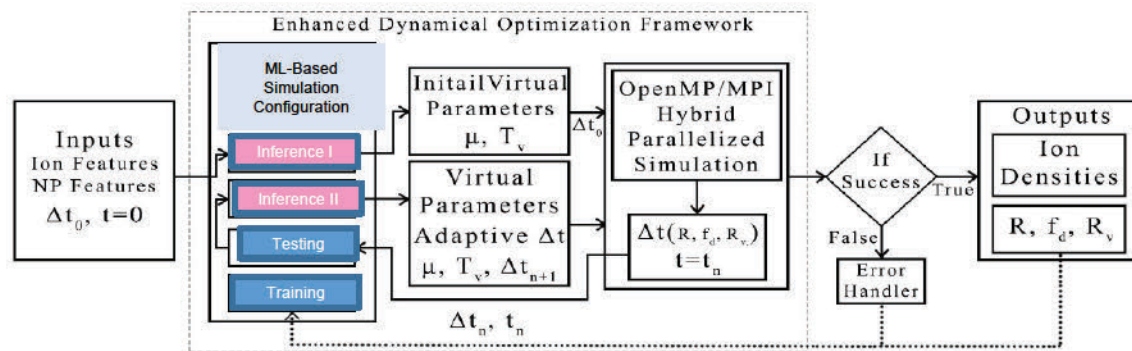
MLAutotuned HPC. Machine Learning for Parameter Auto-tuning in Molecular Dynamics Simulations: Efficient Dynamics of Ions near Polarizable Nanoparticles

JCS Kadupitiya, Geoffrey Fox,
Vikram Jadhao

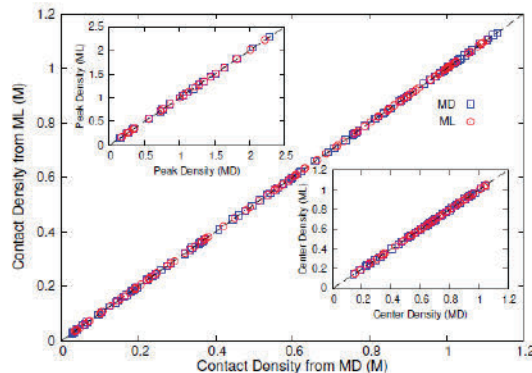
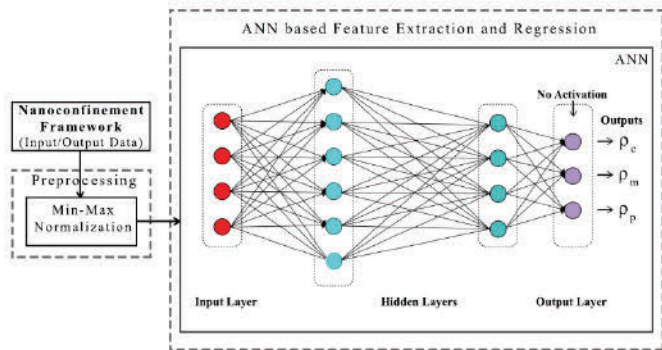


Integration of machine learning (ML) methods for parameter prediction for MD simulations by demonstrating how they were realized in MD simulations of ions near polarizable NPs.

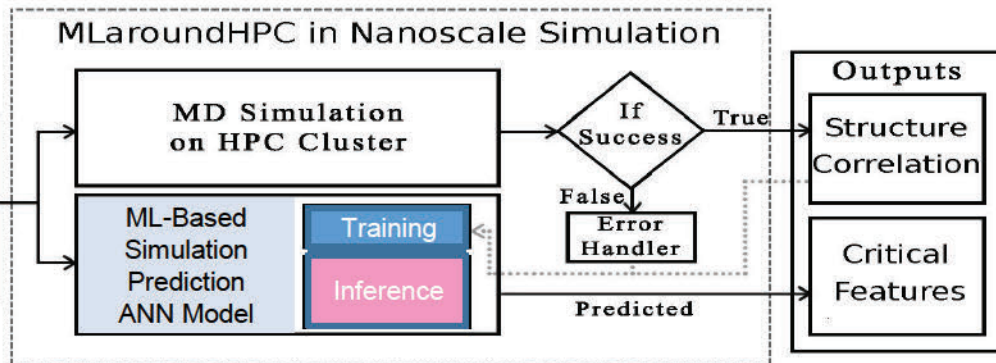
Note ML used at start and end of simulation blocks



MLaroundHPC: Machine learning for performance enhancement with Surrogates of molecular dynamics simulations



- We find that an artificial neural network based regression model successfully learns desired features associated with the output ionic density profiles (the contact, mid-point and peak densities) generating predictions for these quantities that are in excellent agreement with the results from explicit molecular dynamics simulations.
- The integration of an ML layer enables real-time and anytime engagement with the simulation framework, thus enhancing the applicability for both research and educational use.
- Deployed on nanoHUB for education



ML used during simulation

Speedup of MLaroundHPC

- T_{seq} is sequential time
- T_{train} time for a (parallel) simulation used in training ML
- T_{learn} is time per point to run machine learning
- T_{lookup} is time to run inference per instance
- N_{train} number of training samples
- N_{lookup} number of results looked up

N_{train} is 7K to 16K in our work

$$\text{Effective Speedup } S = \frac{T_{seq}(N_{lookup} + N_{train})}{T_{lookup}N_{lookup} + (T_{train} + T_{learn})N_{train}}$$

- Becomes T_{seq}/T_{train} if ML not used
- Becomes T_{seq}/T_{lookup} (10^5 faster in our case) if inference dominates (will overcome end of Moore's law and win the race to zettascale)
- This application deployed on nanoHub for high performance education use

Computational Media

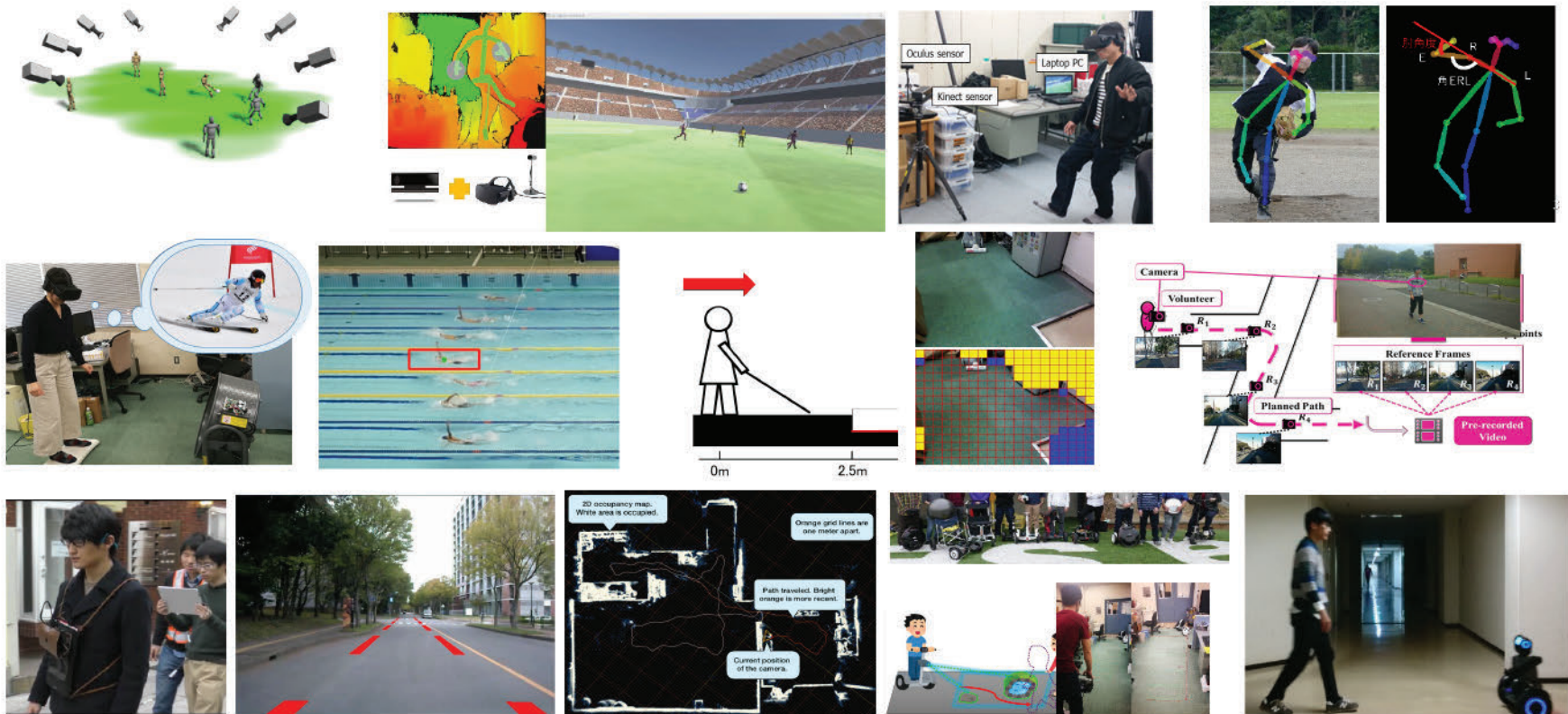


Yoshinari Kameda (University of Tsukuba, Japan)
Computer Vision and Image Media Lab.

Research background

- Computational media group
at division of computational informatics
at Center for Computational Sciences, U-Tsukuba.
- Tech: Computer vision, pattern recognition, Mixed Reality, Augmented Reality, free-viewpoint visualization by a large number of cameras
- Apps: Cyber-Medical issues, sport analysis, intelligent transport systems, empowerment of human functions, etc.

From computer vision to CHI



Towards real-world application

1. Human interface

- Instant reaction to input from human (10-1000ms)
- Sensing: multiple cameras, audios, other sensors
- Presenting: visual/auditory feedback via AR/MR

2. Continuous data analysis

1. Reinforcement learning
2. Scalability and flexibility for data input channels

Development scheme

1. Making on-line programs
 - Trial-and-error to cope with unexpected input data
 - Adaptation for new sensors and visualization devices (new cameras, HMD, etc)
2. Sharing R&D environment
 1. OS / Libraries / Docker
 2. Commonly available (transferrable) codes
 3. App level support (e.g. github, google colab, etc)

Optimized Images via ManifestList

BDEC2, Kobe Workshop



DISCLAIMER

Information subject to change

The content of this slide deck represents a expersp of the ongoing discussion within Docker, with customers and partners.

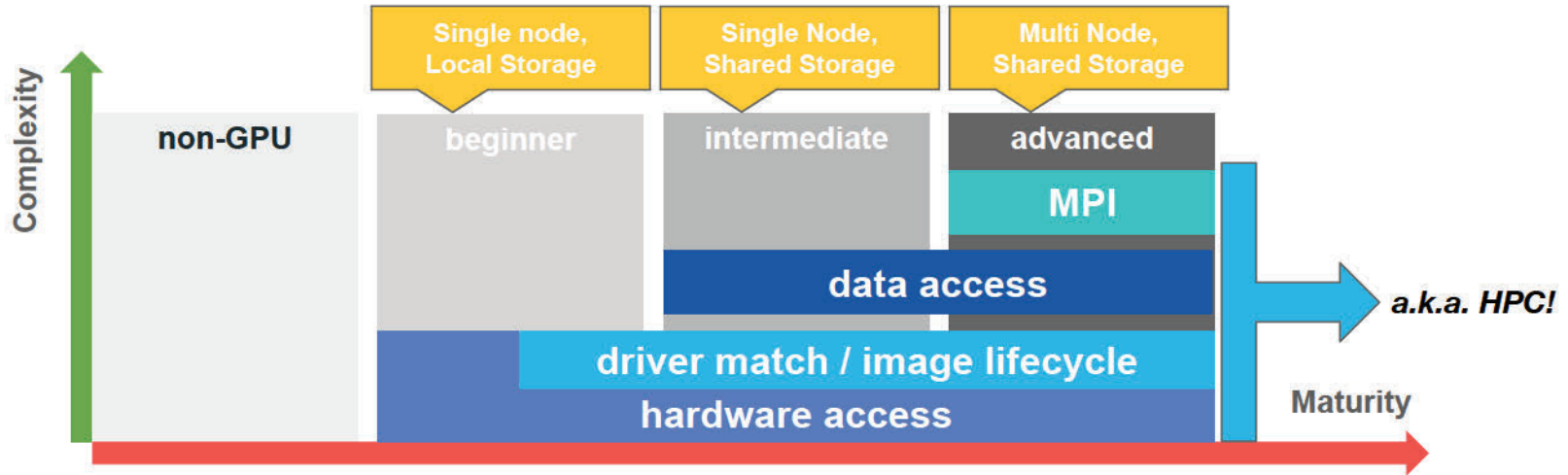
Thus, the information is subject to change and not a committed roadmap.



Convoy from GPU to HPC

In general: three levels to become a traditional HPC workload

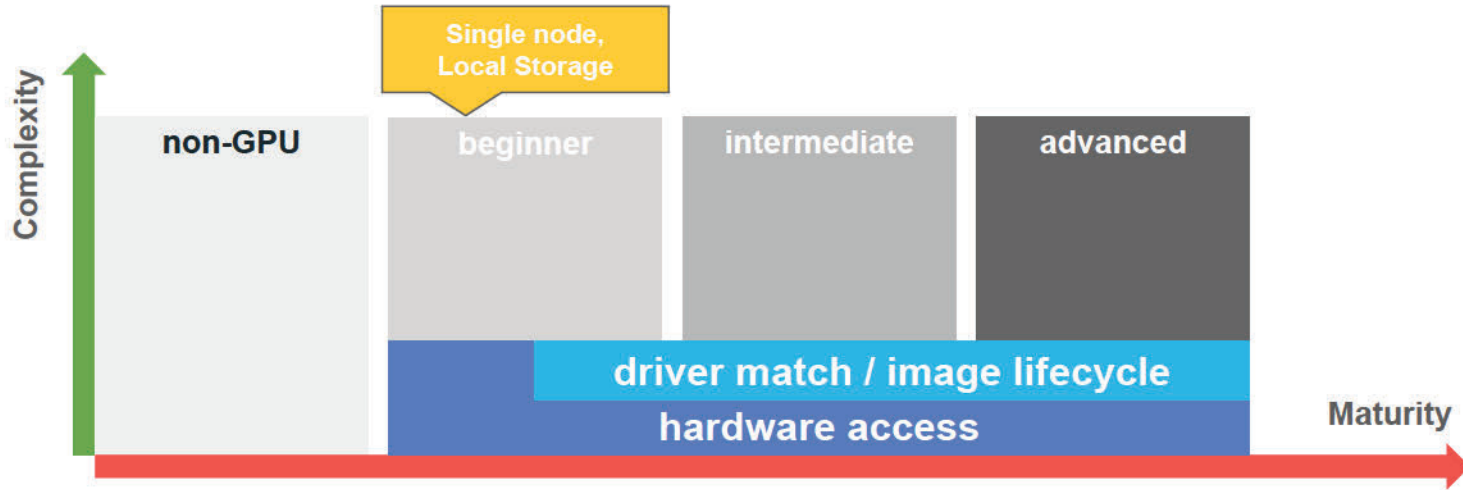
1. **beginner:** Start using GPUs with local storage
2. **Intermediate:** Control access to local mount points via POSIX enforcement
3. **advanced:** Distributed workloads using MPI



Convoy from GPU to HPC

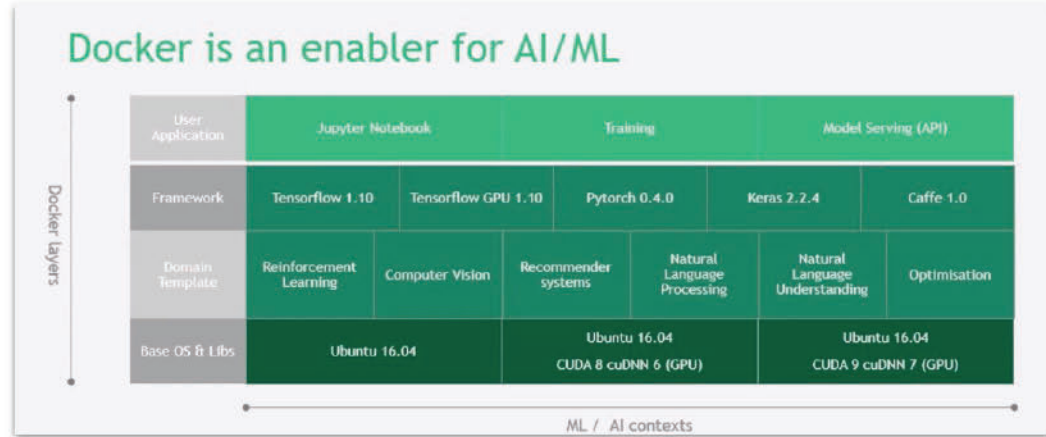
Today: Just a brief discussion about the first step and how to improve

1. **beginner:** Start using GPUs with local storage
2. **intermediate:** Control access to local mount points via POSIX enforcement
3. **advanced:** Distributed workloads using MPI

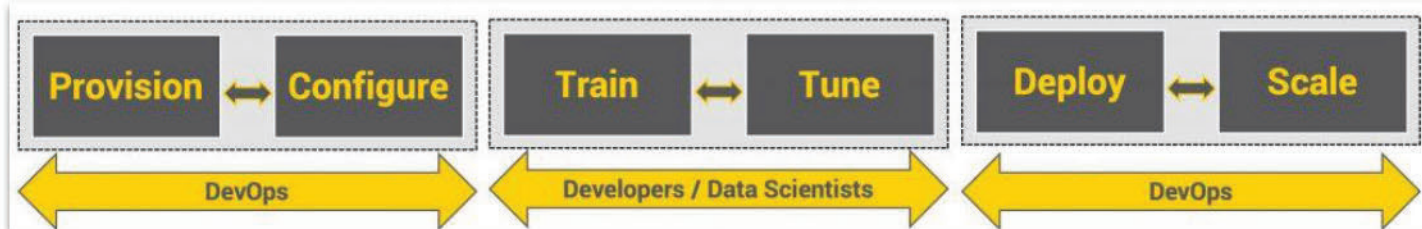


Modular Software Stacks

- Multiple phases in ML lifecycle.
- Modular Assembly of Software Stacks
- With diverse and a growing set of hardware and software, curating stacks is hard.



DockerCon EU 2018: Lessons in Using Docker to Close the Loop on Industrializing AI and ML Applications



source: <https://thenewstack.io/an-introduction-to-the-machine-learning-platform-as-a-service/>



Modular Software Stacks

- Multiple phases in ML lifecycle.
- Modular Assembly of Software Stacks
- With diverse and a growing set of hardware and software, curating stacks is hard.

Docker is an enabler for AI/ML

Docker layers	User Application	Jupyter Notebook		Training		Model Serving (API)	
	Framework	Tensorflow 1.10	Tensorflow GPU 1.10	Pytorch 0.4.0	Keras 2.2.4	Caffe 1.0	
	Domain Template	Reinforcement Learning	Computer Vision	Recommender systems	Natural Language Processing	Natural Language Understanding	Optimisation
	Base OS & Libs	Ubuntu 16.04		Ubuntu 16.04 CUDA 8 cuDNN 6 (GPU)		Ubuntu 16.04 CUDA 9 cuDNN 7 (GPU)	

Ubuntu 16.04

Ubuntu 16.04
CUDA 8 cuDNN 6 (GPU)

Ubuntu 16.04
CUDA 9 cuDNN 7 (GPU)

Provision ↔ Configure

Train ↔ Tune

Deploy ↔ Scale

DevOps

Developers / Data Scientists

DevOps



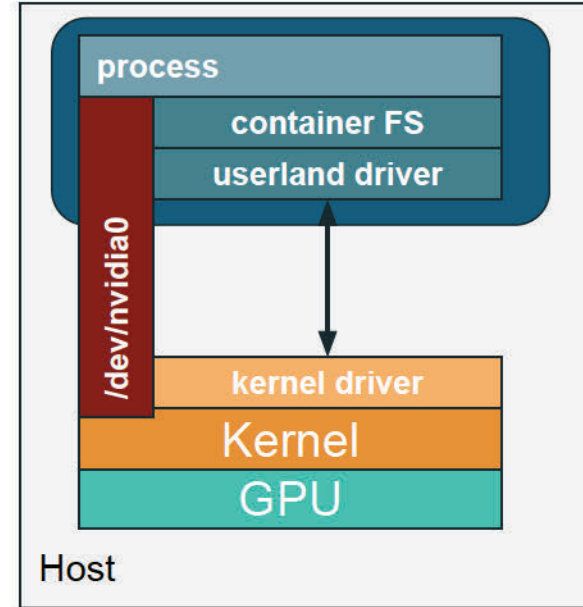
Host-Agnostic vs. Host-Specific

kernel-bypassing devices

- devices and userland drivers have to be present into a container.

The userland driver...

- can be part of the container
 - can be a local volume on the host, which adds/augments the library into the container
- These devices/drivers might be host-specific



Driver/Toolkit Mapping

Making it easy and smooth - and manageable

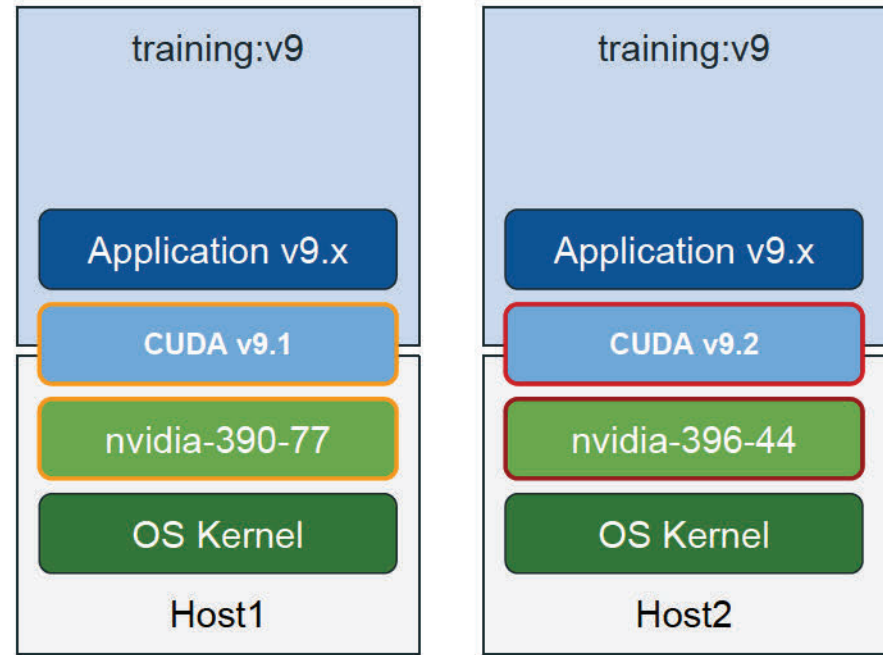


GPU Driver Match

#2 Application / User-land compatibility

Application has to deal with variants of bind-mounted userland driver.

This breaks the immutability of the container FS, forces the application to be flexible wrt the user-land version and thus puts the burden on the runtime.

