Increasing Scientific Data Insights About Exascale Class Simulations Under Power and Storage Constraints

James Ahrens, ahrens@lanl.gov, Los Alamos National Laboratory

February 2014, LA-UR-14-20859

Frame the challenging problems of exascale data analysis and visualization

- Notion of a cost per insight in terms of power and storage used
- Challenges the premise of our traditional workflow.

Power constraints

- Target goals for peak performance to increase three orders of magnitude while system power is only targeted to increase by a factor of two
- Most expensive operation is data movement

Storage constraints

- Gap between both capacity/bandwidth and FLOPS will widen
- Storage system of an exascale supercomputer will be proportionally smaller and slower

Traditional post-processing oriented visualization and analysis approach

- Temporal simulation snapshots are saved at regular intervals
  - Saving checkpoints for later restart in case of errors
- Not with power and storage constraints

In situ visualization and analysis

- During the simulation run while the data is resident in memory

Sampling and Uncertainty Quantification of Simulation Data are Needed

- *In situ* data analysis
  - Access to the entire simulation data
  - Including spatial, temporal, multivariate and variable type domains
  - Only available when in memory
  - Form of sampling
- Analyst explicitly samples…

- Example:
  - Stratified random sampling approach
  - MC^3 cosmological particle simulation
  - Analyze the entire particle population
    - Record full population statistics
    - Quantify sample error

Deliberate Analysis Choices Are Necessary

- Traditional approach
  - During a simulation run, full simulation snapshots are saved
  - Belief: Snapshots can answer arbitrary analysis questions - “All the data has been saved”
    - Not necessarily true for the time domain

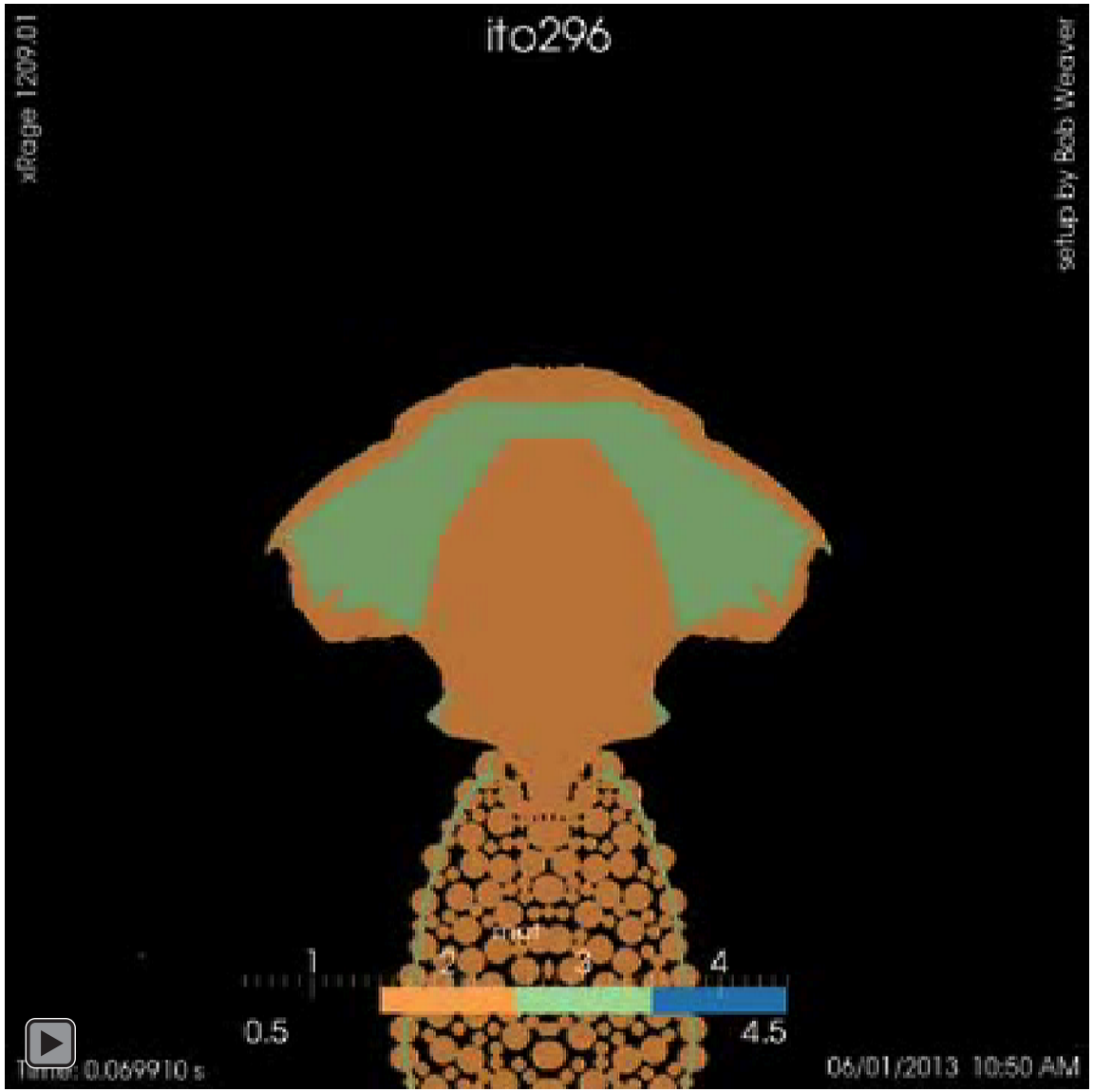
- *In situ* approach
  - Automatic selection of data at runtime
    - Belief: Reduces the type of questions that can be asked about the data during post-processing analysis
  - Make deliberate analysis choices before the simulation is run
    - Constrained by a power and storage budget

- Observational/experimental community -- streaming approaches
  - Accelerator physics, fusion reactors and cyber-security
  - Pre-planned data reducing streaming analysis is common practice
  - Custom software and hardware accelerators
    - Typically employed to reduce and analyze data in real-time

- Key research questions to answer are:
  - How general and with what quality can analysis questions be answered from:
    - Compact data products generated *in situ,* in a post-processing manner?
  - What new mathematical or analysis techniques will support this process?

Data Reduction and Prioritization Is Required

- Simulation data stream significantly reduced into a compact analysis product
  - To fit within the budget
- Collect most important data
  - Prioritization
- For example:
  - Measured spatial and temporal entropy in a running simulation
  - A memory buffer collected time steps
    - higher entropy overwrite ones with lower entropy.
  - Summary of the phases of the simulations in which the most change occurs



Belmont Forum - J.-P. Vilotte, J.Y. Berthou, P. Monfray, M. Girerd

13 world's major and emerging national research agencies and international science councils.

Belmont Challenge

- Priority knowledge supporting action on **societal environmental change challenges**
- Coordinate and stream line international efforts (research, infrastructures)

"To deliver knowledge needed for action to avoid and adapt to detrimental environmental change including extreme hazardous events."

E-Infrastructure and Data Management collaborative research actions (CRA).

- Address the Belmont Challenge priorities.
- Lever Belmont Forum members existing investments through international added value.
- Bring together leading earth & environment scientists, social scientists, computer and research informatics scientists, and users.

Knowledge Hub (http://bfe-inf.org/)

- Community building and strategy development
- Environmental data management and interoperability
- Data exploitation and valorisation: synergistic Data and HPC infrastructures

Build up on existing projects and consortia

Knowledge Hub: Work Packages

WP1: Standards (R. Cossu & M. Mokrane)

- *Integration and interoperability of heterogeneous multidisciplinary datasets*
- *Data platforms allowing exploration and mining of data and derived-data*
- *Credit and Trust of derived research findings through provenance management*

WP2: Data & Compute Infrastructures

(J.-P. Vilotte, T. Koike)

- *Exploitation and valorisation of data generated by observational and extreme-scale simulations;*
- *Synergistic challenges between existing Data and Compute e-infrastructures*
- *End-to-end workflows and data movement across Data and HPC e-infrastructures*

WP3: Harmonisation of Data Infrastructures (C. Waldmann)

- *Global Data Infrastructures interactions and governance;*
- *Minimal common core services*
- *Thematic and Integration core services and tools*

WP4: Data Sharing (D. Peters, A. Treloar)

- *Values and incentives for data sharing and management;*
- *Trust, data quality and curation,*
- *Legal issues: data providers IP rights, liability of infrastructure management*

W5: Open Data Policy

(B. Gemeinholzer, A. Treloar)

- *Values and incentives for Open Data and Open Science for environmental change challenges;*
- *Enable citizen science and crowd sourcing*
- *Liability and uncertainty in decision making*

WP6: Capacity Building (L. Allison, R. Gurney)