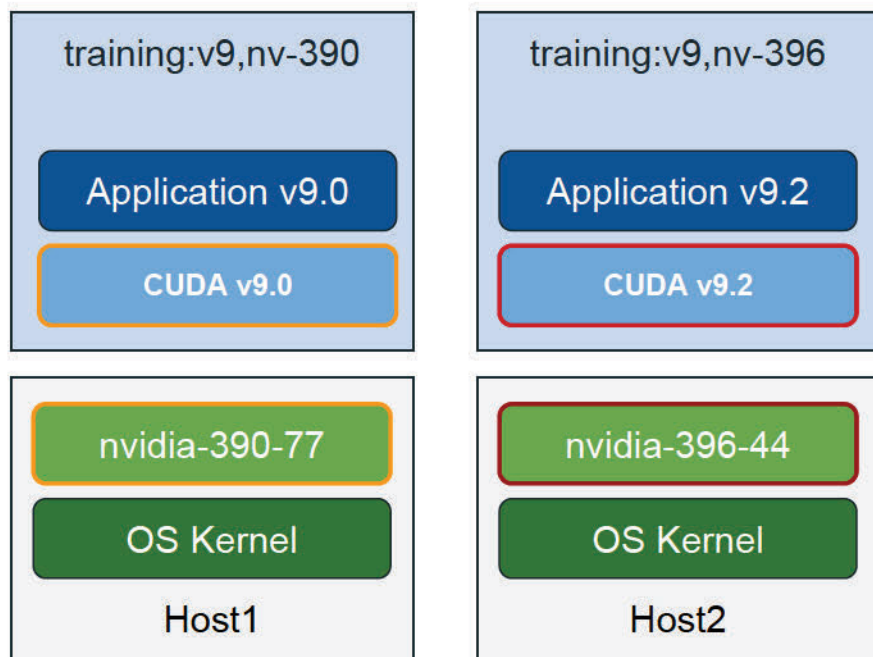


GPU Driver Match

#2 Application / User-land compatibility

Possible Solution: Using image annotations to download the exact image for the underlying configuration.

This allows to optimize down to different versions, GPU/CPU models and such. Forces customers to use proper automation (as a side effect).



System-Optimized Images

```
$ docker run --rm -ti --device=/dev/nvidia{0,ctl,-uvm} qnib/cv-tf-dev:1.12.0
Your CPU supports instructions that this TensorFlow binary was not compiled to use: SSE4.1 SSE4.2 AVX AVX2 FMA
[]
```

```
$ docker run --rm -ti --device=/dev/nvidia{0,ctl,-uvm} qnib/cv-nccl90-tf-dev:broadwell_1.12.0
libcuda reported version is: 390.30.0
kernel reported version is: 396.44.0
kernel version 396.44.0 does not match DSO version 390.30.0 -- cannot find working devices in this configurat
[]
```

```
$ docker run --rm -ti --device=/dev/nvidia{0,ctl,-uvm} qnib/cv-nccl92-tf-dev:broadwell_1.12.0
Ignoring visible gpu device (name: Tesla M60, compute capability: 5.2) with Cuda compute capability 5.2.
The minimum required Cuda capability is 7.0.
[]
```

```
$ docker run --rm -ti --device=/dev/nvidia{0,ctl,-uvm} qnib/cv-nccl92-tf-dev:broadwell_nvcap52_1.12.0
Created TensorFlow device (/job:localhost/replica:0/task:0/device:GPU:0 with 6723 MB memory)
-> physical GPU (name: Tesla M60, compute capability: 5.2)
['/job:localhost/replica:0/task:0/device:GPU:0']
```



System-Optimized Images [cont]

Using a meta-Image (Manifest List)
to use a common name.
Depending on the node-config the
node downloads the right Manifest.

```
$ sudo cat /etc/docker/daemon.json
{
  "platform-features": [
    "broadwell",
    "nv-compute-5-2",
    "nvidia-396-44"
  ], ...
}
```



```
image: qnib/cv-tf:1.12.0-rev9
manifests:
  image: qnib/cv-tf-dev:1.12.0-rev11
  platform:
    architecture: amd64
    os: linux
  image: qnib/cv-nccl92-tf-dev:broadwell_1.12.0-rev8
  platform:
    features:
      - broadwell
      - nvidia-396-44
  image: qnib/cv-nccl92-tf-dev:broadwell_nvcap52_1.12.0-rev2
  platform:
    features:
      - broadwell
      - nv-compute-5-2
      - nvidia-396-44
```



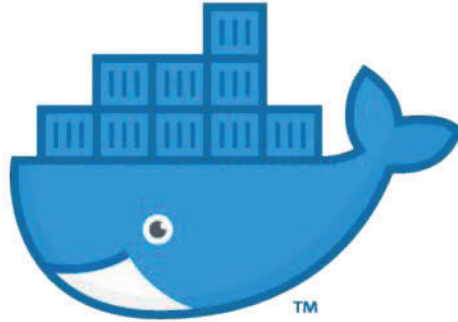
Implications

HW and OS Optimizations

This scheme can not only be employed for different GPU drivers, but

- highly optimized code for CPUs, Storage (gcc flags are the limit)
- No generic container images across multiple host configurations, stable image even if new nodes are added or old are removed
- Needs to be evolved into decision tree
 - a. Try to fetch most specific image
 - b. Iterate to generic image if no specific image is available





THANK YOU :)

More Info:

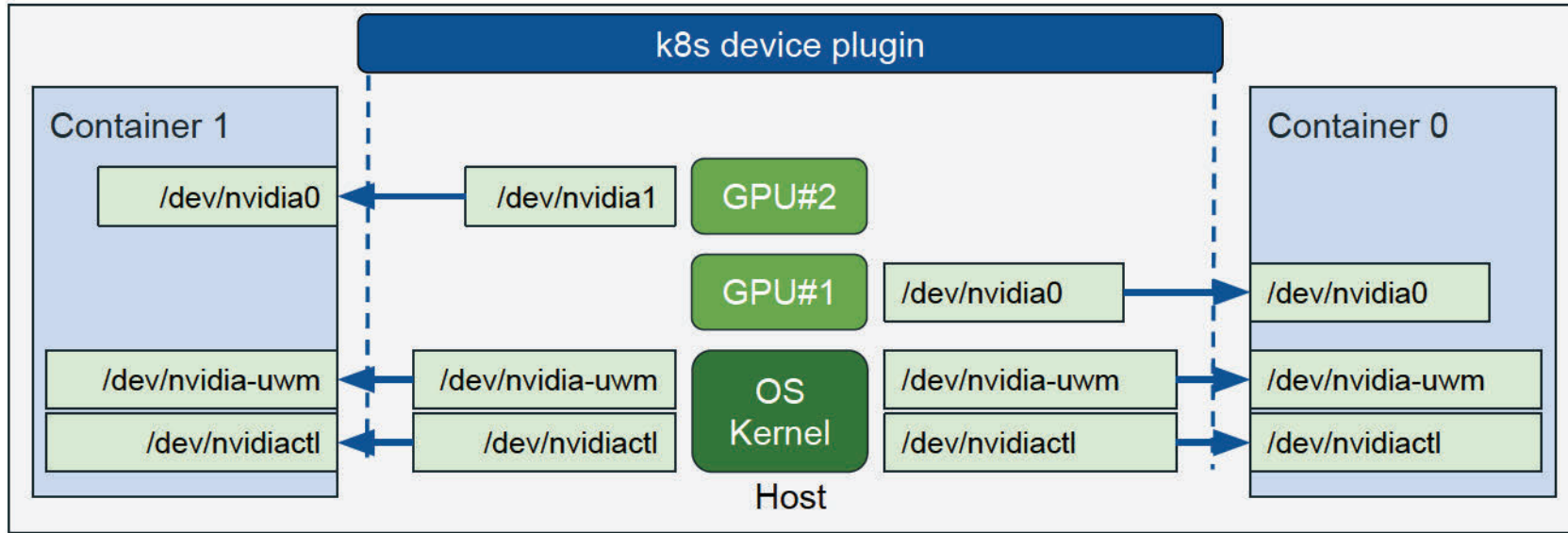
- <http://qnib.org/2019/02/14/manifest-list-to-pick-optimized-images/>
- <https://github.com/moby/moby/issues/38715>

Device Access

#1 Device Passthrough

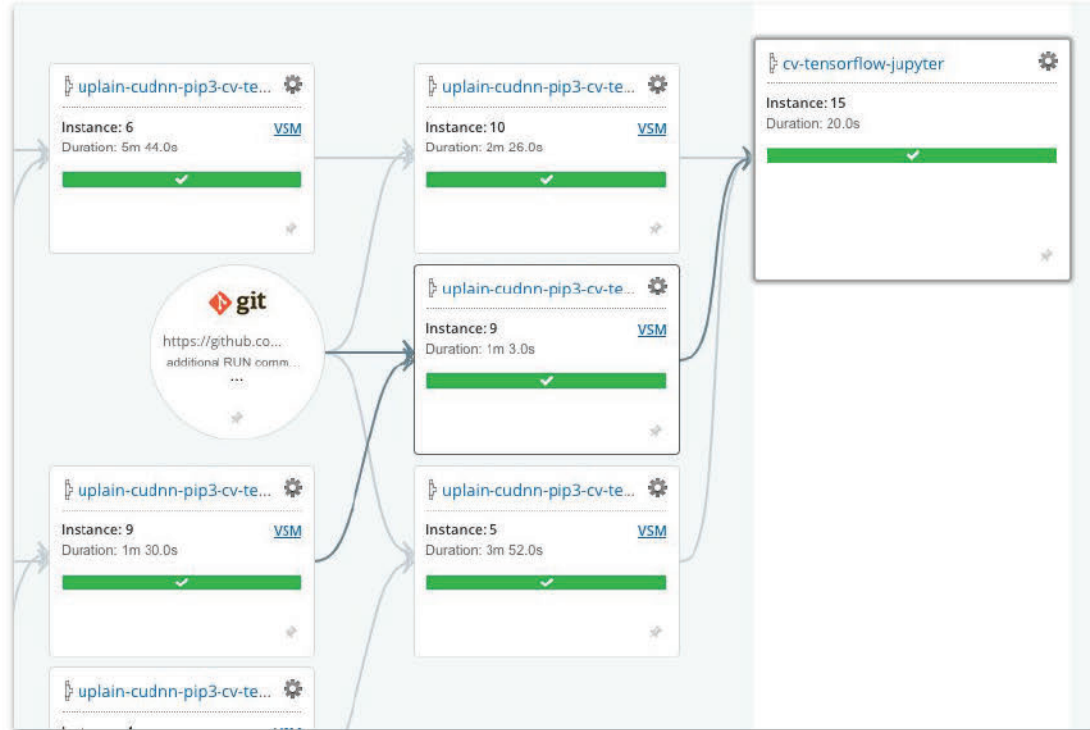
Device Passthrough

- Available GPUs are mapped in, depending on scheduler
- Auxiliary devices are mapped in, depending on use-case



CI/CD is King

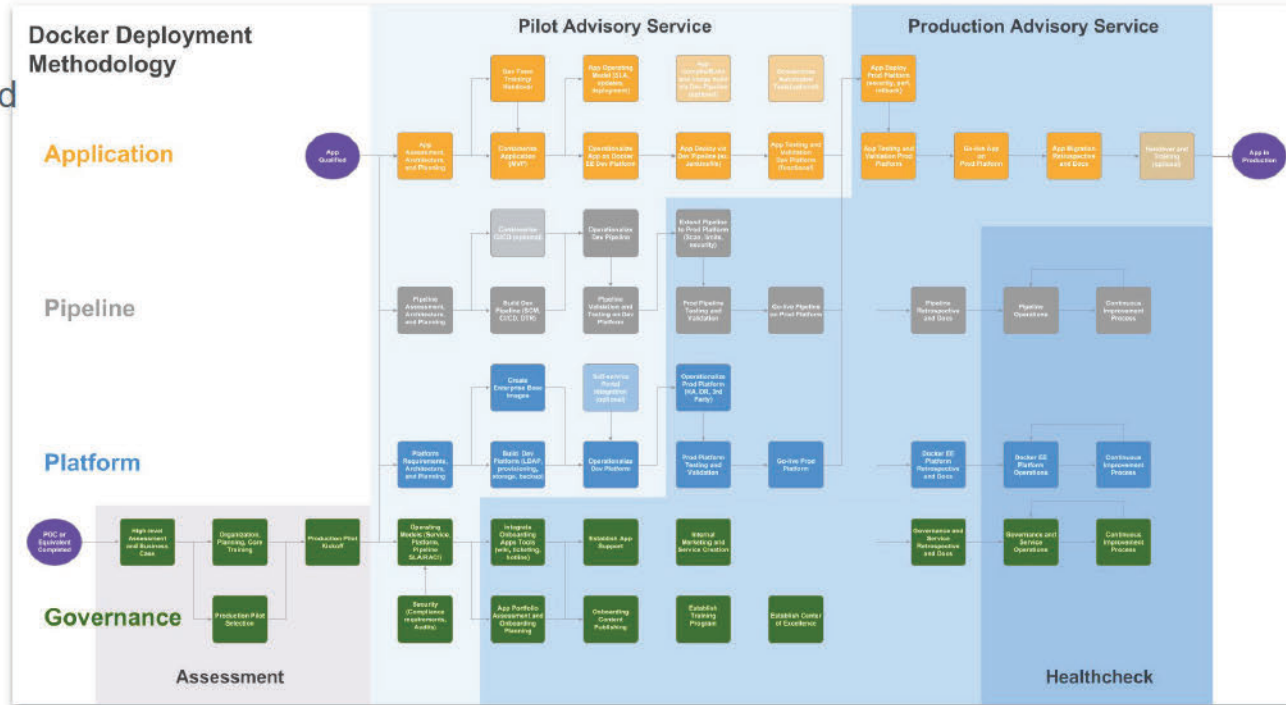
- build Images
 - TF w/ CUDA 9.0
 - TF w/ CUDA 9.2
 - TF w/ CUDA 10
- Merge into ManifestList
 - qnib/cv-tensorflow-jupyter



Docker Deployment Methodology

The Docker Deployment Methodology is prescriptive and comprehensive process enabling customers to deploy Docker Enterprise Edition at scale to run production workloads

The Methodology was developed directly from 3+ years of operationalizing containers with enterprise customers



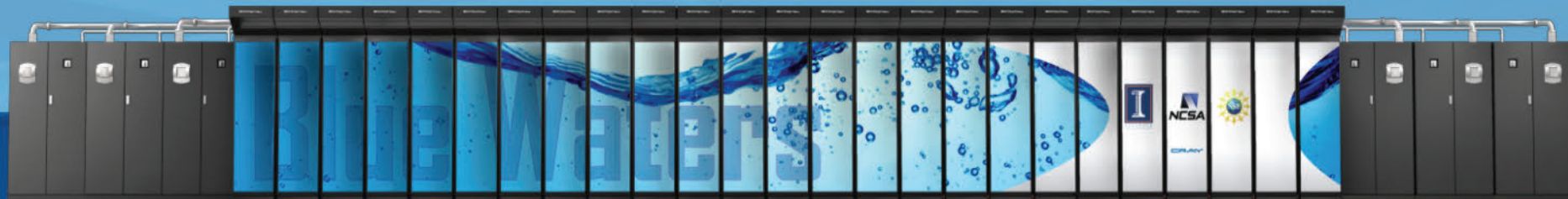
BLUE WATERS

SUSTAINED PETASCALE COMPUTING

Geospatial and Global Earth Mapping and Modelling Applications Requirements for the BDEC2 Workshop

William Kramer, University of Illinois at Urbana Champaign

with contributions from Paul Morin, Jonathan Pundsack, Claire Porter and others at the Polar Geospatial Center at the University of Minnesota, and Kiayu Guan, Brett Bode and Greg Bauer, NCSA at University of Illinois at Urbana Champaign and Phil Maechling and Cristine Goulet from the Southern California Earthquake Center at the University of Southern California.



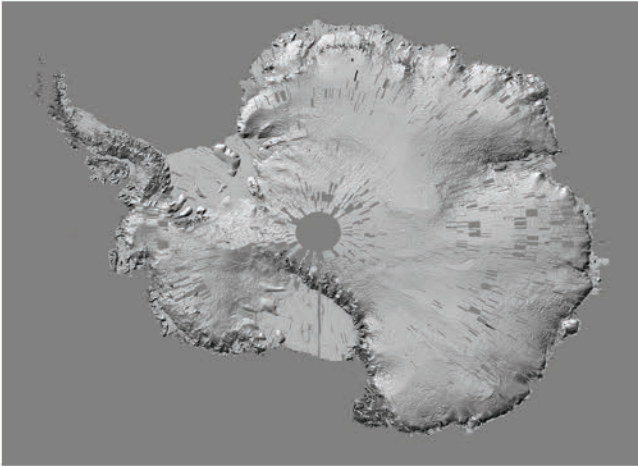
GREAT LAKES CONSORTIUM
FOR PETASCALE COMPUTATION

CRAY

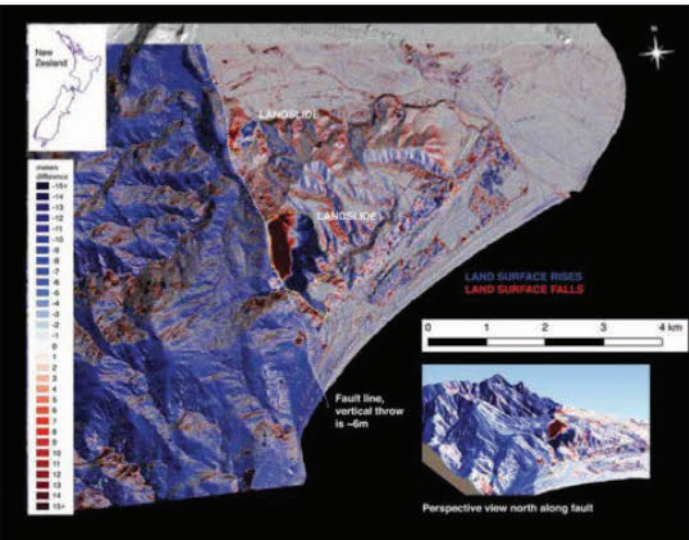
- In the last few years, using leadership computing, new best of breed approaches for creating Geo data, model and mapping (DMM) products created paradigm shifting products that many other use.
 - Digital Surface and Digital Elevation maps
 - Seismic Hazard Maps
 - Crop planting and yield maps
 -
- Example - The digital elevation map creation capabilities greatly reduce the cost and improves the timeliness and resolution of traditional map and been shown to be 58,000 times more productive (faster, cheaper)
 - The improvements include increasing the resolution of current elevation maps by more than 3 orders of magnitude (12.5^2 improvements in resolution),
 - improving the time to production 58,500x in time to solution compared to a single workstation, and
 - a 220 times reduction in cost, resulting in 9 orders of magnitude of overall productivity improvement.
- The consumers have grown increasingly reliant on timely products
- Initial this was a research investigation to create applications that would prove effective but also had a serendipitous benefit of creating use DMM products that other areas use.
 - Water resource managers, environmental research, building engineers, farmers,

	Strips	Area – km ²	Estimated BW node-hours required to process one time	Data Required (PB)
In-track	336,492	538 million	54 million	5.38
Cross-Track	2,956,131	2,365 million	473 million	47.30
Total	3,292,623	2,903 million	527 million	52.68

- The computational needs to create a single, 2 meter global set of DEMs one time is
 - 527 million Blue Waters X86 node hours – 3.2-3.5 Blue Waters Years
 - .5 meter resolution is 16 time the computational requirement – 51.2-56 Blue Waters years
- Requirement is to redo this processing every 2-3 year, with more frequent processing in high interest areas



- Each stripe averages 4 GB in size, and two strips are needed for every DEM.
- Two meter DEMs average 8 GBs.
- So, the processes consumes 8 GB and produces 8 GB per sample.
- One time world map is requires 26-30 PB is consumed and 26-30PB is produced.
- Since the original strip data flows from repositories specific for the satellites, and is stored in open access repositories, these 50+ PBs of data has to move within the period of the campaign.
- If you assume this is a yearly campaign, the average sustained data rates are ~10 kbps, but will have peaks where multiple streams of 8 GBs need to move before or after a job initiated.

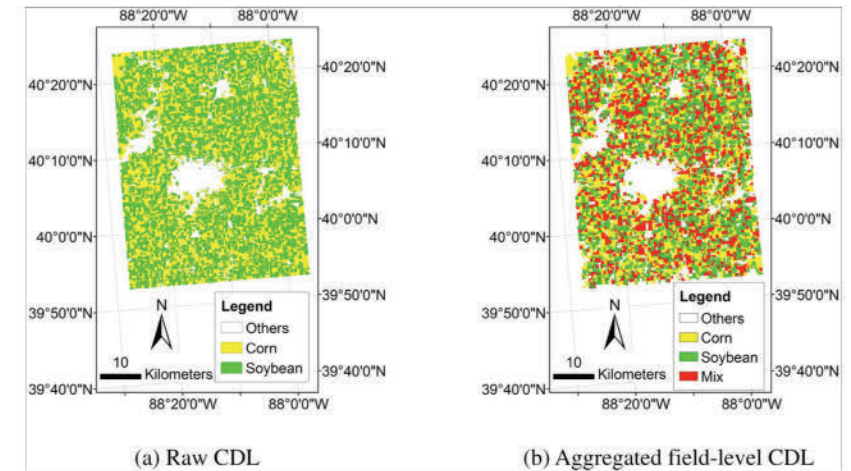
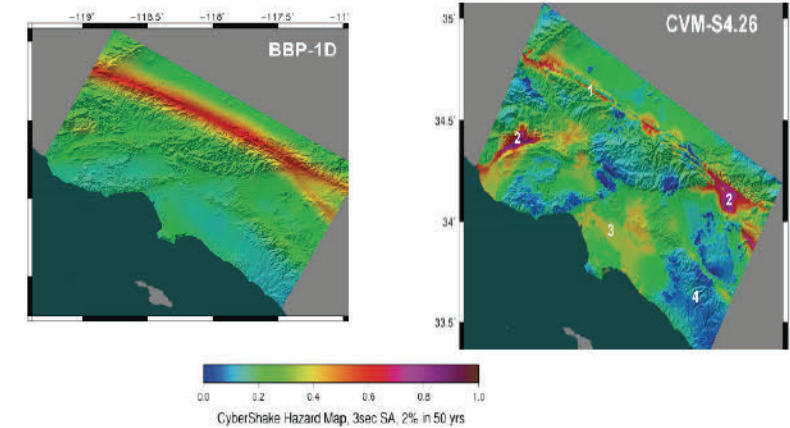


• **Seismic**

- Develop physics-based earthquake simulations that are more accurate than the current empirical NSHMP standard
 - Challenges – resolution, shaking frequency proportional to structure height, structure analysis
- Many engineering and social applications: performance-based design, seismic retrofitting, resilience engineering, insurance rate setting, disaster preparation and warning, emergency response, and public education
- Requirements same order as DEMs, more areas (entire Pacific Northwest, data, improvements)

• **Agriculture**

- Developed modeling framework which combines the strengths of earth system model and agronomy crop model to conduct parameter sensitivity analysis and spatially explicit optimization
- Using three different cuDNN-accelerated deep learning frameworks with of satellite data at a 30-meter resolution, achieve well over 95% accuracy.



Example of the 2015 CDL of Champaign County. (a) Raw CDL; (b) the aggregated field-level CDL, where the CLU is used to provide the field-level boundaries.

- Codes and methods that were research first time frontier science have now moved to best of breed sustaining production
 - Workflows are established to being established.
- New versions of the codes with improvements (e.g. high frequency for seismic modes, better identification for crop yields, etc.)
- But the success of frontier science has made DMMs for these areas now required and expected by a broad, diverse communities

The background features a large, faint watermark of the Shanghai Jiao Tong University logo. The logo is circular, containing a gear, a book, and a scale, with the year '1896' at the bottom. The text 'SHANGHAI JIAO TONG UNIVERSITY' is written around the perimeter, and Chinese characters '上海交通大学' are at the top.