- *Holistic education and training of new generation data-intensive scientist and data curators in environmental sciences;*
- *Security issues and legal frameworks*
- *Sustainable human resource*

futurearth
research for global sustainability

RDA Research Data Sharing without barriers
RESEARCH DATA ALLIANCE

COOPEUS

EarthCube

GEO GROUP ON EARTH OBSERVATIONS

WP2: Interfaces between Data and HPC Infrastructures

Coordinators: Jean-Pierre Vilotte & Toshio Toike
(Roberto Cesar, Andrew Treloar)

Topics:

- *Enable exploitation and valorisation of large volumes of* **data generated by high-throughput instruments, observational and monitoring systems, extreme-scale computing***;*
- *Synergistic challenges between existing Data and Compute e-infrastructures*
- *End-to-end workflows and data movement across Data and HPC e-infrastructures*

- **Data movement across Data and HPC e-infrastructures**
- **Big Data analytics and data-intensive extreme computing;**
- **Orchestrated data-streaming and data-shipping workflows and execution models;**
- **Distributed parallel staging and compute data management that meets data life-cycles;**
- **Providers policy and AAI services**
- **Data provenance and Data identifiers**

Objectives

- Identify existing related **groups and projects** in Earth environment sciences and Natural Hazards including **e-infrastructure providers**
- Bring together **users and experts** on Data and HPC infrastructures
- Survey of **good practices** and identify **use cases or proxy mini-apps**
- Identify **synergistic gaps and barriers** between Data and HPC infrastructures in support of orchestrated data-intensive and extreme-computing workflows
- Innovative tools and methods for complex **Big Data analytics**
- Barriers in **integrated services** across data and compute e-infrastructures
- User-driven performance and quality indicators for Data and HPC infrastructures interface