Benchmarking Huawei ARM Server Processor for HPC Workloads

James Lin

Shanghai Jiao Tong University (SJTU), Center for HPC

BDEC2, Kobe, JP
February, 2019



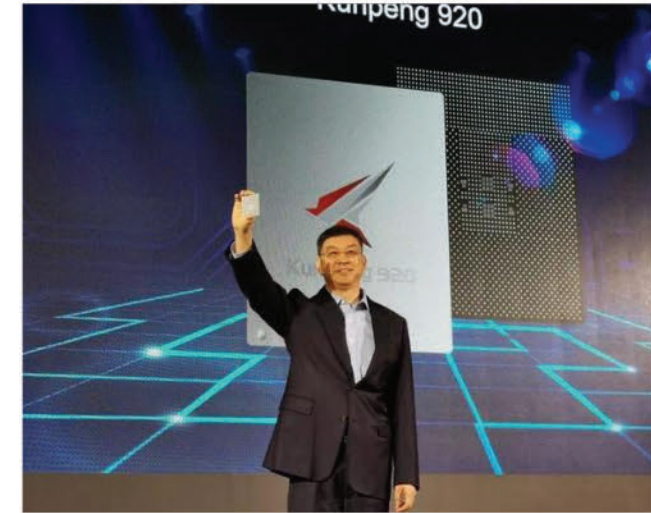
The First ARM HPC Innovation Center in China (hosted in SJTU)



- Supported by major ARM vendors in China, including Huawei, Phytium/NUDT

Huawei Kunpeng 920 ARM Sever Chip

Model	Intel Xeon Gold 6148	Hi1616	Kunpeng 920 (Engineering Sample)
Arch	Skylake-SP	ARMv8-A	ARMv8.2-A
Lithography	14nm	16nm	7nm
Main Frequency(GHz)	2.4	2.4	2.0
Num of Cores	20	32	48
Vectorization Ins/Width	AVX512/512bits	ASIMD/128bits	ASIMD/128bits
Theoretical DP Peak Performance (GFLOPS)*	1536	307.2	768
L1 Cache	32KB Ins + 32KB Data	48KB Ins + 32KB Data	64KB Ins + 64KB Data
L2 Cache	1 MB	1MB (shared)	512KB
L3 Cache	1.375 MB	32MB (shared)	64MB (shared)
DRAM Support	6 x DDR4-2666	4 x DDR4-2400	8 x DDR4-3200
TDP	150	70	150
Launch Time	2017	2016	2019



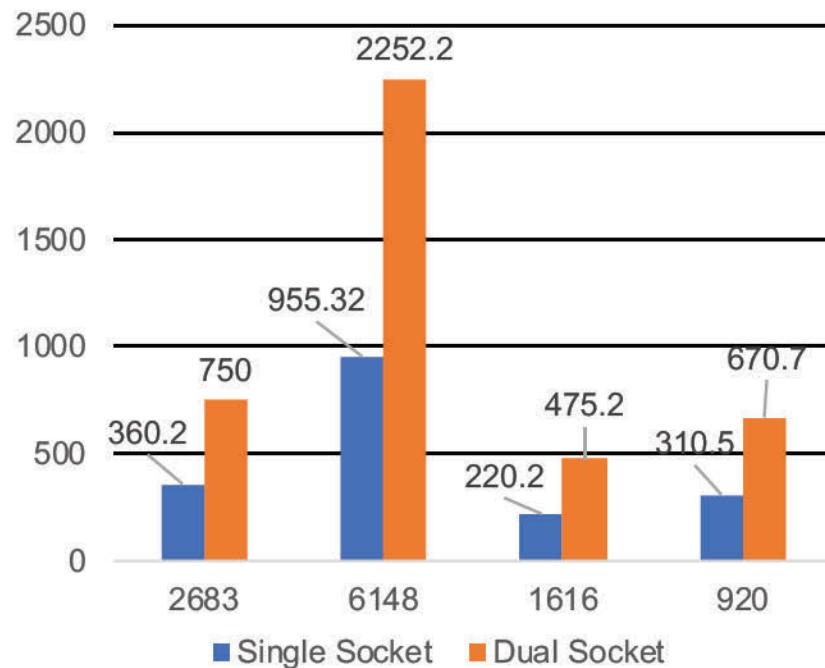
Announced on Jan 7, 2019

* Theoretical DP peak performance is calculated based on the frequency we test during chips running their best vectorization instruction set.

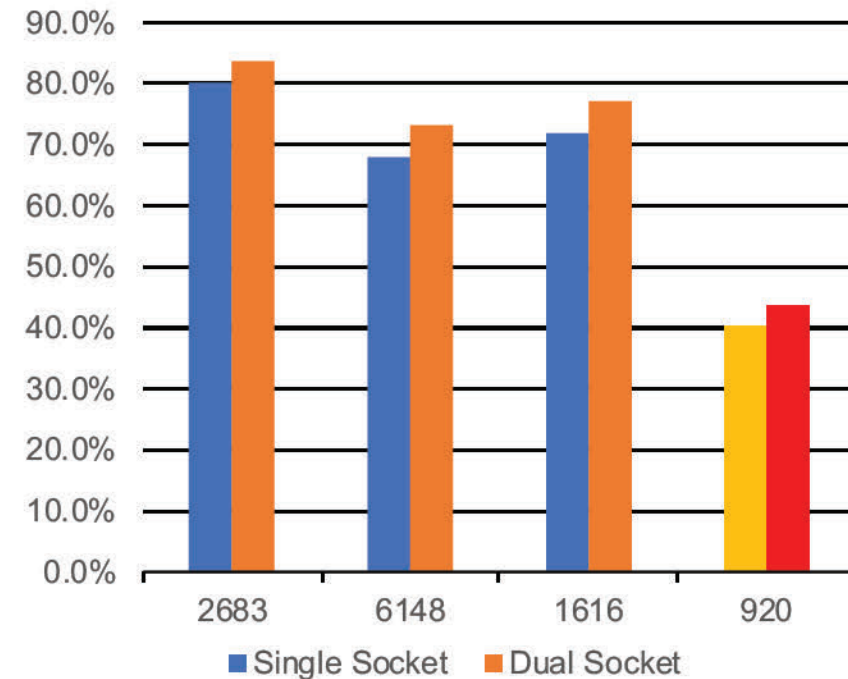
Float-point Results

- 41.1% Better than Hi1616, compared to a 165.3% increase from Haswell to Skylake in 3 years.
- HPL efficiency on Kunpeng 920 is around 40% compared to more than 70% on other chips.

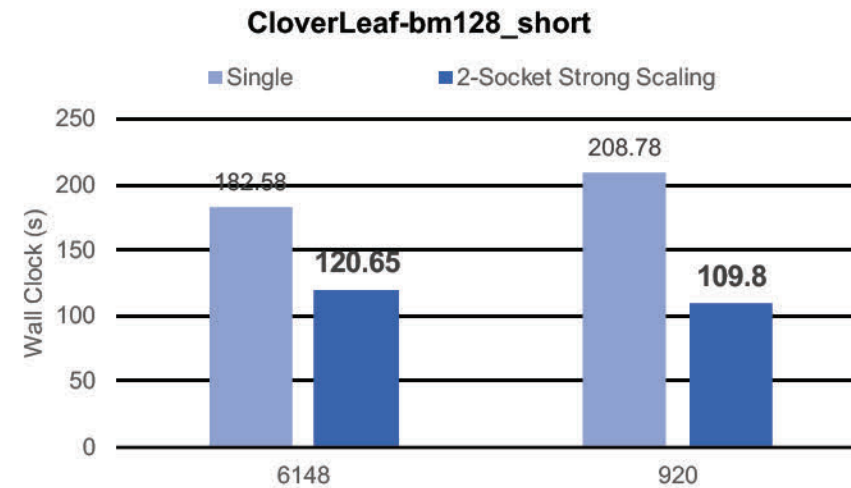
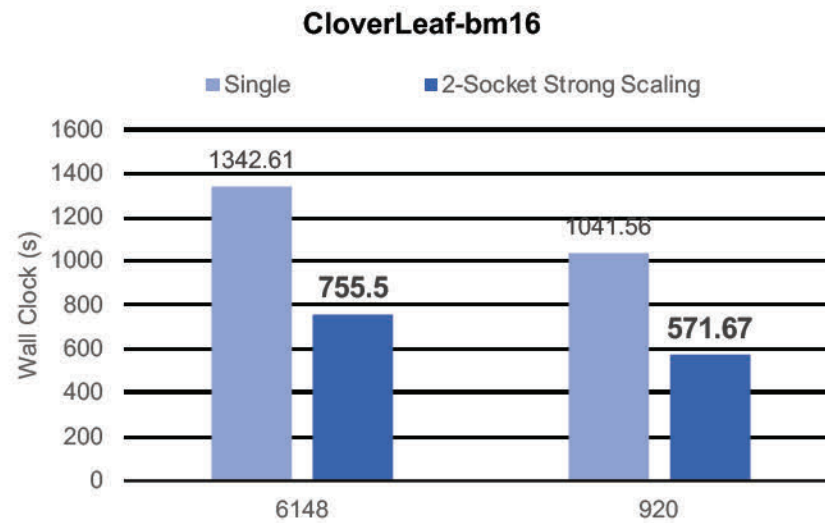
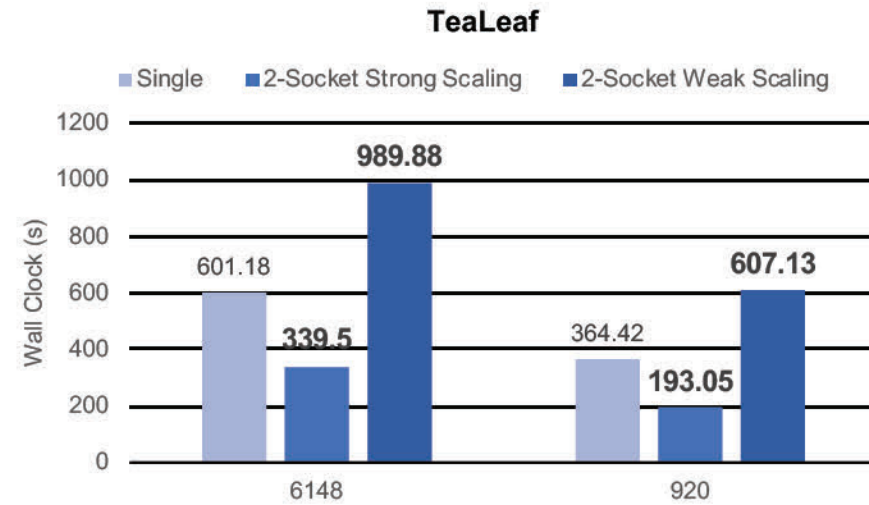
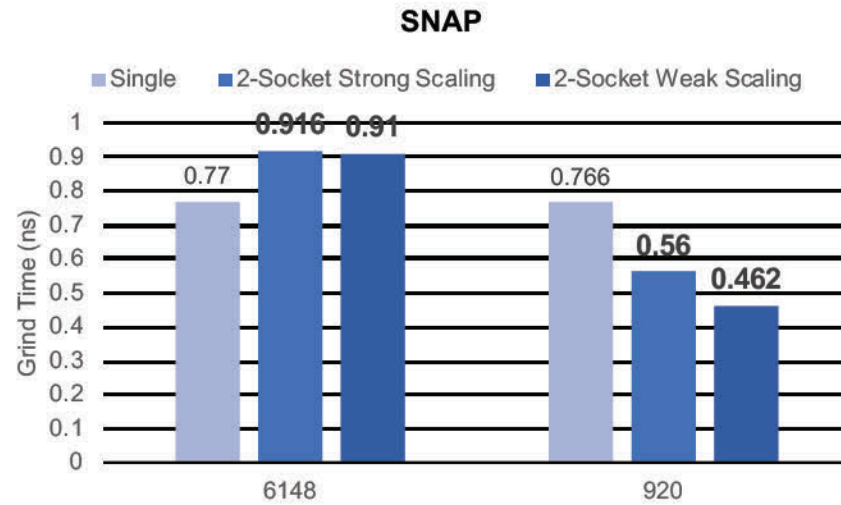
HPL Benchmark on Four Platforms



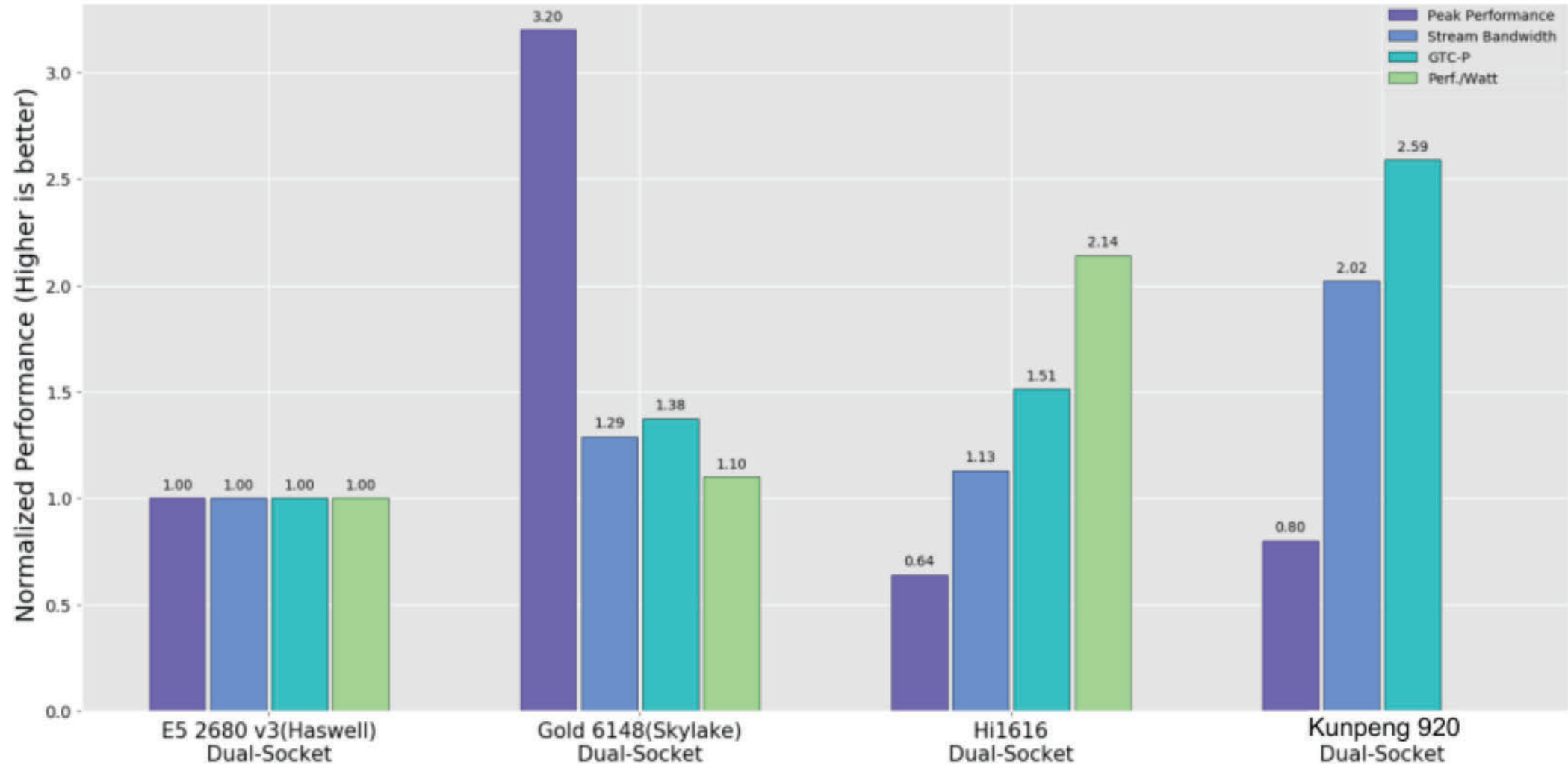
HPL Efficiency on Four Platforms



Mini-apps Results (lower is better)



Real Application Result: GTC-P (Gyrokinetic Toroidal Code - Princeton)

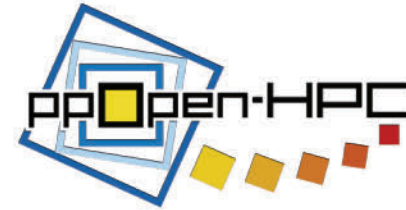




東京大学
THE UNIVERSITY OF TOKYO



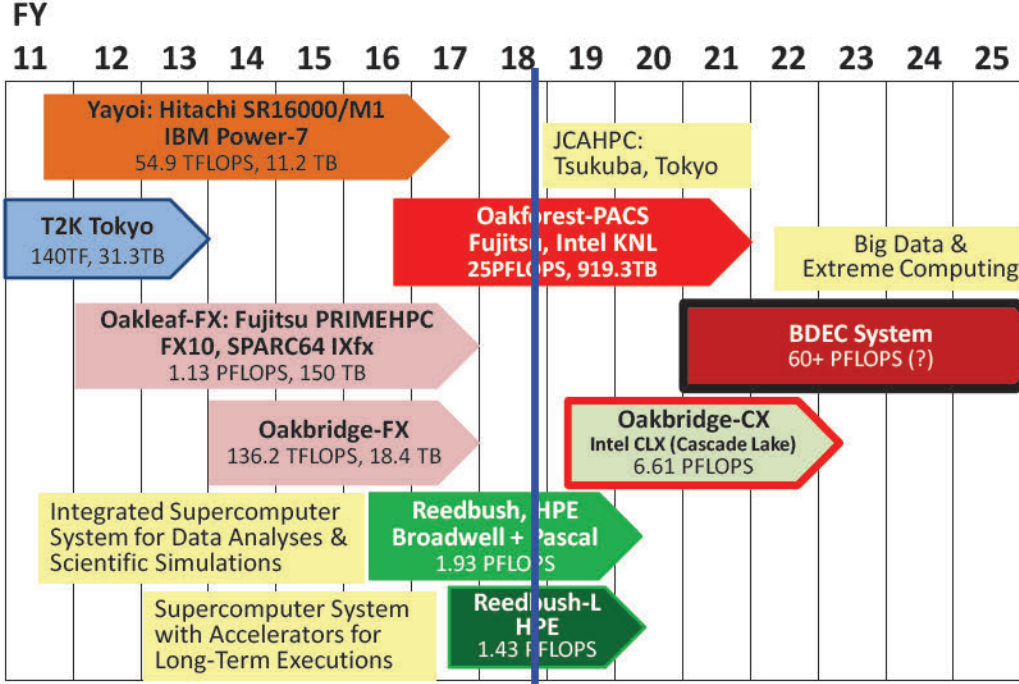
東京大学情報基盤センター
INFORMATION TECHNOLOGY CENTER, THE UNIVERSITY OF TOKYO



Innovative Method for Integration of Simulation/Data/Learning in the Exascale/Post-Moore Era

Kengo Nakajima
Information Technology Center
The University of Tokyo

BDEC2 Workshop, Kobe, Japan
February 19-21, 2019

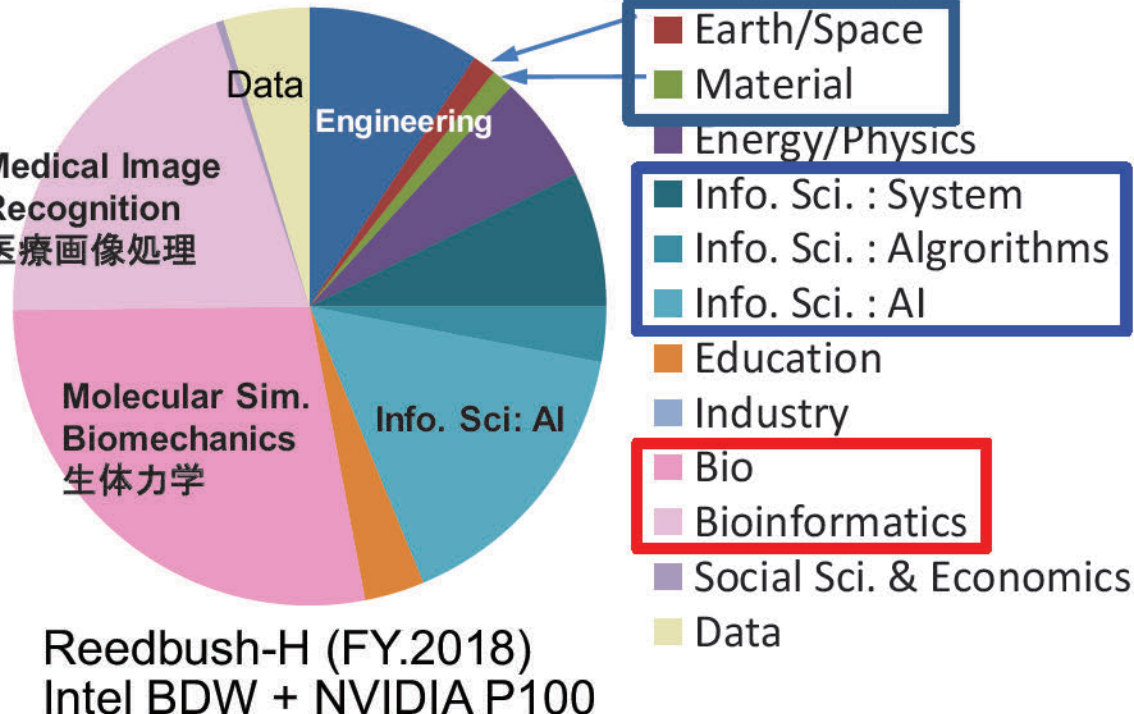
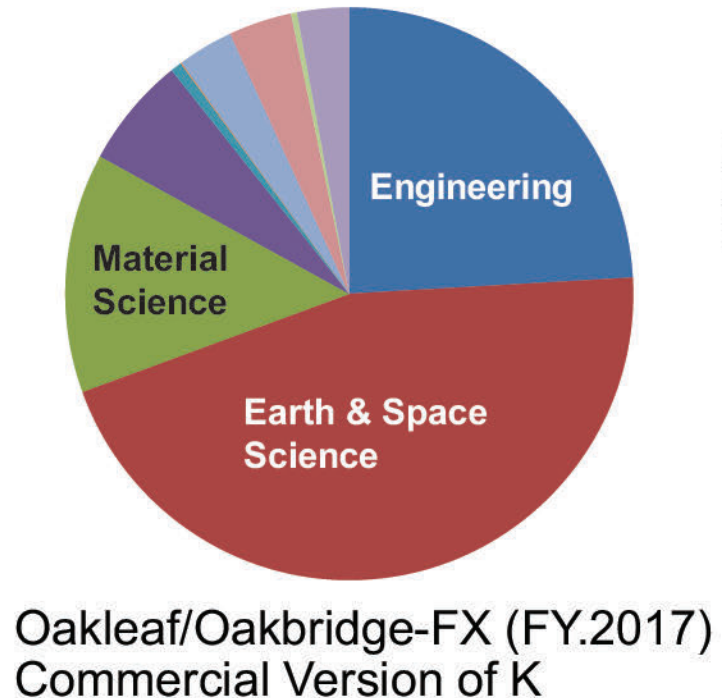


New Types of Users

- Mostly CSE, so far
- Data, ML, AI etc.
 - Genome Analysis
 - Medical Image Recognition

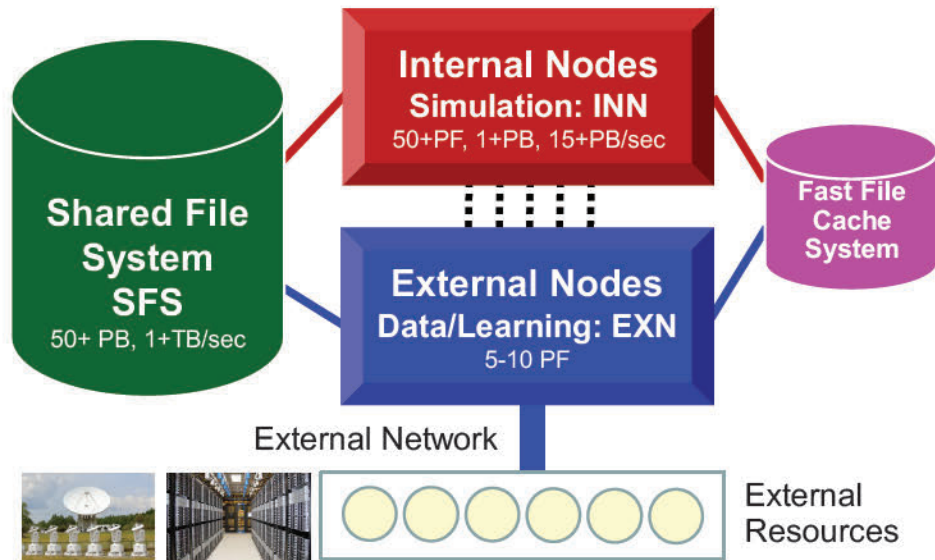
New Methods

- Integration of CSE (Simulations) + Data + Learning



BDEC System at ITC/U.Tokyo¹¹⁷

- April 2021
- Platform for Simulation + Data + Learning (S+D+L)
- 60+ PF, 3.5-4.5 MW
 - External Nodes for Data Acquisition/Analysis (EXN)
 - 5-10 PF, 200+ TB
 - Internal Nodes for CSE/Data Analysis (INN)
 - 50+ PF, 1+ PB, 15+ PB/sec.
 - Shared File System (50+PB, 1+TB/sec) + File Cache
- Architectures of EXN and INN could be different
 - EXN could include GPU, FPGA, Quantum Device



- Possible Applications
 - Atmosphere-Ocean Simulations with Data Assimilation
 - Real-Time Disaster Sim. (Flood, Earthquakes, Tsunami)
 - Earthquake Simulations with Data Assimilation
 - Data Driven Approach

Real-Time Earthquake Simulation with Data Assimilation

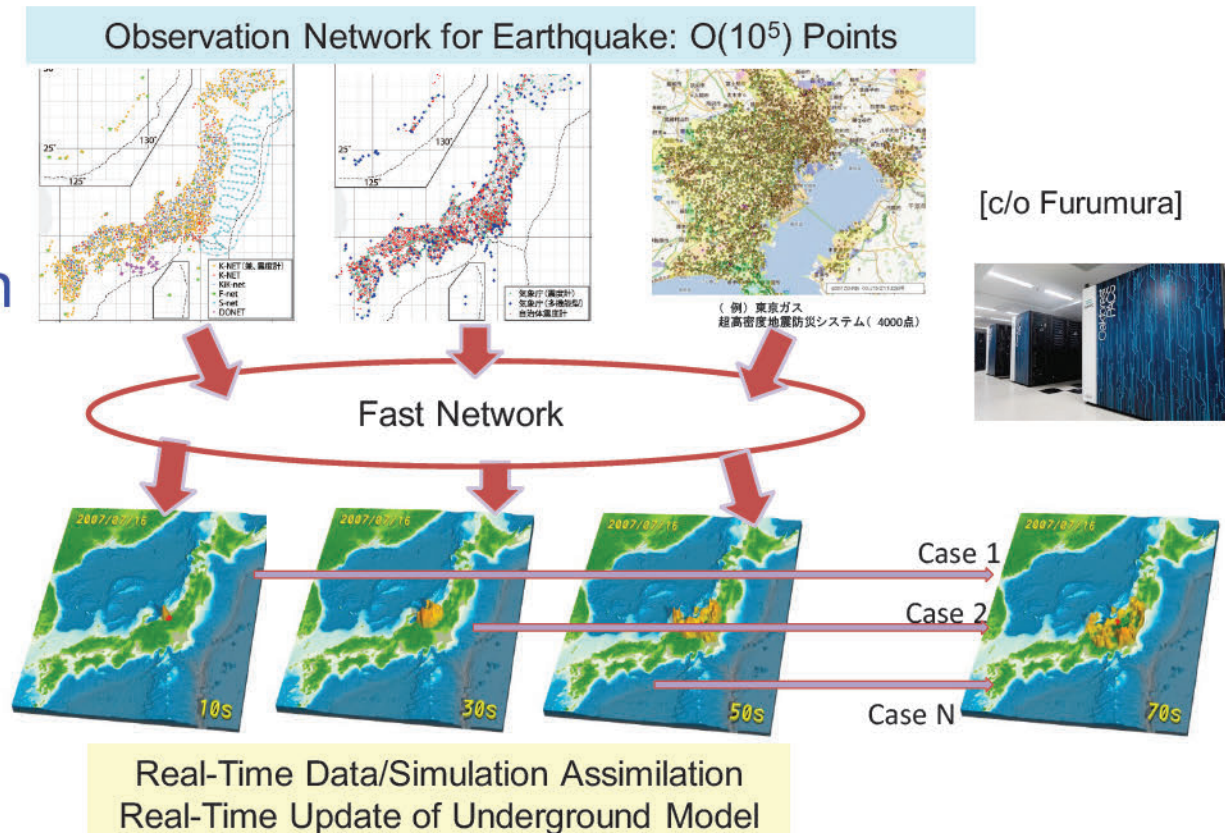
- Seismic Observation Data (100Hz/3-dir's/O(10³) pts) by JDXnet is available through SINET in Real Time
 - Peta Server in ERI/U.Tokyo: O(10²) GB/day ⇒ EXN of BDEC
 - O(10⁵) pts in future including stations operated by industry

- External Nodes

- Real-Time Data Acquisition
- Data Assimilation
- Update of Underground Model

- Internal Nodes

- Large-Scale Multiple Simulations



Data Driven Approach

DDA, Integration of (S+D+L)

- Real-World Simulations

- Non-Linear: Huge Number of Parameter Studies needed

- Reduction of cases is a very crucial issue

- Data Assimilation

- Mid-Range Weather Prediction: 50-100 Ensemble Cases, 1,000 needed for accurate solution.
- 50-100 (or fewer) may be enough for accurate solution, if opt. parameters are selected (e.g. by ML),

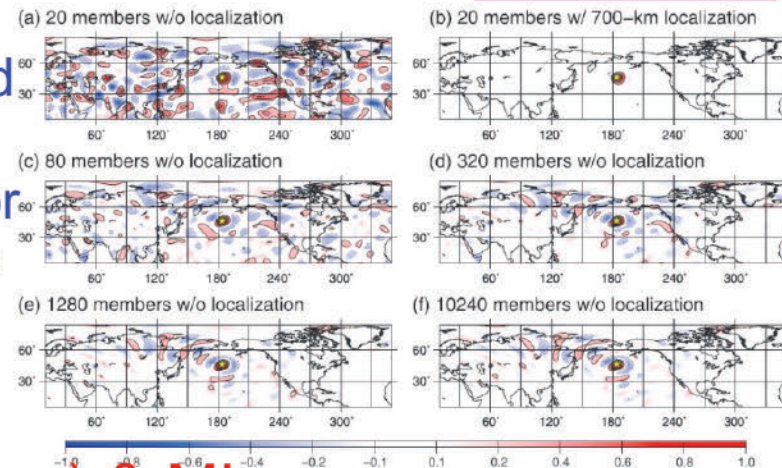
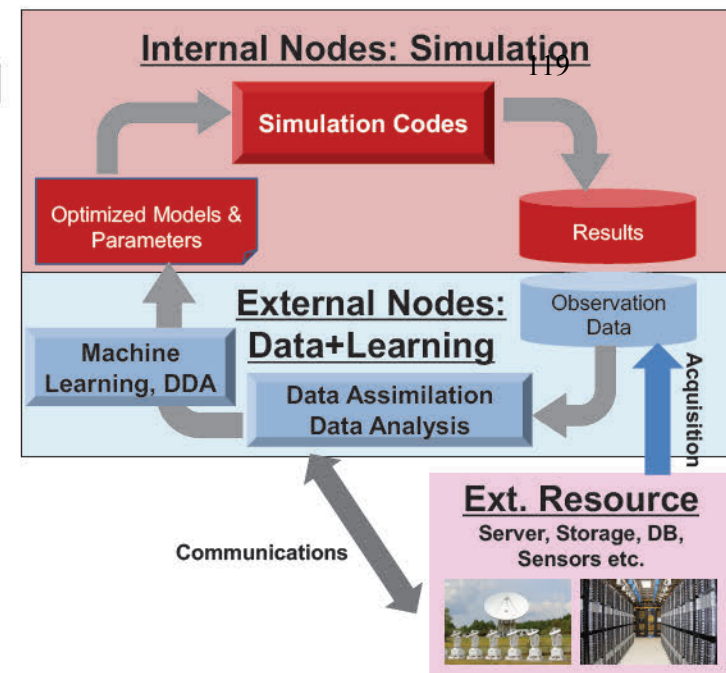
- Data Driven Approach (DDA)

- Integration of CSE & (Observation) & ML

- $O(10^3-10^4)$ Training Data Sets: Difficult

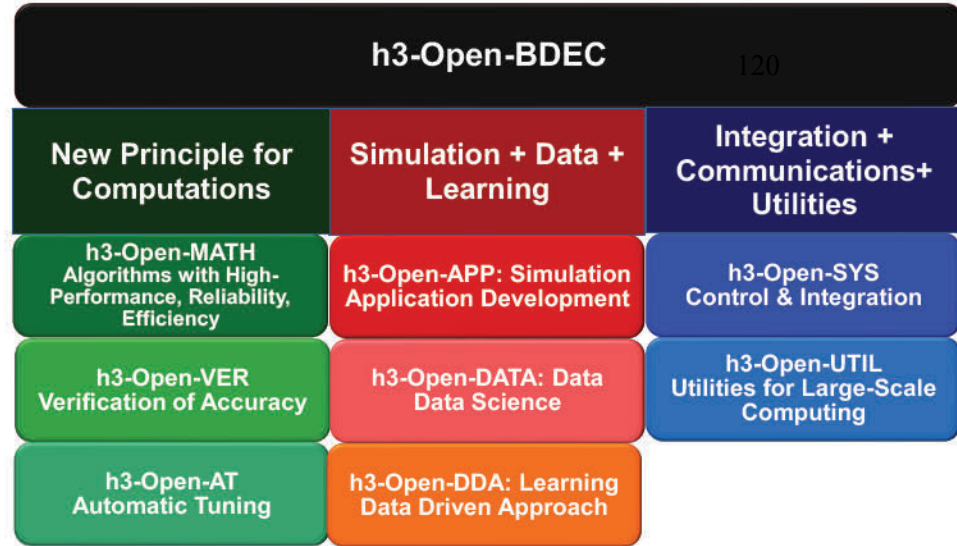
- Successful under Only Limited Conditions using Simplified Models

- hDDA: Hierarchical DDA by More Efficient Training Approach

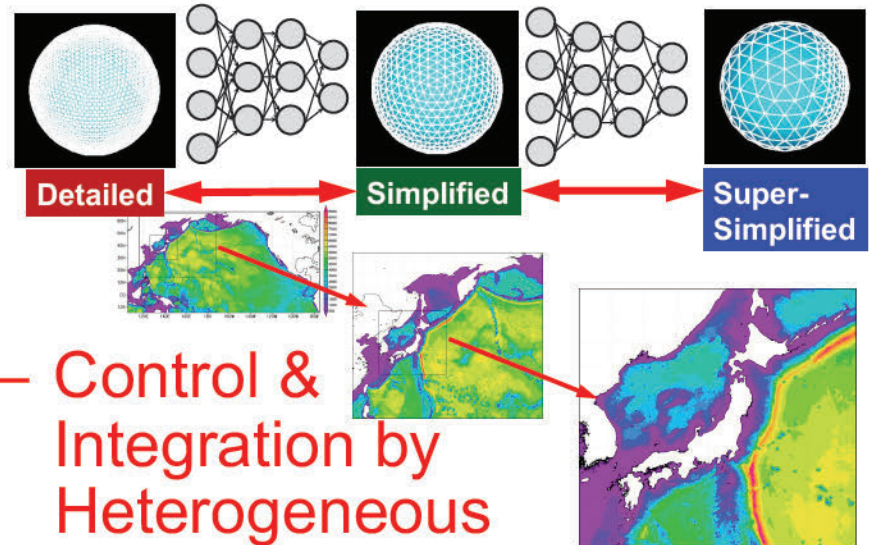


[Miyoshi et al. 2014]

h3-Open-BDEC



- Innovative Software Infrastructure for (S+D+L)
 - h3: Hierarchical, Hybrid, Heterogeneous
- Innovative/New Ideas
 - Adaptive Precision + Accuracy Verification + AT
 - Appropriate Computing
 - hDDA for General Problems by Machine Learning -> Reduction of Computations
 - Generation of Simplified/Local/Surrogate Model by ML
 - Multilevel/Multi-nested Approach using AMR
 - MOR (Model Order Reduction)
 - UQ (Uncertainty Quantification)



- Control & Integration by Heterogeneous Containers
 - Various Functions on Heterogeneous Architectures
 - Including CPU, GPU, FPGA, Quantum Devices

Farming the Environment into the Electrical Power Grid

- Offshore Wind Farm Simulation & Control



Dr. Yiwei Qiu
ywqiu@mail.tsinghua.edu.cn
BDEC2-Kobe
Feb. 20th, 2019

Smart Grid Operation and Optimization Laboratory

- Smart Grid Operation and Optimization Laboratory (SGOOL) was established in Tsinghua and Zhejiang in 2009 and 2012 respectively
- 2 Professors, 4 Associate Professors, 1 Assistant Professor, 5 Postdoctoral Research Fellows, about 20 Ph.D. Students and 30 M.S. Students



Tsinghua University



Dr. Yonghua Song
Professor, Tsinghua University/Zhejiang University
Rector of Macau University from Jan, 2018



Dr. Zechun Hu
Associate Professor, Tsinghua University
Research Interests: Electric Vehicle, Energy Storage



Dr. Jin Lin
Associate Professor, Tsinghua University
Research Interests: Industrial Microgrid, Active Distribution Network, Hydrogen EDN



Zhejiang University



Dr. Yi Ding
Professor, Zhejiang University
Research Interests: Power System Reliability, Market



Dr. Hao Wu
Associate Professor, Zhejiang University
Research Interests: Power System Operation/Security



Dr. Shufeng Dong
Associate Professor, Zhejiang University
Research Interests: State Estimation, Electrical Information



Dr. Can Wan
Assistant Professor, Zhejiang University
Research Interests: Renewable Prediction, Optimization

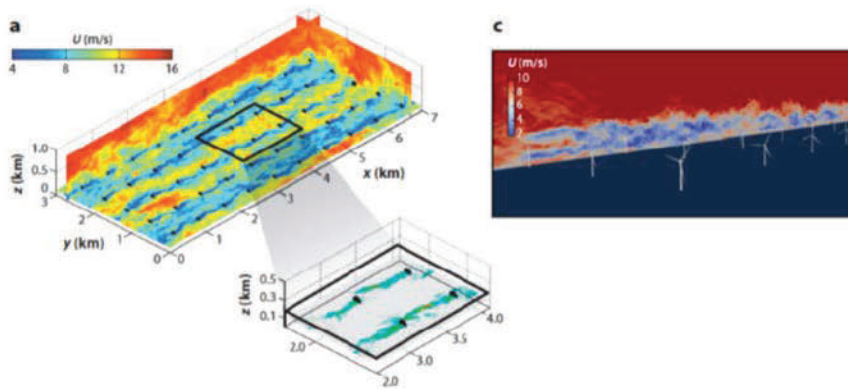
Objective of the FENGBO Project ¹²³

Farming the Environment into the Grid: Big data in Offshore Wind



- Complex **aerodynamic interaction** among turbines caused by the **wake effect**
- **Electromechanical interaction** among turbines through the power grid
- Investigate the interactions and **coordinated control** among different wind turbines
- Funded by Natural Science Foundation of China

Figures from Goit and Meyers [2015], Churchfield et al. [2012b]

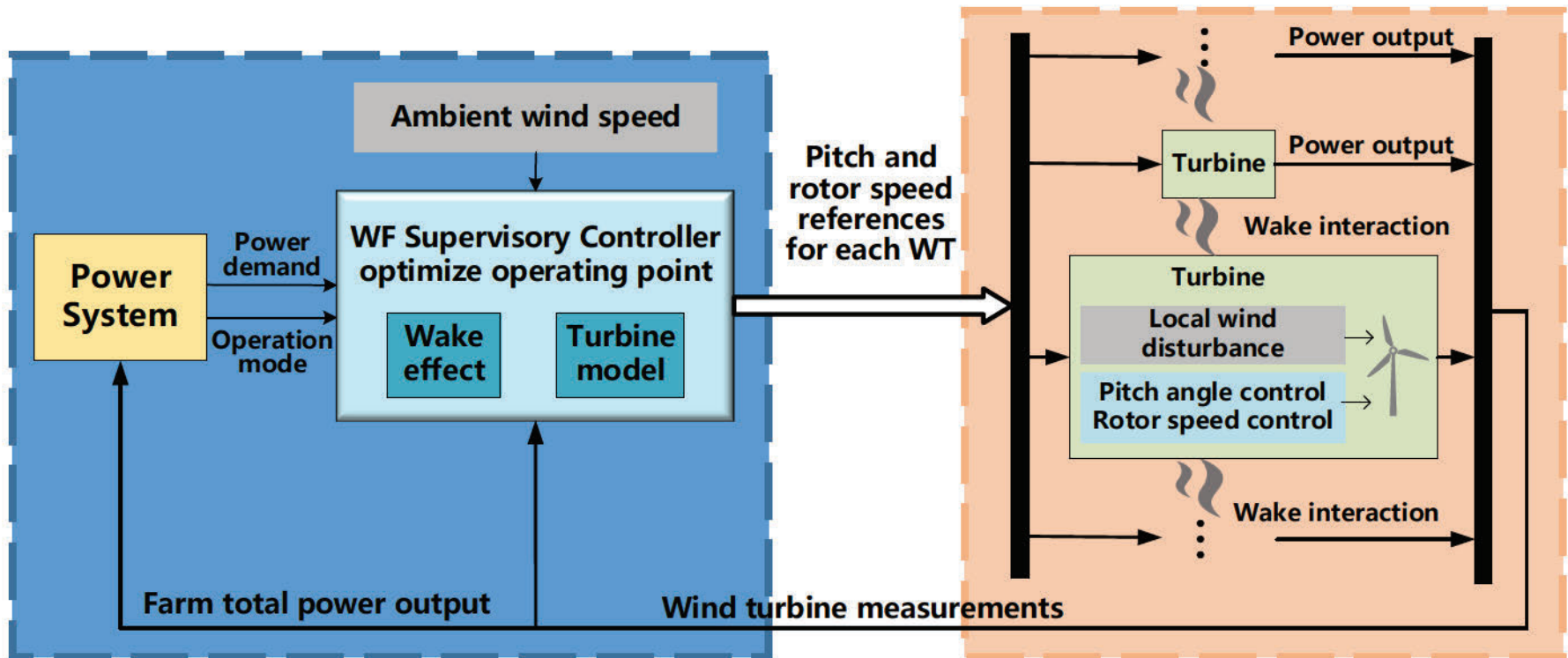


Hierarchical Wind Farm Control Scheme

124

Turbine level control — operating performance; local wind turbulence;

Farm level control — aerodynamic interaction among WTs; respond to meet the requirements (power, spinning reserve) of the power grid



Electromechanical dynamics
Compatible with PCs

Large-scale aerodynamics
Only with supercomputers

Simulation & Verification

- Problem—How does the WF **aerodynamics** interact with the **electro-mechanical dynamic processes** of the power system? Does the widely used **wind farm model** (single aggregated turbine) in electrical power industry valid?
- Solution—

A joint simulation platform

wind flow field

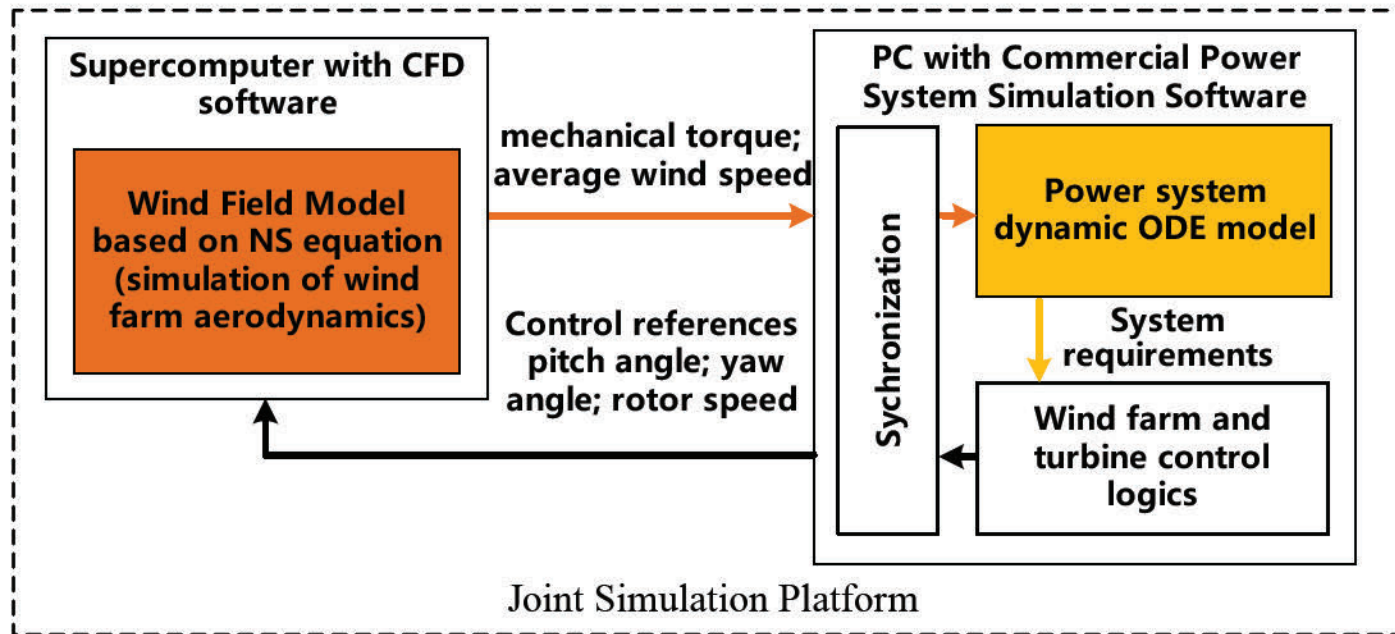
power system



Control & Optimization

- Problem—How to utilize the interaction between the electro-mechanical system and the wind flow field to improve WF operation performance?
- Solution—Are the PDE governing equations practical to be incorporated into the control loop? Or should data-driven approaches be used to fit the aerodynamic interactions based on field data?

Joint Simulation Platform on Sunway Taihu Light ¹²⁶

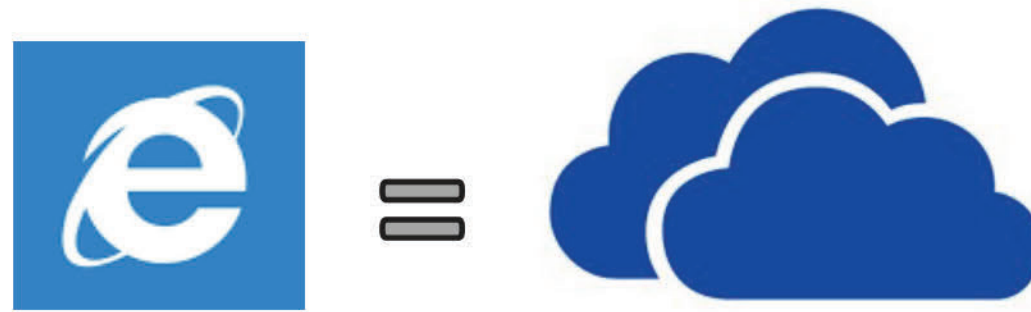


- **Sunway Taihu Light, China's National Supercomputing Centre in Wuxi**

Thank you!