WP2: Interfaces between Data and HPC Infrastructures

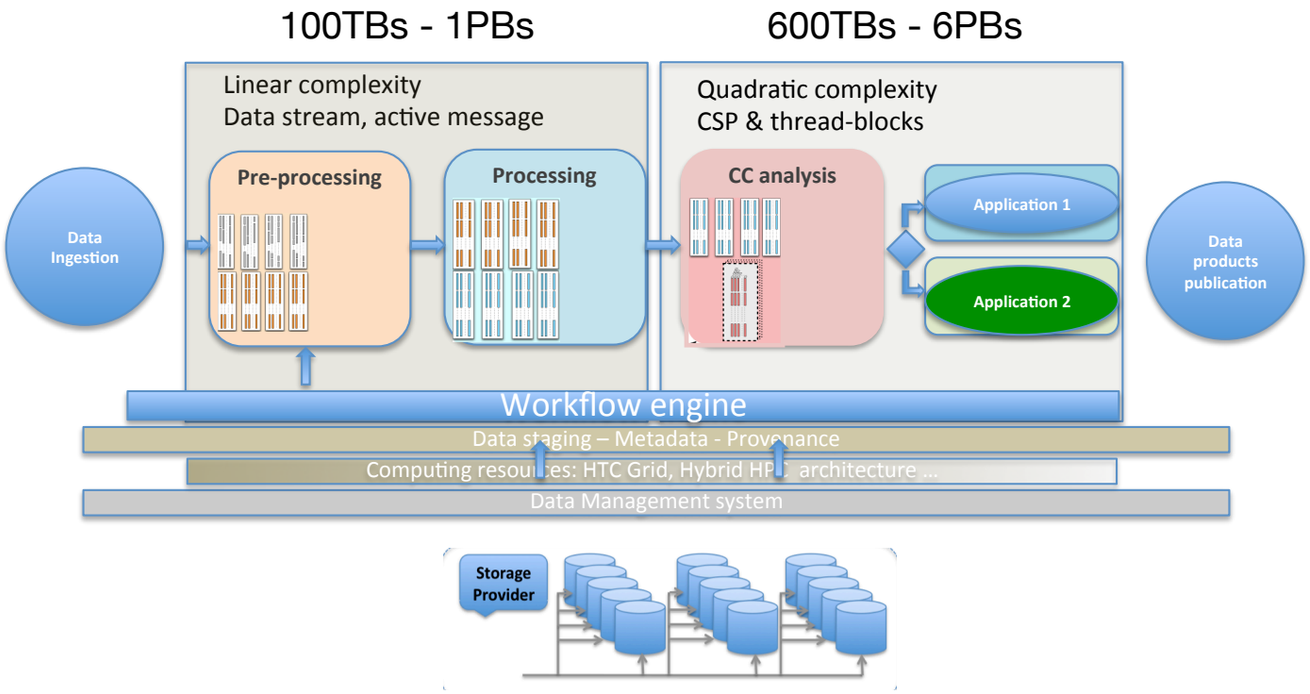
Data centres

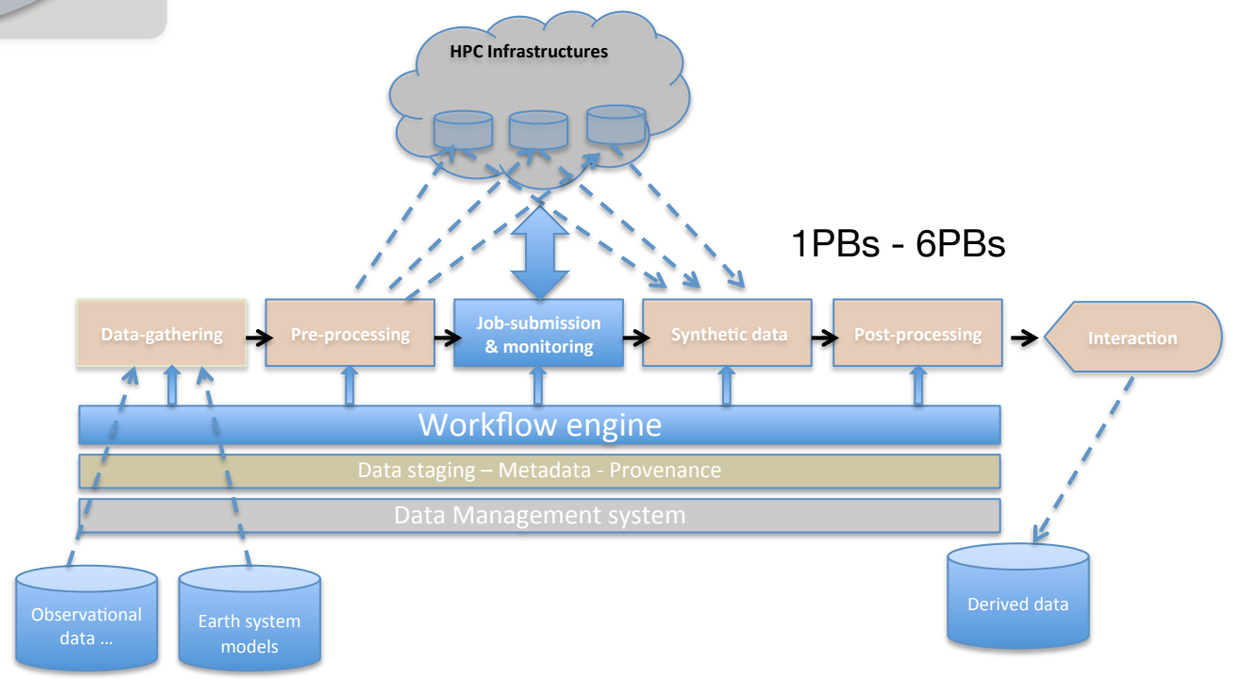
- data and derived data archiving and preservation
- Curation, annotation, PIDs, Provenance
- data,meta-data, distribution standards
- continuous or real time data stream
- complex and multi-disciplinary data

HPC Data staging infrastructure

- Complex data-intensive analysis
- In-situ data production: (ensemble) simulation, assimilation, inversion (particle filtering)
- Staging storage management (safe replication, **data life cycles**)
- Distributed compute storage management (fast and large number sequential IOs, vertical reuse)

End-to-end Workflows



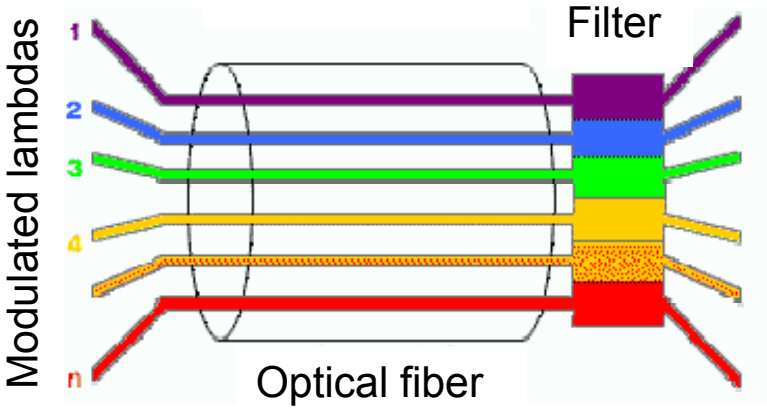


- Data streaming and data shipping workflow engines, different execution models
- Orchestrated workflows and data movement across infrastructures
- Large semi-structured binary objects with fine-grained access dynamically reconfigurable
- Data movement across infrastructures (AAI policy), data provenance and workflow metadata