**Farming the Environment into the
Electrical Power Grid**
- Offshore Wind Farm Simulation & Control

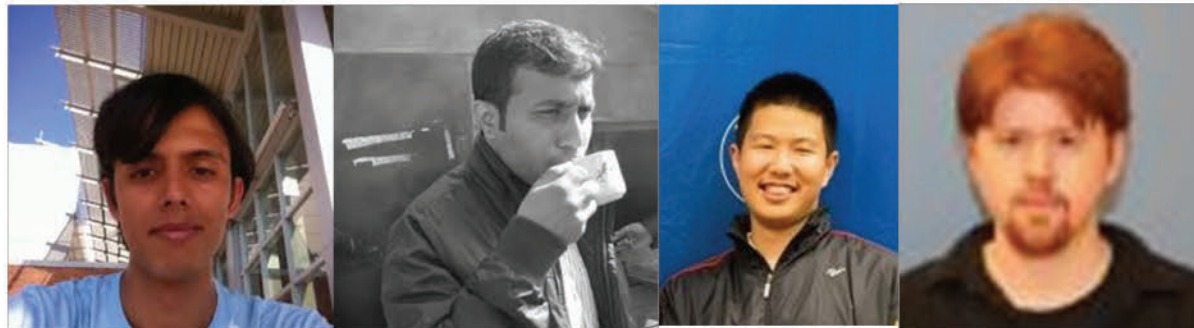
Dr. Yiwei Qiu
ywqiu@mail.tsinghua.edu.cn
BDEC2-Kobe
Feb. 20th, 2019



System Infrastructure for Elevating Edge to be a Peer of the Cloud

Kishore Ramachandran

Embedded Pervasive Lab, Georgia Tech



Enrique Saurez

Harshit Gupta

Zhuangdi Xu

Adam Hall

Fog/Edge Computing

Extending the cloud utility computing to the edge

Fog/Edge computing today?

- Edge is slave of the Cloud
 - Platforms: IoT Azure Edge, CISCO Iox, Intel FRD, ...
- Mobile apps beholden to the Cloud

Vision for the future?

- Elevate Edge to be a peer of the Cloud
 - Prior art: Cloudlets (CMU+Microsoft), MAUI (Microsoft)

In the limit

- Make the Edge autonomous even if disconnected from the Cloud

Problem

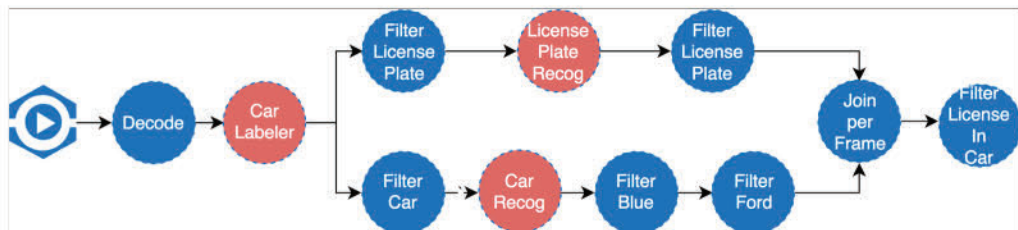
- Densely geo-distributed devices are hard to program:
 - How to distribute the computation?
 - How to move the information?
 - How to tune the parameters?
 - How to support multiple hardware platforms?
- How do we coordinate hundreds of application simultaneously?

Solution

- Declarative languages designed for geo-distributed setting
- Automatic decomposition of stages
- Optimizing for WAN configurations
- Efficient reuse of data and computation

```

SELECT FRAME, OBJECT
WHERE OBJECT.LABEL == "car"
  AND OBJECT.COLOR == RED
  AND OBJECT["LICENSE"] == "LIC"
FROM CAMERAS NEAR GATECH
    
```



Problem

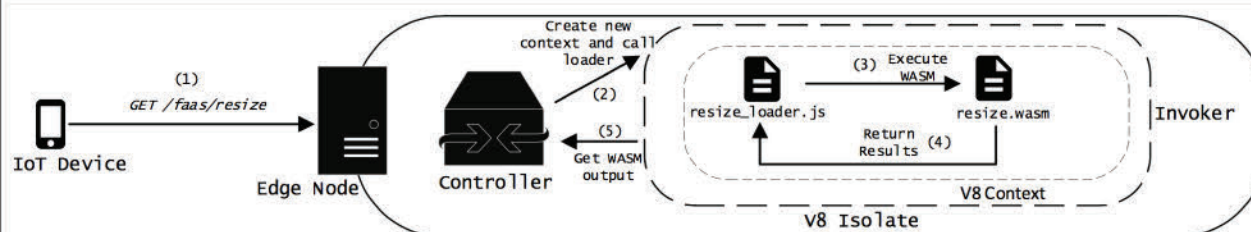
- Resource-constrained Edge nodes need:
 - High multi-tenancy
 - Isolated execution environments
- Dedicated VMs/containers impractical

Solution

- serverless Computing (aka *Function-as-a-Service*)
 - ◆ Applications executing only when needed
 - ◆ Allows for better resource sharing

Adapting Serverless for Edge

- SoA relies on containers (e.g., Docker) to isolate functions
 - Suffers from *cold start problem*
 - ◆ 300ms+ of setup time
 - ◆ Problem for latency-sensitive apps
- Two approaches:
 - ◆ Reduce cold start time by modifying Docker's *runc*
 - ◆ Alternative format (*WebAssembly*) for functions



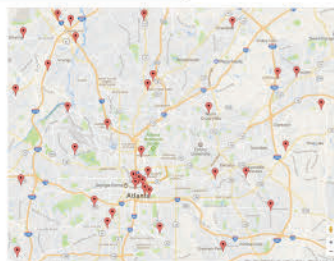
Objective : To provide cloud-like data abstractions (e.g. key-value stores, publish-subscribe) on edge + cloud infrastructure continuum

Aims of the project

- Low-latency data access to real-time apps running in edge
- Support intensive OLAP queries in cloud
- Abstractions similar to cloud platforms
 - get/put key-value pairs, publish-subscribe data
 - don't worry about data placement/migration/....

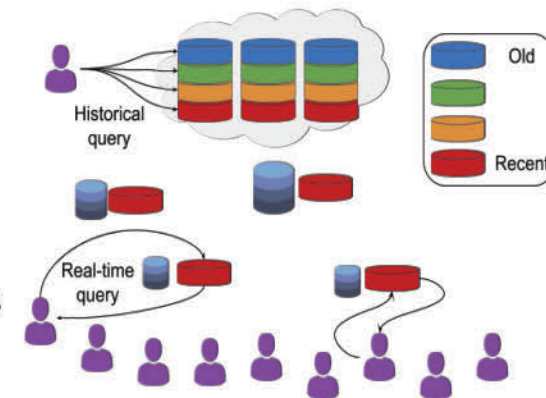
Peculiar challenges in edge computing environments

1. Widely geo-distributed resources
2. Constraints on resource capacity
3. Low statistical multiplexing in client workload
4. Geographically correlated failures are more likely



Interplay between edge and cloud : to alleviate capacity constraints

- Utilize temporal locality of queries
 - Store temporally relevant data on edge nodes
 - Useful for real-time queries
- Older data available on cloud nodes
 - For historical (OLAP) queries

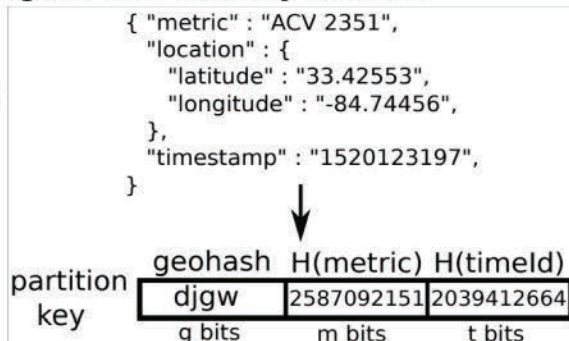


Multiple levels of load balancing for tolerating skews

- Load-balanced data partitioning for long term skews
 - Some areas generate more data than others ⇒ spatial skew in client workload
 - Adding additional capacity should evenly balance the data
- Resource sharing b/w nodes for short-term load surges
 - Temporary surge in traffic at subset of nodes
 - Offload storage temporarily to nearby nodes to maintain throughput

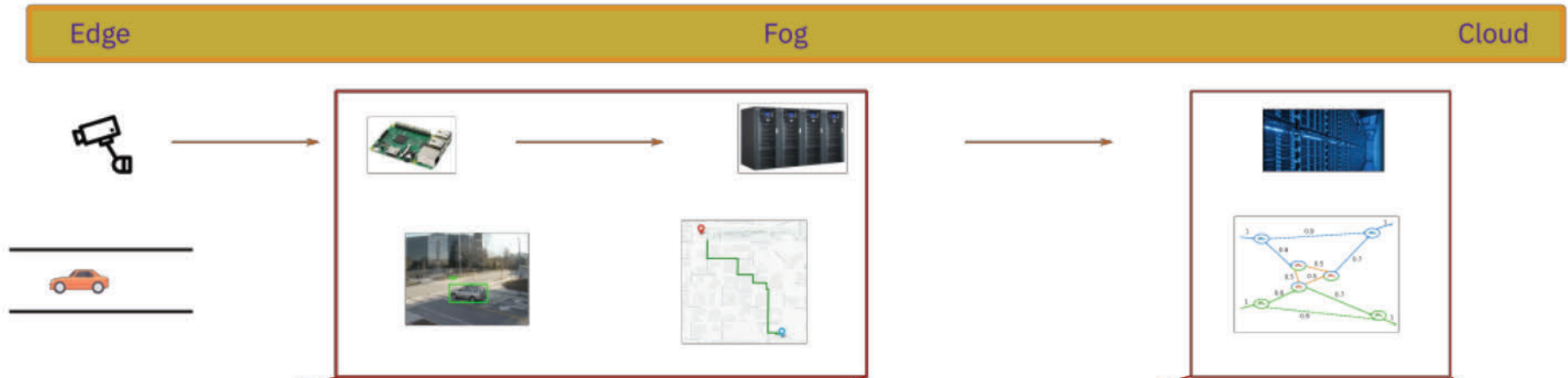
Location-aware data partitioning for low-latency access

- Keep data close to potential users
 - For low latency access
- Use of spatial attributes to distribute data-items



Aims of the project

- Track **all** vehicles over time.
- Store the space-time **trajectories** instead of videos.
- Answer queries **directly** from those trajectories.

**Real-time vehicle detection and re-identification**

- “Each camera” processes (detecting vehicles) its stream at the edge of the network
 - No wasteful network bandwidth to send “everything” to the cloud
- Efficient communication policy between “cameras”
 - Forward and backward propagation to exchange detection objects for re-identification
- Efficient storage policy across “cameras”
 - Construct the trajectories from the result of re-identification
 - Use aggregation to limit the storage requirements for each “camera”

High-performance query and graph processing

- Adopt probabilistic graph to store the trajectories of all vehicles
 - Encode extra information (confidence from re-identification) for better query answering.
- Explore graph processing techniques to relieve the errors from detection and re-identification
 - Example: maximum-probability path for one vehicle versus maximum-probability paths for all vehicles (Each vehicle should occupy an unique disjoint path in the graph)

BDEC2 Kobe: Ricart

24 February 2019



Glenn's 2019 In-and-Out List (Internet Systems)

<u>Out</u>	<u>In</u>
People use the Internet	Devices use the Internet
Move data to computing	Move computing to the data
Exploit massive datacenters & networks	Exploit data locality
Validated datasets	Perishable data streams
Abundant inter-city backbone bandwidth	Abundant intra-city access bandwidth
Bandwidth is the key net measurement	Latency is the key net measurement
Best effort	Predictable, deterministic response time
Task Scheduling	Packet Scheduling
Computers model and monitor real world	Computers are integral parts of real world

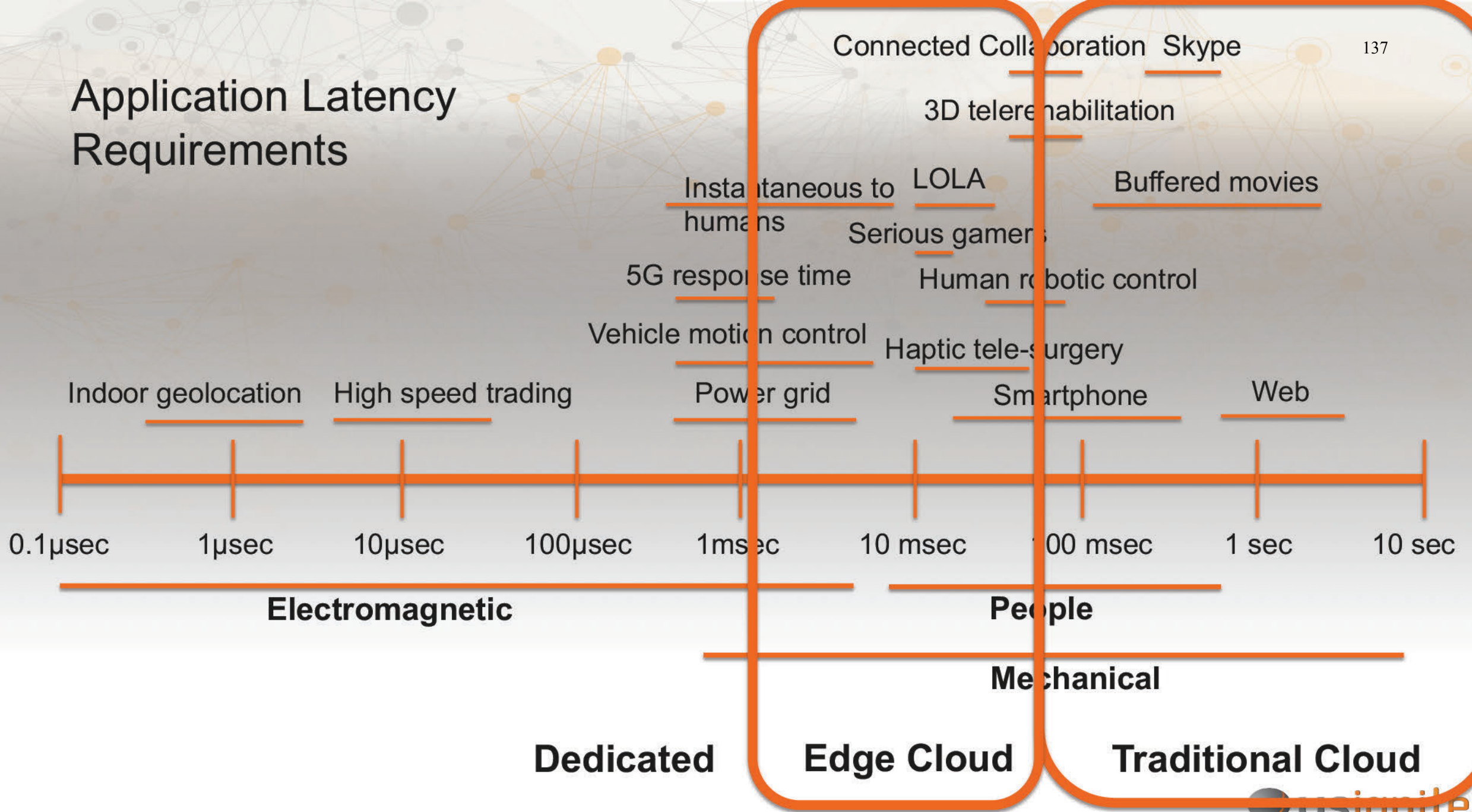
Glenn's 2019 In-and-Out List (Technology)

<u>Out</u>	<u>In</u>
CPUs	GPUs and TPUs
Deductive Programming	Inference Engines
Virtual Machines	Containers, Microservices, and Lambdas
Datasets	Data flows
Location Driven by Economy of Scale	Location Driven by Required Latency
Discrete runs	Continuous operation
Edge Computing	Multiple Collaborating Edges
Vertically-integrated clouds	Horizontally-cooperating clouds
North-South Network Traffic	East-West Network Traffic

Glenn's 2019 In-and-Out List (Programming)

<u>Out</u>	<u>In</u>
User Interface	Device and environmental interfaces
Interact with user	Interact with whole environment
Respond	Anticipate
Explicit Programming	Stating intentions
Explicit intentions	Machine learning (discovery)
Focus on apps	Focus on data streams

Application Latency Requirements





**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



Earth System Modelling: requirements and challenges

Kim Serradell

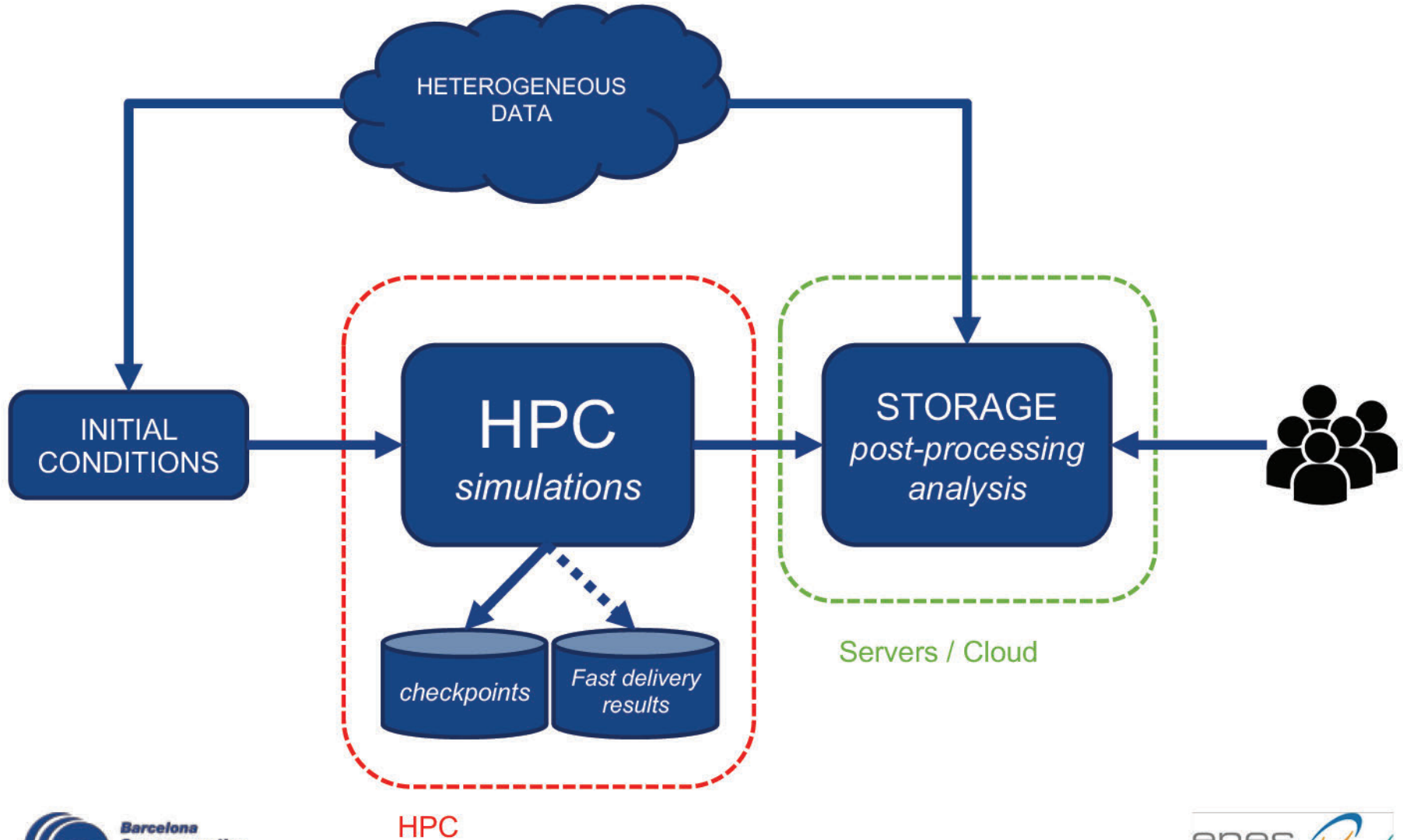
Computational Earth Sciences

19th-21st February

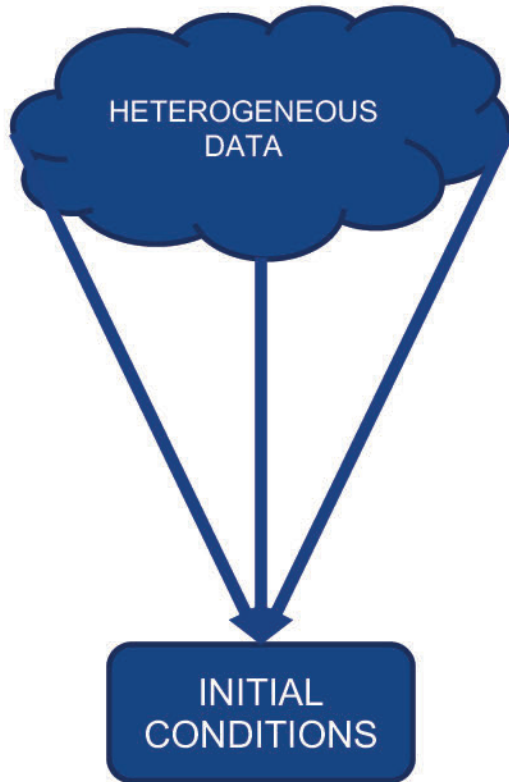
BDEC2 Kobe 2019



Earth System Modelling Workflow



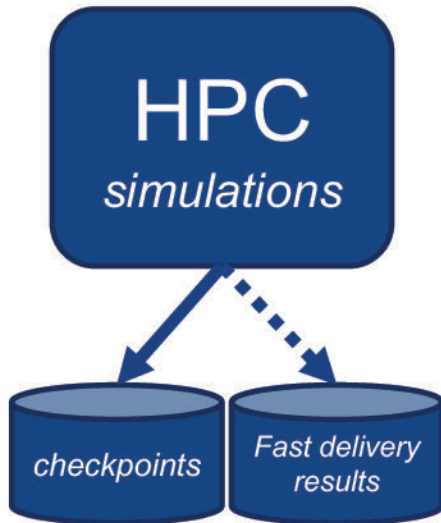
Pre-processing



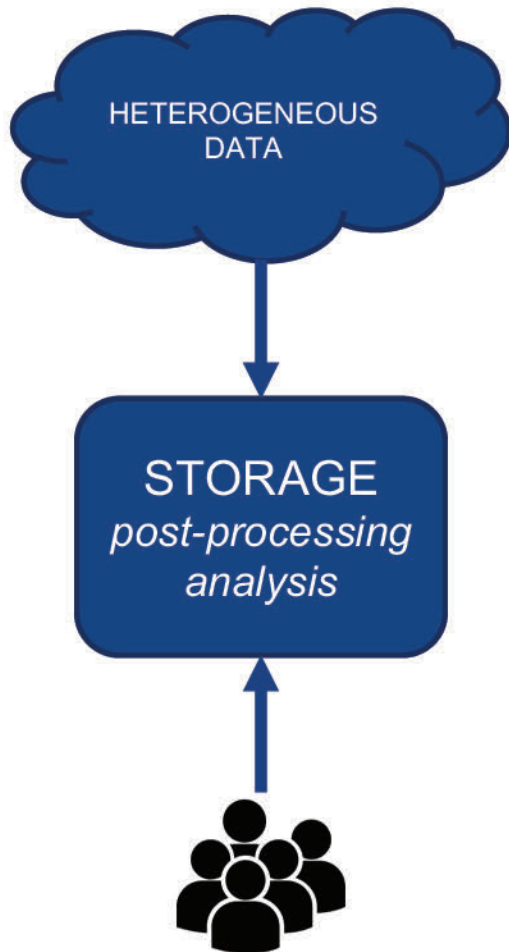
- Deal with massive and heterogenous amount of data (Earth Observations)
 - Sensors (with valid and non valid data)
 - Satellite data
- Complex processes to build initial conditions
 - Real-time data
 - Data assimilation
 - Model checkpoints as initial conditions

Model simulation

- Traditional HPC
 - High-powered nodes, large batch jobs, low-latency networks to deal with increasing problem size (resolution, ensembles, ...)
 - Programming models to deal with heterogenous architectures (DSLs and "separation of concerns")
 - Reproducibility
 - Software Stack (using tools like Spack or Containers)
 - Results (CMIP6 exercise → ~ 100 models)
 - Operational services
 - Storage
 - Periodic output of selected variables
 - Fast delivery (for meteorological applications)
 - Traditional (mix disk/tapes) for later analysis



Post-processing



- Multiple data sources to validate results
- In-situ analysis - visualization
 - Reformatting, sub-setting, re-gridding, averaging...
 - Limit as possible data transfers
 - On user demand analysis and needs
- Reliable dissemination platforms
 - Earth System Grid Federation (CMIP6: 15 to 30 Pb of data)
 - Curated archive, identification and citation
- Efficient and timely handling
- Required throughput for real world applications
- Climate services
- Machine Learning growing need

And last but not least...

- We need powerful, reproducible and easy to adopt workflows to orchestrate the full earth modelling
- Don't forget "Human factor"
 - One individual, multiples roles
 - Training Research Engineers, Computer Scientists...
 - Identify end-users: Earth Scientists, Data Scientists, Policy Markers...
- Extreme Earth Flagship
 - Technology case (Science Cloud, Big Data handling and Distributed extreme-scale computing)
 - <http://www.extremearth.eu/technology-case>





**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



**EXCELENCIA
SEVERO
OCHOA**

Thank you

kim.serradell@bsc.es



COMMENTS FOR BDEC MEETING

Dan Stanzione

Executive Director, TACC

Associate Vice President for Research, The University of Texas at Austin



Follow the
STAMPEDE

Powering Discoveries That Change The World

THESE WORKFLOWS ARE ALREADY HERE

- ▶ We already see this kind of work at TACC
- ▶ An exemplar project, DARPA's Synergistic Discovery and Design
 - ▶ Multiple robot-driven experimental facilities feeding data into an HPC-enabled repository, with hundreds of analysts consuming via web services.



Synergistic Discovery and Design

Objective: *Develop data-driven methods to accelerate design in domains that lack complete models.*

Design in domains with
robust scientific theories

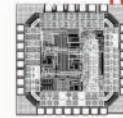
Historically

Mathematical models
&
Simulation

+

Computer aided
design tools

=



Integrated Circuits



Aeronautics

Design in domains with
incomplete scientific theory

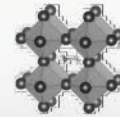
SD2

Data &
Discovery Algorithms

+

Data-driven
design tools

=



Perovskite
Chemistry



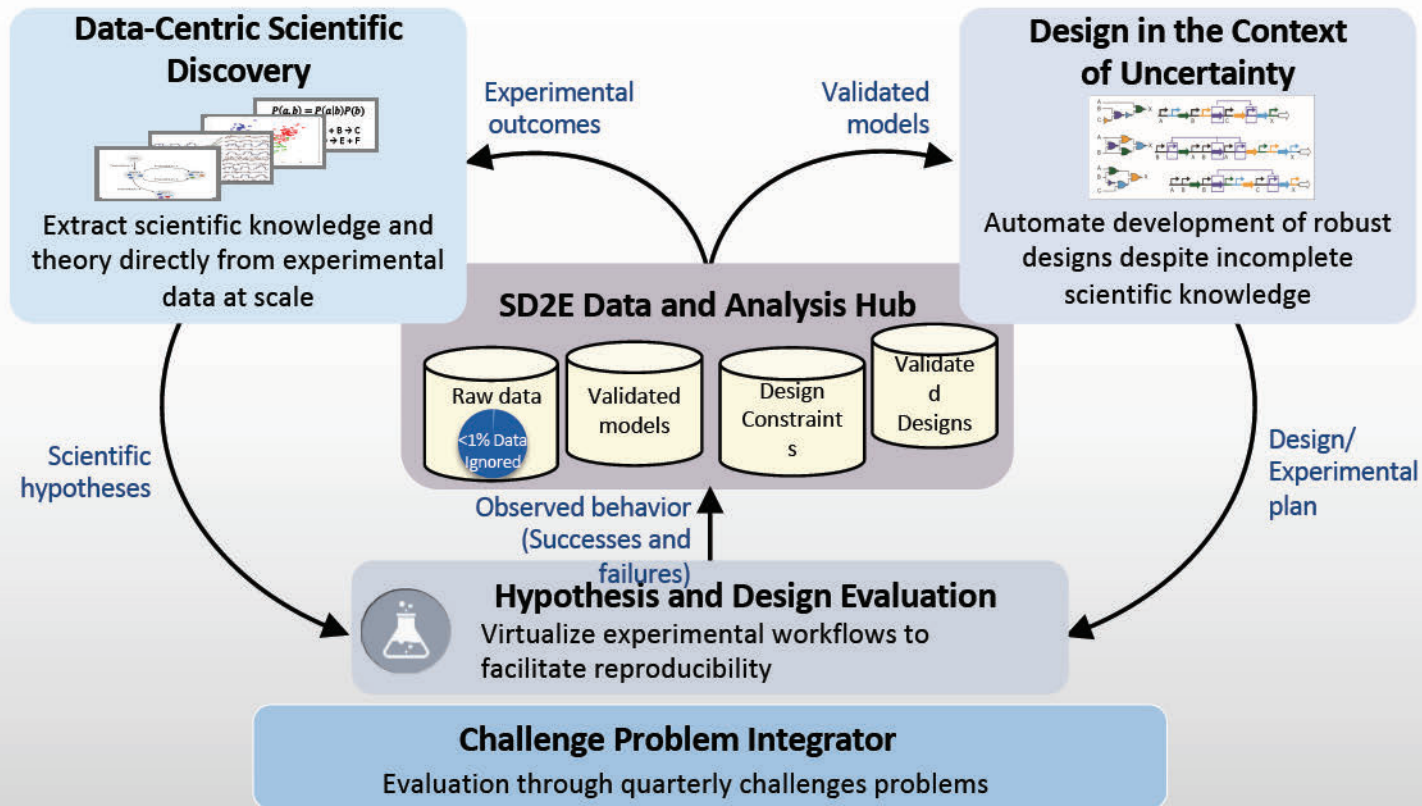
Protein Design

DoD Applications: Foundational technologies with use cases relevant to national security, such as:

- Sensors to detect Chemical, biological, radiological, and nuclear (CBRN) agents
- Organisms that detect and metabolize nuclear waste to remove radiological threats
- Inexpensive, efficient solar materials

Algorithms for scientific discovery and design

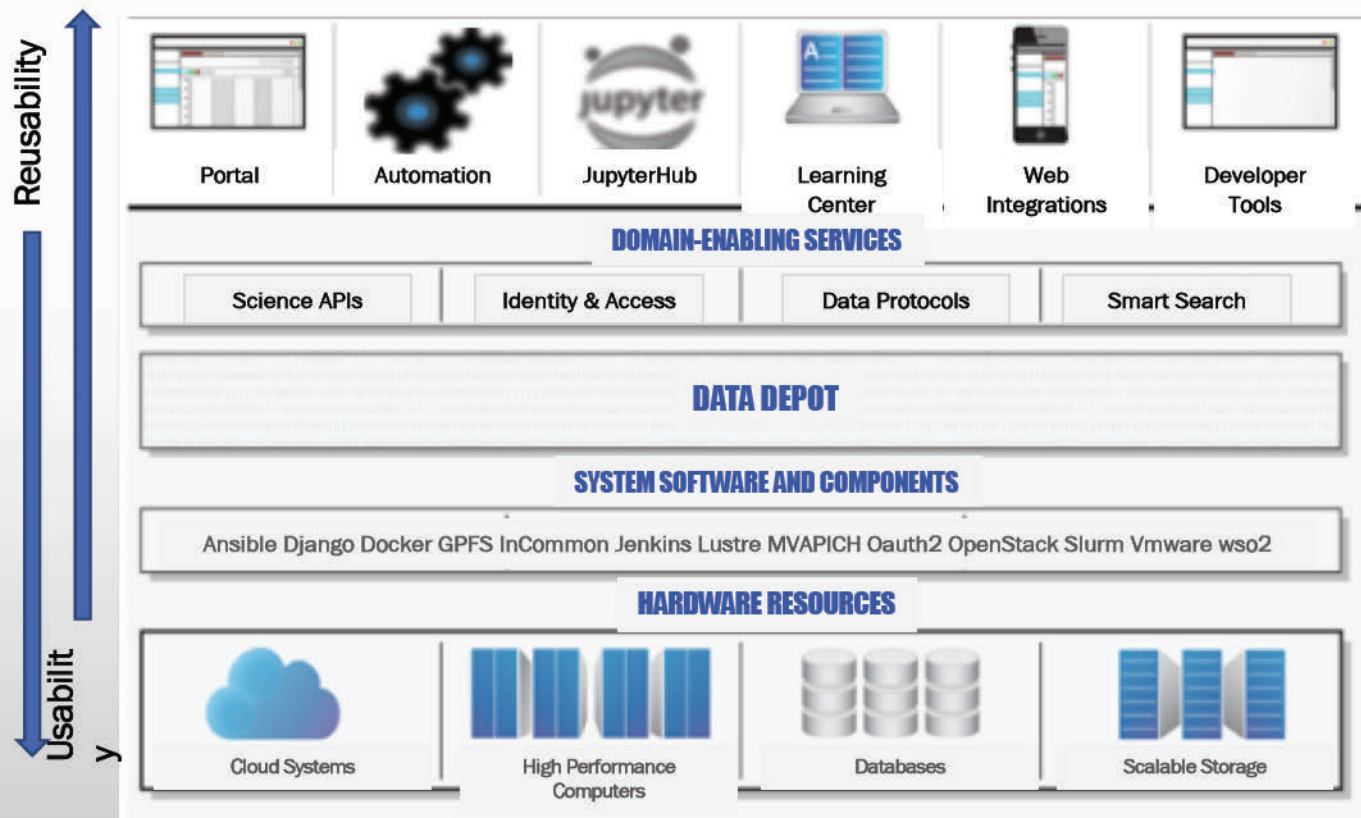
- AI methods that considers trillions of theories and converges on ones that fit the data
- AI methods that meets or exceeds human performance in discovery and design
- Human-computer team that discovers a new subfield of science through analysis of failed experiments at scale



SD2E Integrated Infrastructure Components

- **User Applications**
 - Discovery Workspace
 - JupyterHub
 - 100+ shared notebooks
 - SynBioHub
 - 25,000+ synthetic biology components
 - SD2E CLI
 - Comprehensive scripting support
- **Developer Services (Hosted)**
 - GitLab
 - 100 active repositories
 - Jenkins
 - 70 CI/CD pipelines
 - Portainer
 - 31 user-managed Docker Swarm services
- **Web Service APIs**
 - Agave Science API
 - 45+ user apps
 - Abaco Functions-as-a-Service
 - 125 deployed functions
- ElasticSearch/Logstash/Kibana
 - 400,000 log events/day
- Data Catalog API
 - JSON schema-informed metadata management
- TACC S3
 - Performant S3-compatible uploads
- **Computing Resources**
 - Stampede2
 - Wrangler
 - Maverick 1 & 2
 - Lonestar5
 - Rodeo & Jetstream
- **Data Resources**
 - Stockyard
 - Corral
 - AWS S3+Glacier

Integrated Cyberinfrastructure Design



Web-based Apps



Context-aware Web Services



Pervasive and Performant Storage



Software Foundations

Exascale Physical Systems + Cloud



ARCHITECTURAL APPROACH

- ▶ This requires some new top level architectural approaches, but many things stay the same:
 - ▶ HPC apps should still be programmed like HPC apps
 - ▶ AI/DL/ML apps should still be programmed like HPC apps ;)
 - ▶ Web services provide the top-level glue between subsystems, security architecture, etc.
 - ▶ Data transfer may not be performant in these environments.
- ▶ We have building blocks enabling this now:
 - ▶ All our systems have a front end, persistent REST API
 - ▶ All systems now support container technology, including grabbing from repos.
- ▶ Policies still need to evolve – a single security standard would be nice.

NEW SYSTEM SUPPORT ACTIVITIES

- ▶ Full Containerization support (this platform, Stampede, and *every other* platform now and future.
- ▶ Support for Controlled Unclassified Information (i.e. Protected Data)
- ▶ Application servers for persistent VMs to support services for automation.
 - ▶ Data Transfer (ie. Globus)
 - ▶ Our native REST APIs
 - ▶ Other service APIs as needed – OSG (for Atlas, CMS, LIGO)
 - ▶ Possibly other services (Pegasus, perhaps things like metagenomics workflows)

THANKS!

DAN STANZIONE
DAN@TACC.UTEXAS.EDU



国家超级计算无锡中心
National Supercomputing Center in Wuxi

A Brief Report on the Utilization Status of Sunway TaihuLight

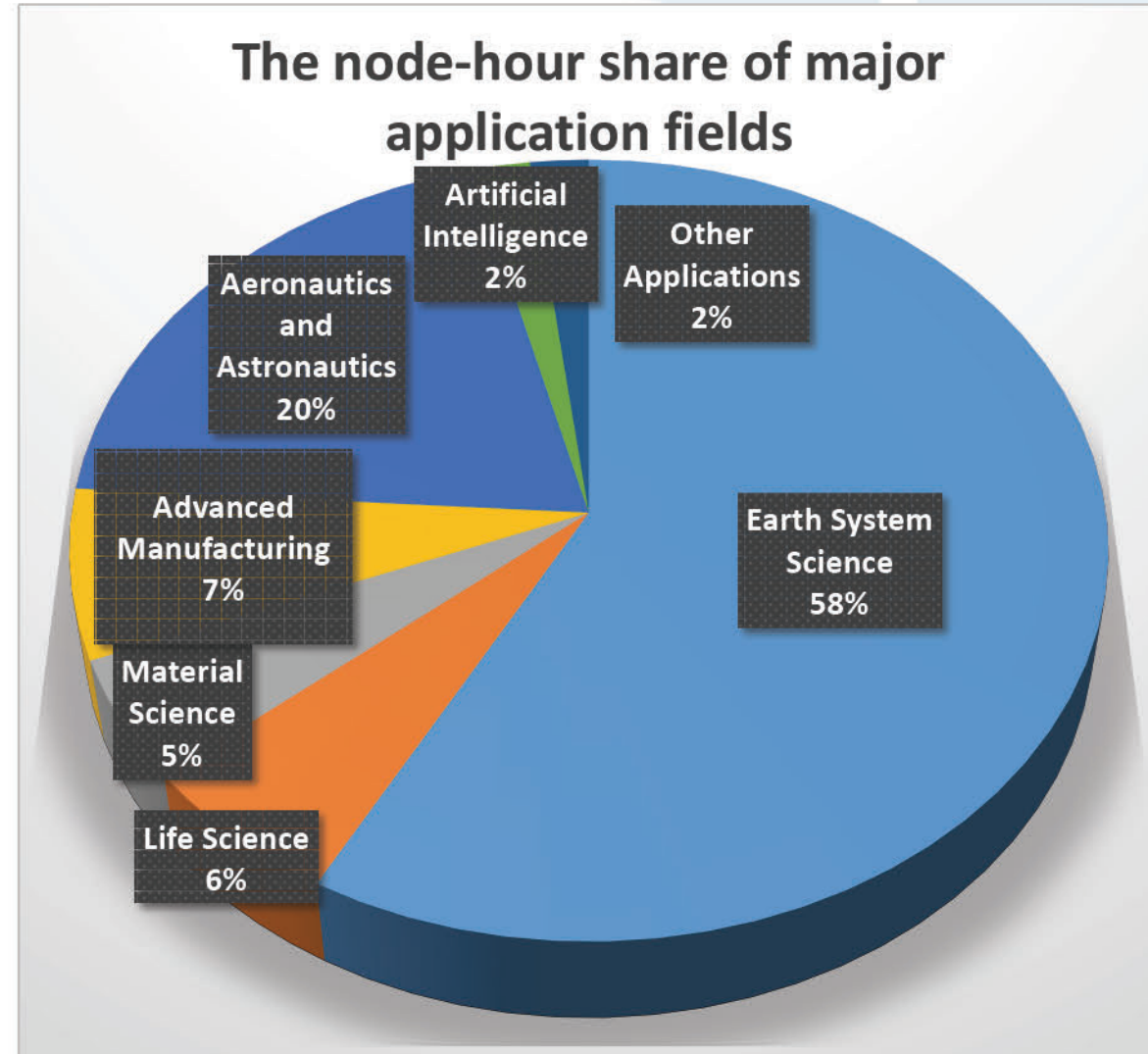
Overview

■ Sunway TaihuLight

- Over 40,000 computing nodes
- Consumed over 210,000,000 effective node-hour in 2018 (about 60% utilization rate)
- Occupied 8.4 PB out of the total 20 PB storage

■ Major application fields

- Earth System Science: CESM, WRF, CWRF etc.
- Life Science: GROMACS, DOCK, Gene Screening, etc.
- Material Science: VASP, LAMMPS, etc.
- Advanced Manufacturing: OpenFOAM, Palabos, swLBM, etc.
- Aeronautics and Astronautics
- Artificial Intelligence



Earth System Science

156

■ CESM

- ❑ Community Earth System Model (CESM)
- ❑ Consumed around 75,000,000 node-hour
- ❑ Generated 400 TB data, most in the form of NetCDF
- ❑ The major users are the Qingdao National Laboratory for Marine Science and Technology, and Tsinghua University

■ WRF & CWRF

- ❑ Weather Research and Forecasting Model (WRF)
- ❑ Consumed around 50,000,000 node-hour
- ❑ Generated 2.5 PB data, most in the form of NetCDF
- ❑ The major users are the China Meteorological Administration and local meteorological companies



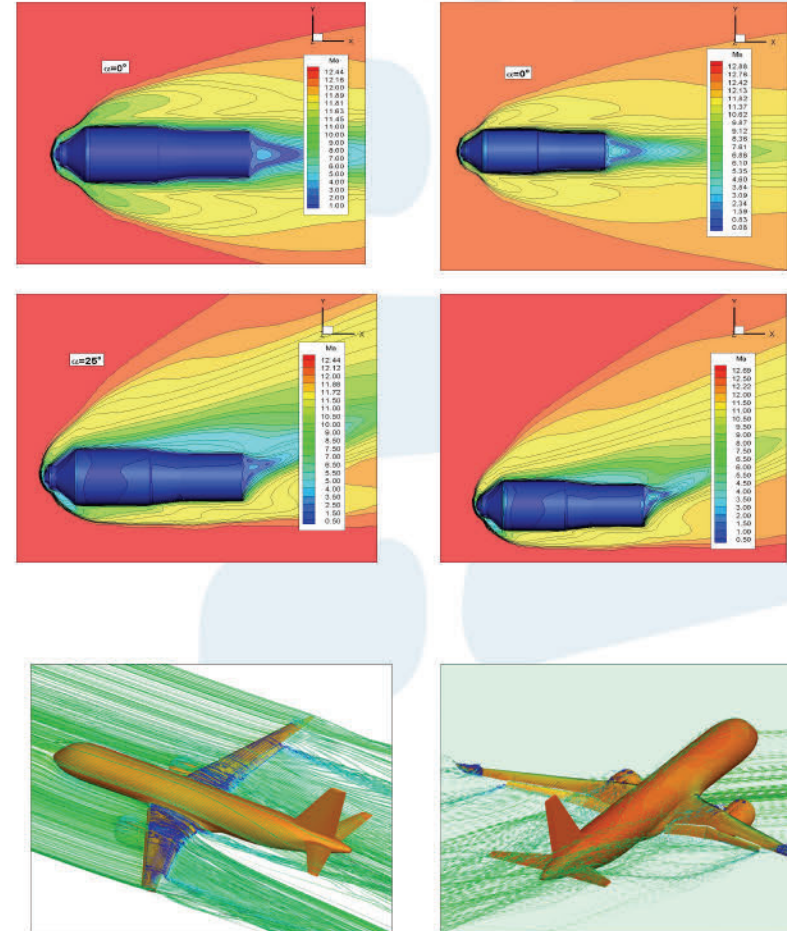
Aeronautics and Astronautics

■ Unified Astronautics Numerical Simulation Software

- Consumed around 40,000,000 node-hour
- Generated 200 TB data
- The major user is the National Laboratory for Computational Fluid Dynamics of Beihang University.

■ Airliner Design

- C919 Airliner design
- Consumed around 2,000,000 node-hour
- Generated 50 TB data
- The major user is the China Aerospace Establishment



Conclusion

- The Sunway TaihuLight is mainly used in the following fields:
 - Earth System/ Aeronautics and Astronautics/ Material Science/ Life Science/ Advanced Manufacturing/ Machine Learning/ ..
 - Most of the applications are based on open-source codes
- The requirements from scientific and engineering applications for the ACP
 - Tremendous computational capacity, storage capacity and network bandwidth for ultra-high resolution and large-scale simulation
 - Data compression, data backup, and redundancy deletion
 - Ability of conducting real-time and continuous post processing and visualization
 - Support for interdisciplinary with certain flexibility



THANK YOU





Convergence of computing and storage models in Big Data and Extreme computing

Martin Swany, Intelligent Systems Engineering

INDIANA UNIVERSITY

Convergence of Domains and Functionality

- Computing and Networking
 - Virtual bridge encapsulation is essential in many data center networks
- Cloud, HPC and Embedded Systems
 - Supercomputers and cloud engines built out of ARM cores
- Network Functions Virtualization (NFV) is the networking equivalent of what we are doing here
 - Common hardware elements can realize routers, firewalls, load-balancers with software changes



Commonalities in execution across the spectrum

- Serverless computing / Function as a service
- Over-decomposition in asynchronous many task programming
- Stream processing operators
- ***Looked at correctly, these are all the same design patterns – small computational kernels of data (message) driven operations***



Storage is the same

- Simple storage elements with simple semantics compose to provide a range of functionality
- Storage and networking for content distribution and streaming data
- Storage plus processing, for serverless, stream and everything else!



Converged functionality for an advanced cyberinfrastructure platform

- We have a chance (and a charter) to design something that is perhaps the first radical departure from existing abstractions in many years
- We should focus on applications that can take advantage of it
- Bare metal, choose your own stack is closer to indifferent cohabitation than convergence

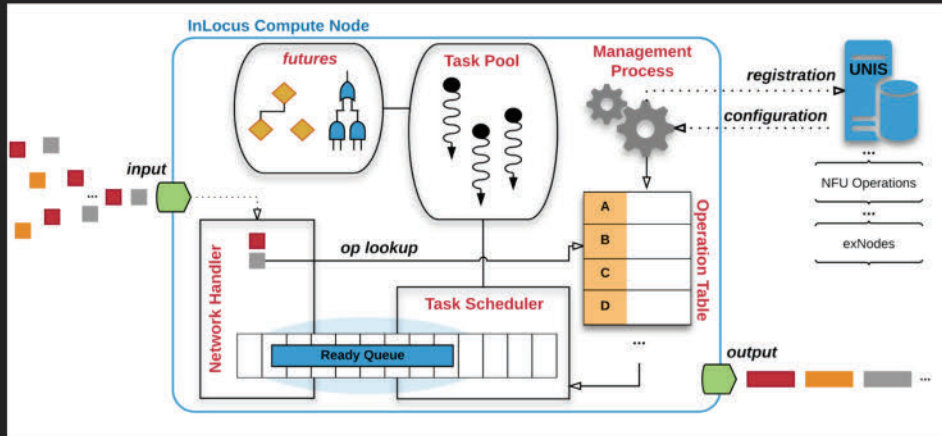
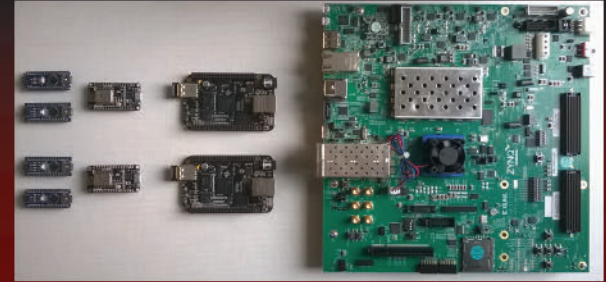


InLocus

InLocus is an architecture that allows streaming data to be processed in the network. InLocus targets microcontrollers, microprocessors, network processors and Field Programmable Gate Arrays (FPGAs) that can be embedded in the network for highly efficient data processing at the edge.

Edge Devices

- Microcontroller
- SoC Computer
- FPGA



Motivation • Smart Cities and the Internet of Things generate massive data. Moving processing to the edge improves latency and reduces network usage.

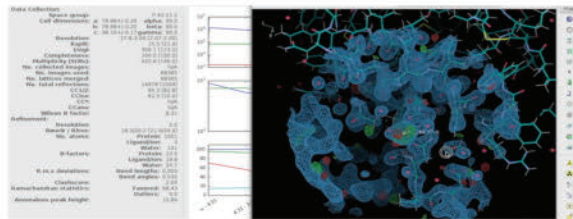
Implementation • Network Functional Units form a distributed, heterogeneous compute fabric of MCUs, FPGAs, and resource-constrained devices, programmed and dynamically routed by a central server (Unified Network Information Service or UNIS).

Future Work • Benchmarking against VM-based applications like Heron to compare performance of C and FPGA implementation with traditional cloud architectures.



Data Analytics at the Exascale for Free Electron Lasers (ExaFEL)

- SLAC-to-NERSC interfacility workflow <https://vimeo.com/slaclab/review/242658577/8aaa06acfb>
- Acceleration of nanocrystallography and single particle imaging analyses



Requirements:

- High throughput data streaming and fast data analysis (real-time)
- Real-time data reduction
- Interfacility data flow from user facility to analysis at supercomputer or cloud

PI: A. Perazzo, LANL PI: C. Sweeney

ExaLearn: Codesign Center for Machine Learning Technologies

- Closed loop control of experiments via real-time reinforcement learning
- Surrogate model development
- Learning on multi-modal data

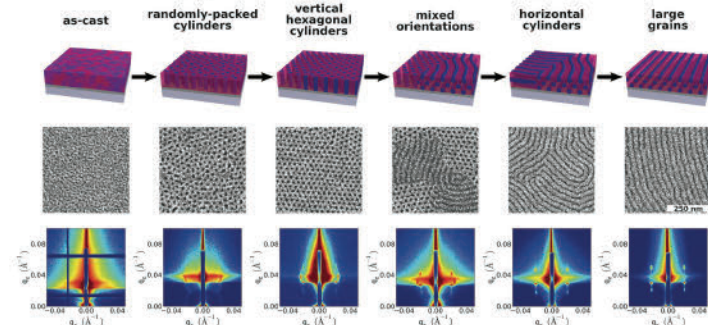


Image courtesy Pawel W. Majewski and Kevin G. Yager

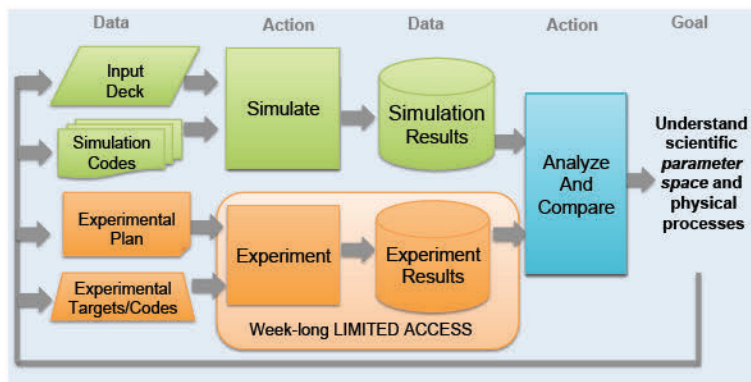
Requirements:

- Model trained at supercomputing facility and deployed at user facility
- ML control of devices at the edge
- Ensemble runs to create synthetic training data.

PI: F. Alexander, Control Use case Lead: C. Sweeney

Real-time Adaptive Acceleration of Dynamic Experimental Science

- Experiment/simulation/emulation for dynamic compression light source experiments.



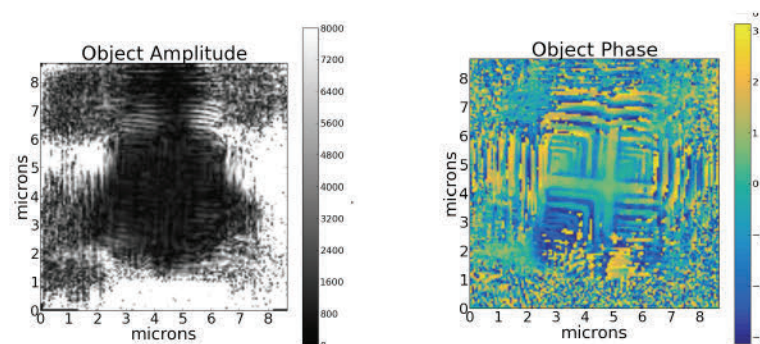
Requirements:

- Real-time emulation and data analysis for human decision-making
- HPC for ensemble simulations
- Integration of data from materials repositories

PI: J. Ahrens, Co-Pis: C. Bolme, R. Sandberg

Coherent Diffraction Imaging

- Ptychography, laminography and tomography of samples



Requirements:

- High throughput data collection
- Compute-intensive (GPU) local and/or remote analysis
- Fast data reduction and data tidying
- Streaming data and data formats

PI: R. Sandberg

- Flow of raw and/or processed data between many locations: edge hardware, user facility, supercomputing center, cloud and users
- Analysis on streaming data in real-time (fast feedback and reduction)
- High-bandwidth data transfer between edge, local and nonlocal compute and storage
- Human-in-the-loop and collaborative (with potentially remote users) decision-making
- On-demand and reserved provisioning of supercomputing resources
- Simultaneous processing of data at multiple levels (edge (detector), streaming analysis, batch analysis, human-in-the-loop decision-making, closed-loop inference for control)
- Generation and synthesis of multi-modal data (streams) from multiple sources (edge (detector), analytics results from supercomputer, simulation, machine learned models, scientific domain databases, cloud)

Connectivity:

- Novel streaming protocols and abstractions to enable streaming between heterogeneous resources and with constraints.
- Heterogeneous data lakes to support data stored in distributed locations, on various storage platforms and with various security constraints.
- Schedulers that operate on heterogeneous workloads (containers, tasks, etc.)
- Composition of services on cloud, mobile and supercomputers

Facility support:

- Resource provisioning for semi-scheduled, on-demand, bursty, and pseudo real-time workloads
- Effective but non-intrusive authentication between facilities
- Front-end services at supercomputing facilities

Usability:

- Recovery mechanisms for computations that may fail anywhere
- Execution and data version provenance across converged platform
- Mechanism for performance- and arena-portable (cloud, mobile, supercomputer) code

Programming models:

- Parallel programming models that will enable flexible, portable, computations all the way out to the edge.
- Programming models that can extend across arenas (cloud, mobile, HPC) for connecting services or for distribution of compute.
- Programming models that support data models that are abstracted from the storage type.

Workflows:

- Abstractions and interfaces that enable users to specify high-level constraints (result quality, performance, resource usage) on workflow.
- Mechanisms that allow workflows to interface with each other

Domain-specific languages and interfaces:

- Abstracting novel low-level protocols and programming models to aid in the convergence of services in the cloud and HPC arenas



Building the Open Storage Network

Alex Szalay
Christine Kirkpatrick, Kenton McHenry, Alainna White
Steve Tuecke, Ian Foster, Joe Mambretti

Institute for Data Intensive Engineering and Science

idies

Computational and Networking Infrastructure

- The NSF has invested significant funds into high performance computing, both capacity and capability
 - *These systems form XSEDE, a national scale organization with excellent support infrastructure*
- The NSF has invested about \$150M to bring high-speed connectivity to over 200 universities in the CC-NIE and CC* programs
 - *Internet2 provides a stable high-speed backbone at multiple 100G lines*
- Storage infrastructure largely balkanized
 - *Every campus/project does its own specific vertical system, lots of incompatibilities and inefficiencies*
 - *Big projects need petabytes, also lots of 'long tail' data*
- Cloud storage not a good match at this point for PBs
 - *Wrong tradeoffs: cloud redundancies too strong for science*

Opportunity

- The NSF has funded 150+ universities to connect to Internet2 at high speeds (40-100G) for ~\$150M
- Ideal for a large national distributed storage system:
 - *Place a 1-2PB storage rack at each of these sites (~200PB)*
 - *Create a redundant interconnected storage substrate*
Incredible aggregate bandwidth, easy flow between the sites
 - *Can also act as gateways/caches to cloud providers*
 - *Automatic compatibility, simple standard API (S3)*
 - *Add Globus to the top layer (G-Connect, GlobusAuth)*
 - *Implement a set of simple policies*
 - *Enable sites to add additional storage at their own cost*
 - *Variety of services built on top by the community*
- Estimated Cost: ~\$20M for 100 nodes

System could be the world's largest academic storage facility

Transformative Impact

- Totally change the landscape for academic Big Data
 - *Create a homogeneous, uniform storage tier for science*
 - *Liberate communities to focus on analytics and preservation*
 - *Amplify the NSF investment in networking*
 - *Very rapidly spread best practices nationwide*
 - *Universities can start thinking about PB-scale projects*
- Impact unimaginable
 - *Links to XSEDE, NDS, RDA, Globus*
 - *Big Data projects can use it for data distribution*
 - LHC, LSST, OOI, genomics
 - *Small projects can build on existing infrastructure*
 - *Enable a whole ecosystem of services to flourish on top*
 - *Would provide “meat” for the Big Data Hub communities*
 - Enable nation-wide smart cities movement

New opportunity for federal, local, industrial, private partnership

Novel Applications of OSN

Community prototypes for different use cases, e.g.

- i. Move and process 1PB of satellite images to Blue Waters using just-in-time streaming*
- ii. Move specific PB-scale MREFC data from Tier1 to multiple Tier2s at universities for detailed sub-domain analytics (LSST)*
- iii. Create large simulation (cosmology or CFD) at XSEDE, using ML-driven data compression and move to a university to include in a Numerical Laboratory*
- iv. Take a large set of LongTail data with small files through a DropBox-like interface, save them on OSN and organize into larger containers, and explore the emergence of broader context*
- v. Interface to cloud providers (ingress/ egress/ compute), especially with GPU allocation*

What is the Future?

- Over the next 5 years it will host and move much of the NSF generated academic data
- Will establish best practices and standards
- Open Data Services migrate one level up, built over **trusted** storage

- Some time in the next 10 years most academic data will migrate into the cloud due to economies of scale
- The OSN will not become obsolete, but becomes part of a hierarchical data caching system
- It will also provide impedance matching to the Tier0/1 to Tier2 center connectivity of MREFC instruments/projects

Summary

- High end computing has three underlying pillars
 - *Many-core computing/HPC / supercomputers*
 - *High Sped Networking*
 - *Reliable and fast data storage*
- The science community has heavily invested in first 2
 - *Supercomputer centers/XSEDE, Internet 2, CC-NIE, CC**
- Time for a coherent, national scale solution for data
 - *Needs to be distributed for wide buy-in and **TRUST***
- Only happens if the whole community gets behind it
- ***KEEP IT SIMPLE and AGILE!***

openstoragenetwork.org

Scientific Methods Transformation via AI/Deep Learning & Advanced Cyberinfrastructure Platforms (ACP): Fusion Energy Exemplar

William M. Tang

Princeton University/Princeton Plasma Physics Laboratory (PPPL)

**Big Data & Extreme Computing 2nd Series, (BDEC-2)
Workshop 2: Kobe, Japan**

February 19-20, 2019

Princeton/PPPL Fusion AI Team:

Julian Kates-Harbeck (Harvard U/PPPL), Alexey Svyatkovskiy (Princeton /Microsoft)

Eliot Feibush (PPPL/Princeton), Kyle Felker (Princeton/ANL), Ge Dong (Princeton/PPPL)

Dan Boyer (PPPL) & Keith Erickson (PPPL)

AI/Deep Learning: Scientific Methods Transformation Exemplar: Fusion Energy

Most Critical Problem for Fusion Energy →

Accurately predict and mitigate/avoid large-scale major disruptions in magnetically-confined thermonuclear plasmas such as ITER – the \$25B international burning plasma “tokamak”

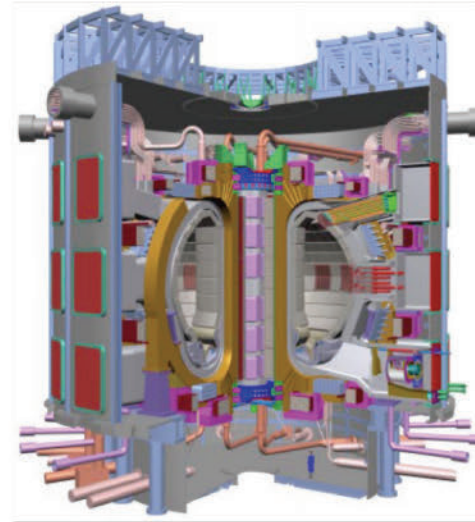
- with goal of 10X “break-even” and operation scheduled for ~ 2026

Technical Focus → development & deployment of advanced Machine Learning Software via AI/Deep Learning Neural Nets – both Convolutional & Recurrent in Princeton’s “FRNN Code”

Recent Status → described in detail in *NATURE* article (accepted for publication, January 2019)

Success of ITER Requires Sufficiently Low Disruption Rate

- Mid-pulse disruptions eliminate planned discharge time following disruptive event → *greatly reduces physics productivity*
- Disruptions can require *long recovery time bad for overall shot frequency*
- Disruption heat fluxes can *reduce component lifetime* (e.g. divertor target ablation)
- Damage to in-vessel components *can require shutdown for repair*

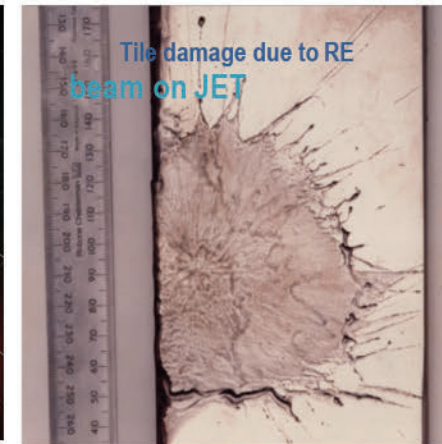


Availability > 80%
(during operation periods)

Design target <10%
disruptivity



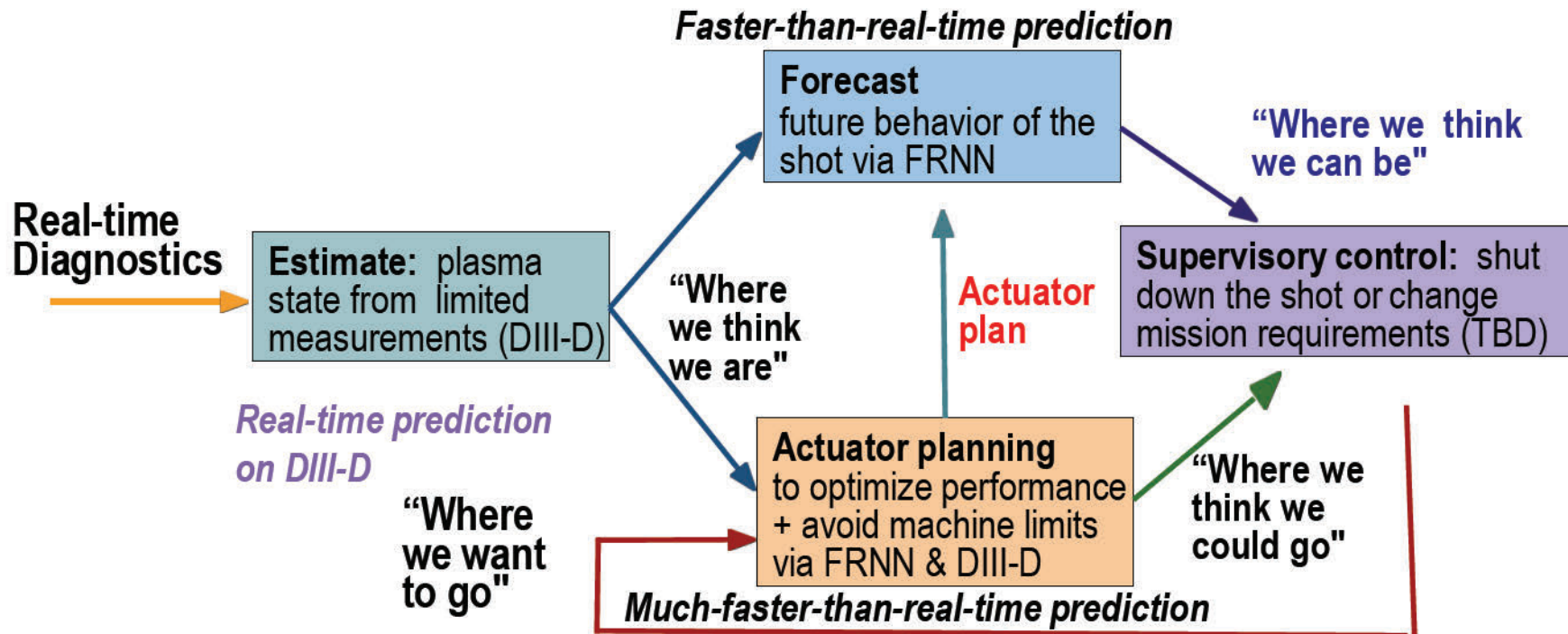
Tile broken by disruption forces in DIII-D



Tile damage due to RE beam on JET

Control Capabilities Needed for Real-Time Experimental Planning

with Dan Boyer, Keith Erickson, ... and especially experimental/advanced diagnostic expertise



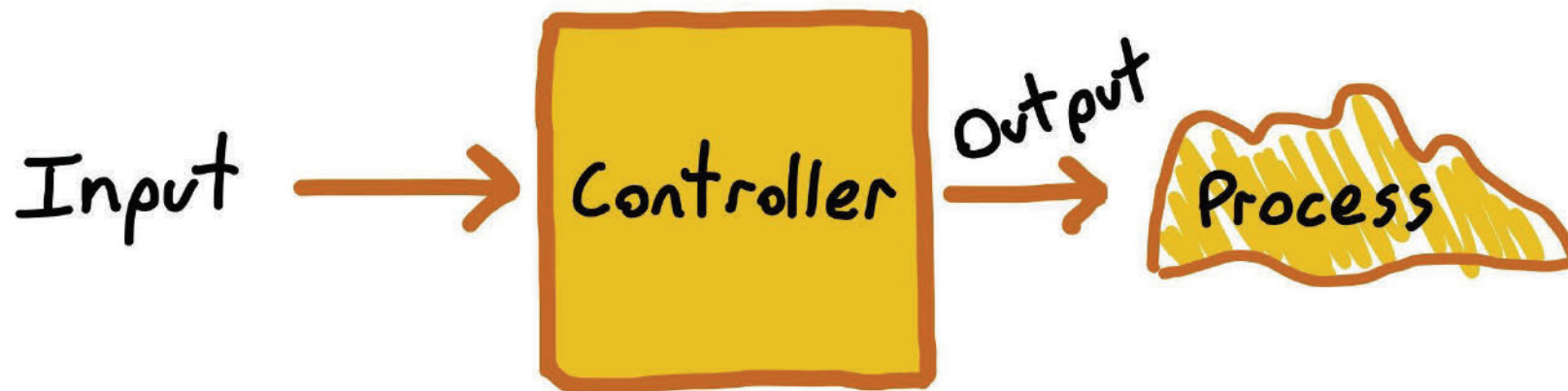
- **Can we make our models fast & accurate enough?**
--- e.g., via reinforcement learning/inference/
- **Can we make our models realistic enough?**
--- e.g., via focused actuator planning with experimental partners

Control Methods with Containers

Ref: Vallery Lancey, Lead DevOps Engineer, "Checkfront"

- Managing a system using human and internal controls
- Inputs dictate what the controller should do (setpoint)
- Outputs dictate what the controlled process should do

Closed Loop Container: (i) Contains feedback from the process to the controller; (ii) Controller able to self-correct to achieve desired outcome



Control System Management

Traditional: a "sysadmin" examines the system, makes a judgement, and performs an action.

Automatic: the system tracks its own state, and translates the state to some internal action.

POSSIBLE FRNN DEPLOYMENT INTO PCS @ DIII-D, JET, KSTAR, ...
(A. Svyatkovskiy,, Princeton U/PPPL/Microsoft)

Suggested Approach: Deploy AI/DL FRNN disruption predictor as a **web service** using "azureml" and Azure Container service (Microsoft)

- 1) Train modes as usual using the FRNN package
- 2) Prepare a "helper code" to deploy the model – (details sketched)
- 3) Interact with the model via "RESTful API" -- by sending input data in as "JSON" operation and Receive prediction as "JSON"
- 4) This approach can serve predictions on the order of a few 100 nano-seconds (<400 nanoseconds), Including network latency *

*** examples available from Microsoft deployment**

In Situ Data Analytics for Next Generation Molecular Dynamics Workflows

Michela Taufer

**Department of Electrical Engineering and Computer Science
The University of Tennessee Knoxville**



THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

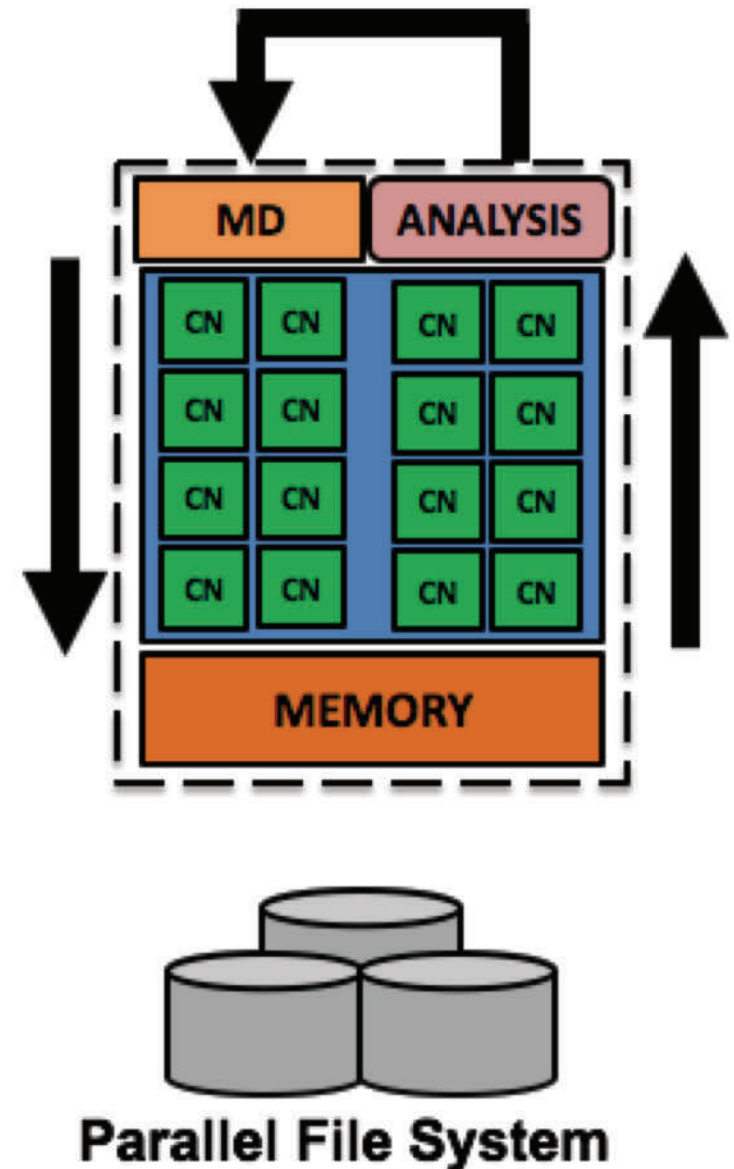
BIG ORANGE. BIG IDEAS.®

Project Overview

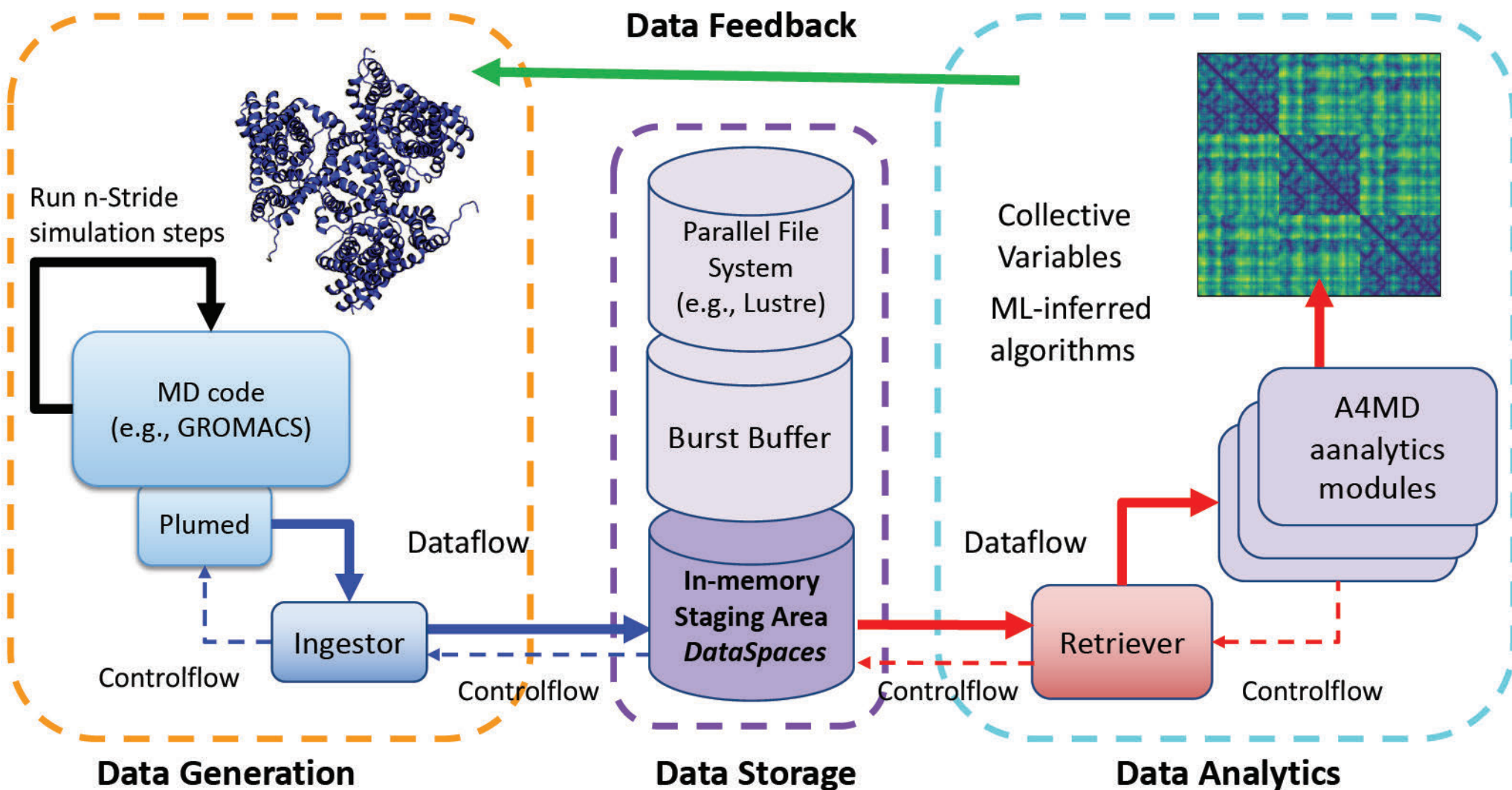
Project goals: (1) create new *in situ* methods to **trace rare events** such as conformational changes in **classical molecular dynamics (MD) simulations at runtime**; (2) design new data representations for **machine learning inferred knowledge** and build an global organization of structural and temporal molecular properties; (3) **integrate simulation and analytics into workflows** for detection of changes in structural and dynamic molecular properties and runtime steering of MD simulations

PIs: *Michela Taufer (UTK), Trilce Estrada (UNM), Ewa Deelman and Rafael Ferreira da Silva (USC), Michel Cuendet and Harel Weinstein (Weill Cornell Medical College of Cornell University)*

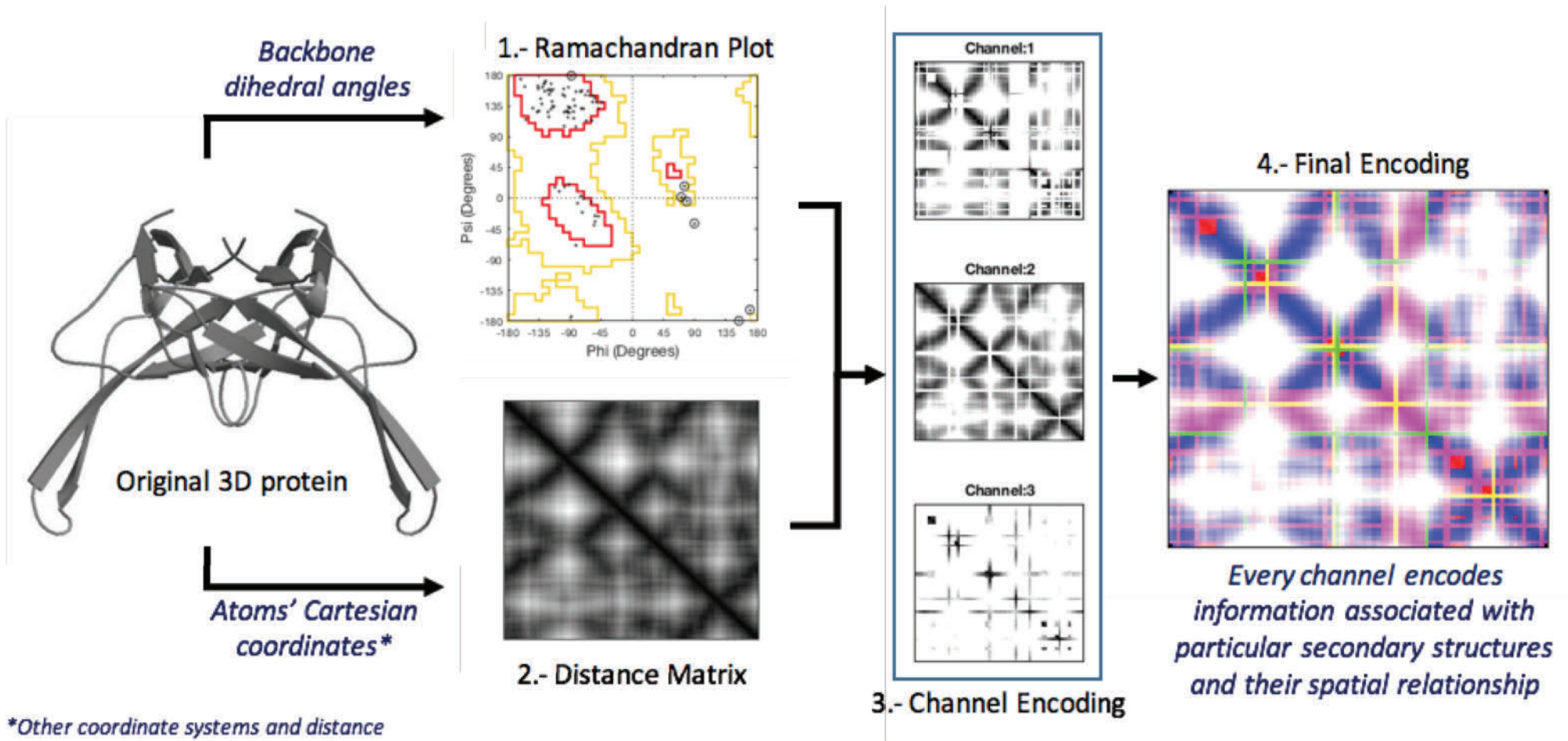
Sponsor: NSF 1841758/1741040/1740990



Building a Close-loop Workflow



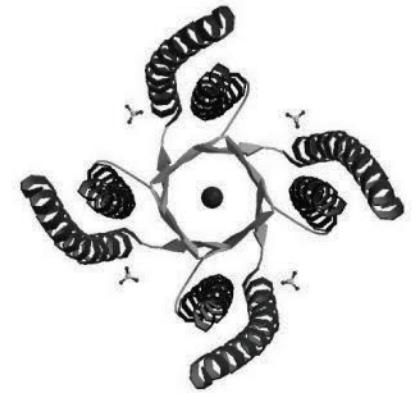
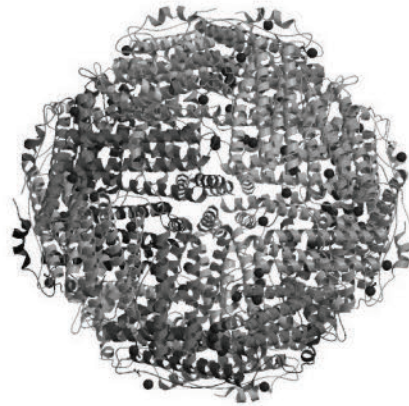
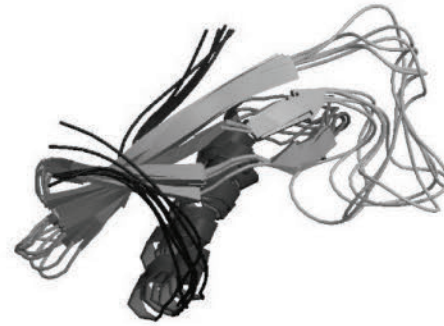
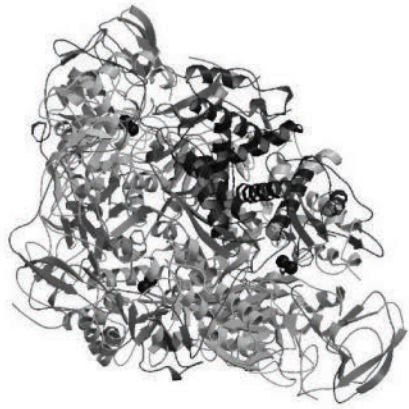
Algorithms for ML-inferred Knowledge



**Other coordinate systems and distance representations could also be used*

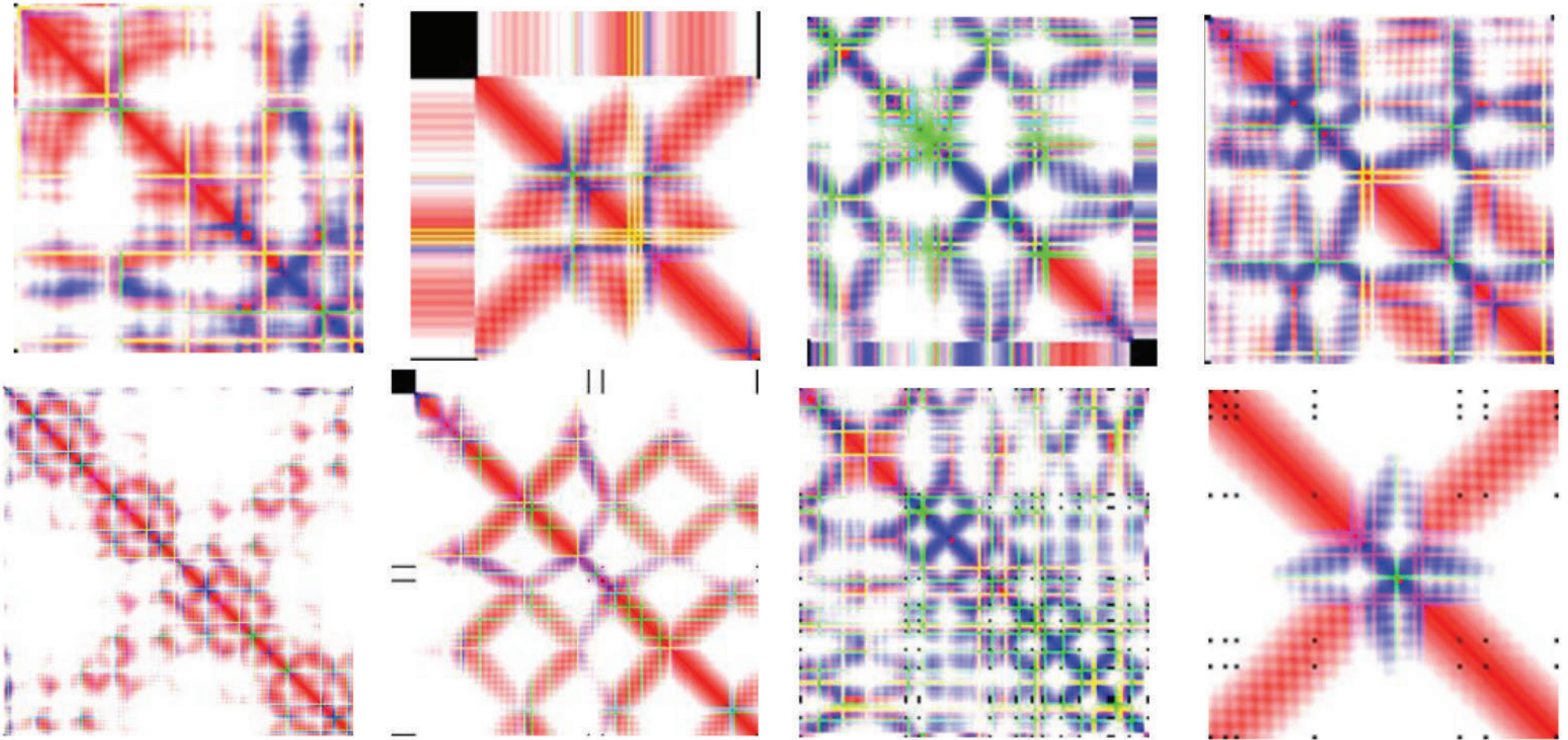
T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer. **Graphic Encoding of Proteins for Efficient High-Throughput Analysis**. ICPP 2018.

From Multi-fold Representation to Image Encoding



T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer.
Graphic Encoding of Proteins for Efficient High-Throughput Analysis. ICPP 2018.

From Multi-fold Representation to Image Encoding



T. Estrada, J. Benson, H. Carrillo-Cabada, A. Razavi, M. Cuendet, H. Weinstein, E. Deelman, and M. Taufer.
Graphic Encoding of Proteins for Efficient High-Throughput Analysis. ICPP 2018.

Challenges and Opportunity

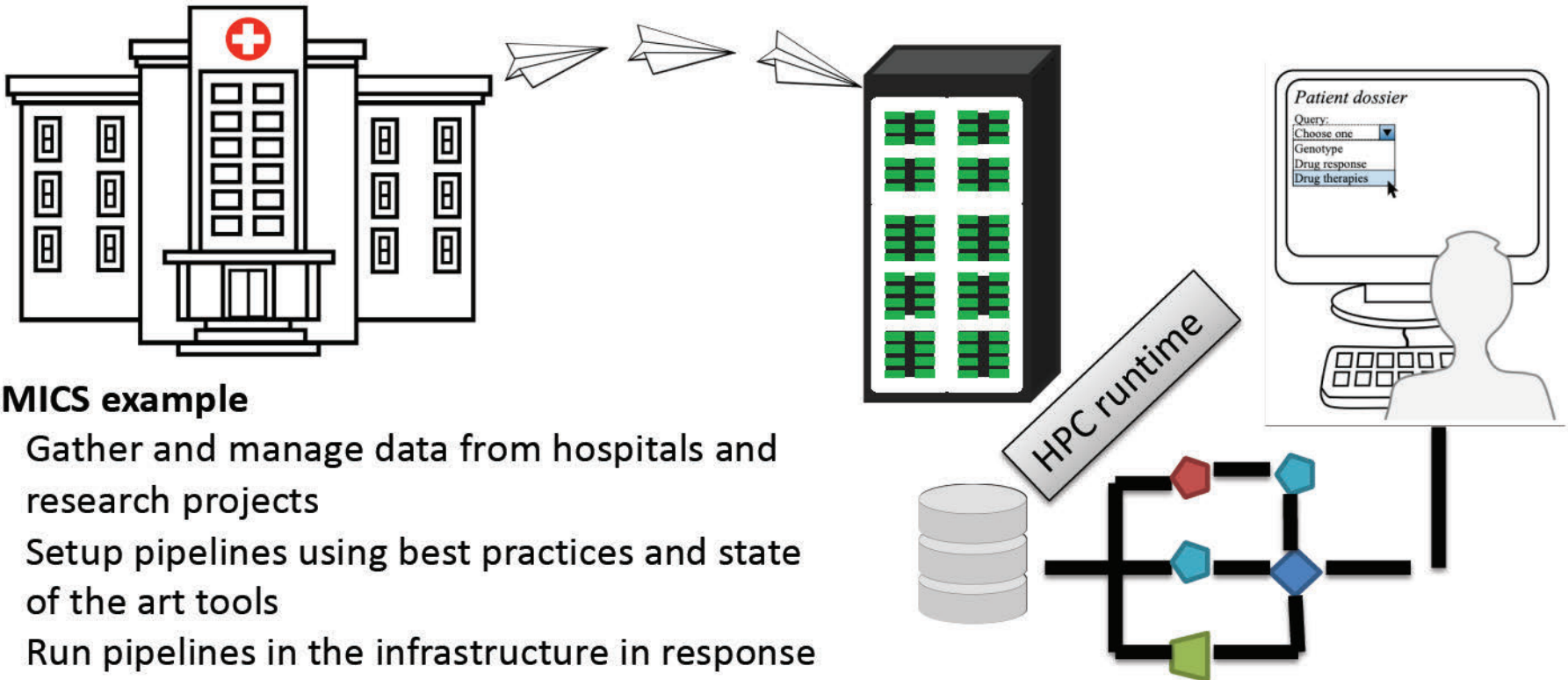
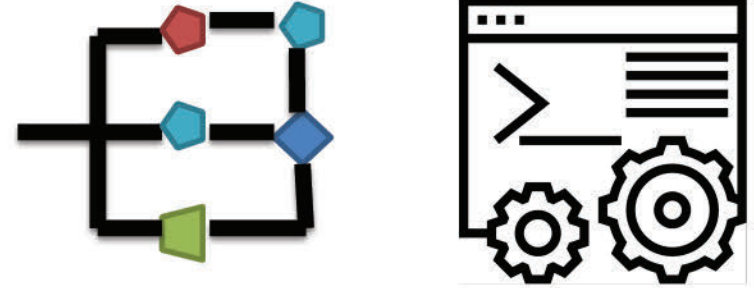
- *Efficiency*: Optimize workflows' performance and power usage associated to data movement and analytics
- *Generality*: Build workflows that support different types of analytics across different MD applications
- *Non-invasive*: Capture data from MD simulations without rewriting legacy codes or simulation scripts
- *Portability*: Execute combined simulations and analytics across different platforms and with heterogenous resources
- *Scalability*: (Re)design ML algorithms for knowledge discovery at scale

Preparing analysis tools



Composing reusable pipelines

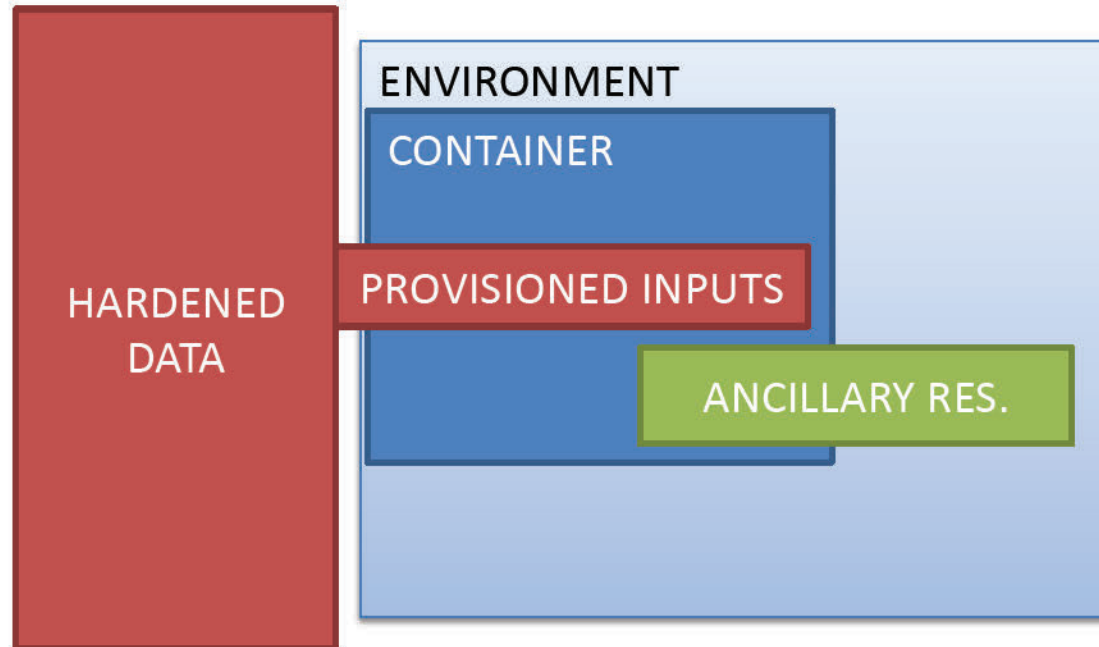
193



OMICS example

- Gather and manage data from hospitals and research projects
- Setup pipelines using best practices and state of the art tools
- Run pipelines in the infrastructure in response to user queries respecting access policies

- Preparing the software tools is often cumbersome:
 - Provisioning tools like puppet might help, also anaconda
- Containers help but:
 - Ancillary resources might make it prohibitively large
- Light containers running on an environment with provisioned ancillary resources helps but:
 - Software resources might need to be compiled through the container and make it awkward
- We are trying a hybrid approach



- Thematic functionalities packaged in pipelines or dependency trees
 - Managed by specific teams
- Possible to tie to particular runtime environments
 - Protect IP of methods and ancillary data
 - Transparent use in the overall picture
- Comfortable development and deployment of updates
 - Develop locally and deployed via containers simply

- Workloads can be very heterogeneous:
 - Expensive long tasks: NGS alignment
 - Exhaustive annotation of annotations: large ancillary resources
 - Creative approaches: Data exploration, interpretation, machine learning, etc
- Workload details and privacy calls for a distributed approach to execution across Cloud/HPC/Workstation
- How to incorporate streaming data is not clear

- Cohort-base queries require:
 - Awareness of data provided. Indices. FAIR Data?
 - Consent. GDPR Compliance?