Impact of huge bandwidth optical interconnection network and a new memory paradigm with global address space for BDEC systems

Tomohiro Kudoh, Shu Namiki, Ryousei Takano, Kiyo Ishii, Yoshio Tanaka, Isao Kojima,
Tsutomu Ikegami, Satoshi Itoh, Satoshi Sekiguchi
National Institute of Advanced Industrial Science and Technology (AIST)

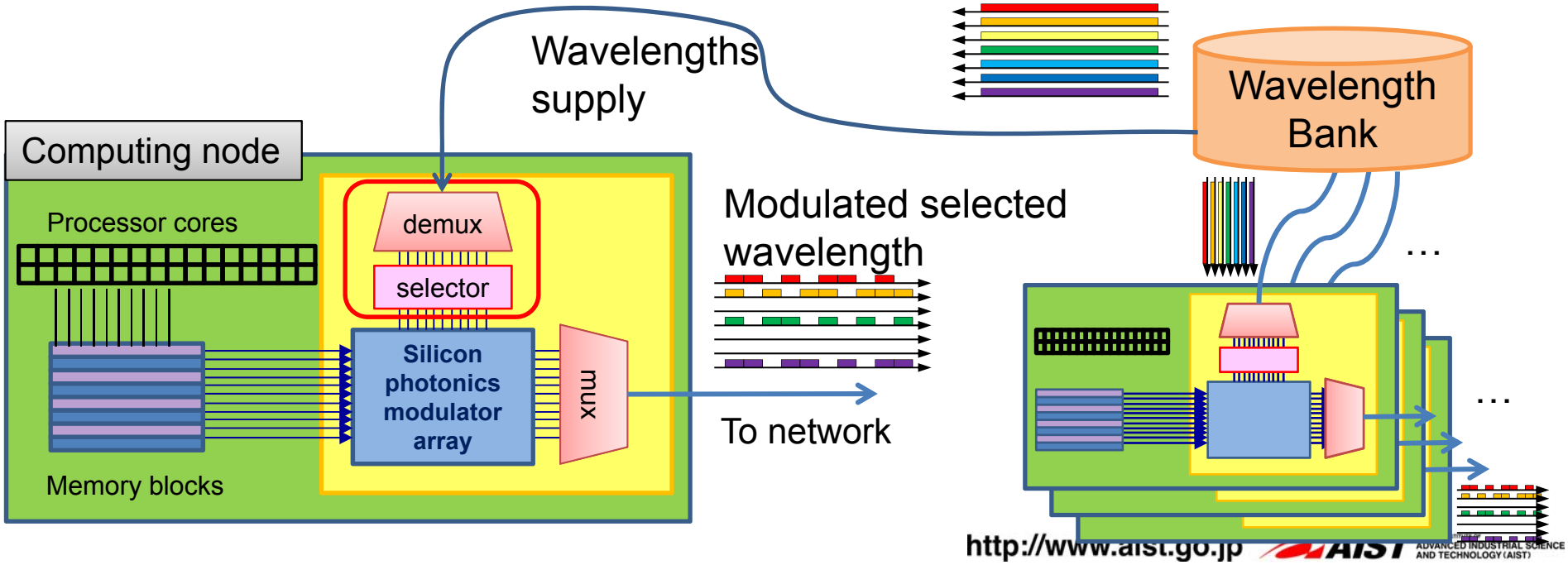
- Conventional HPC systems: the inter-node I/O bandwidth is about 1/10 of the intra-node memory access bandwidth
  - Pin-bottleneck will increase the gap
- Breakthrough in interconnect bandwidth is possible by Dense Wavelength Division Multiplexing (DWDM)
  - cf. 50Gbps x 100 lambda = 5Tbps/fiber
- Problem: DWDM Light source
  - Expensive compound semiconductor
  - Produce heats: Precise temperature control needed for DWDM
- Solution: Wavelength Bank
- Need a new architecture and software to utilize huge interconnect bandwidth
  - Interconnect bandwidth can be comparable to or greater than memory bandwidth



Wavelength Division Multiplexing (WDM)

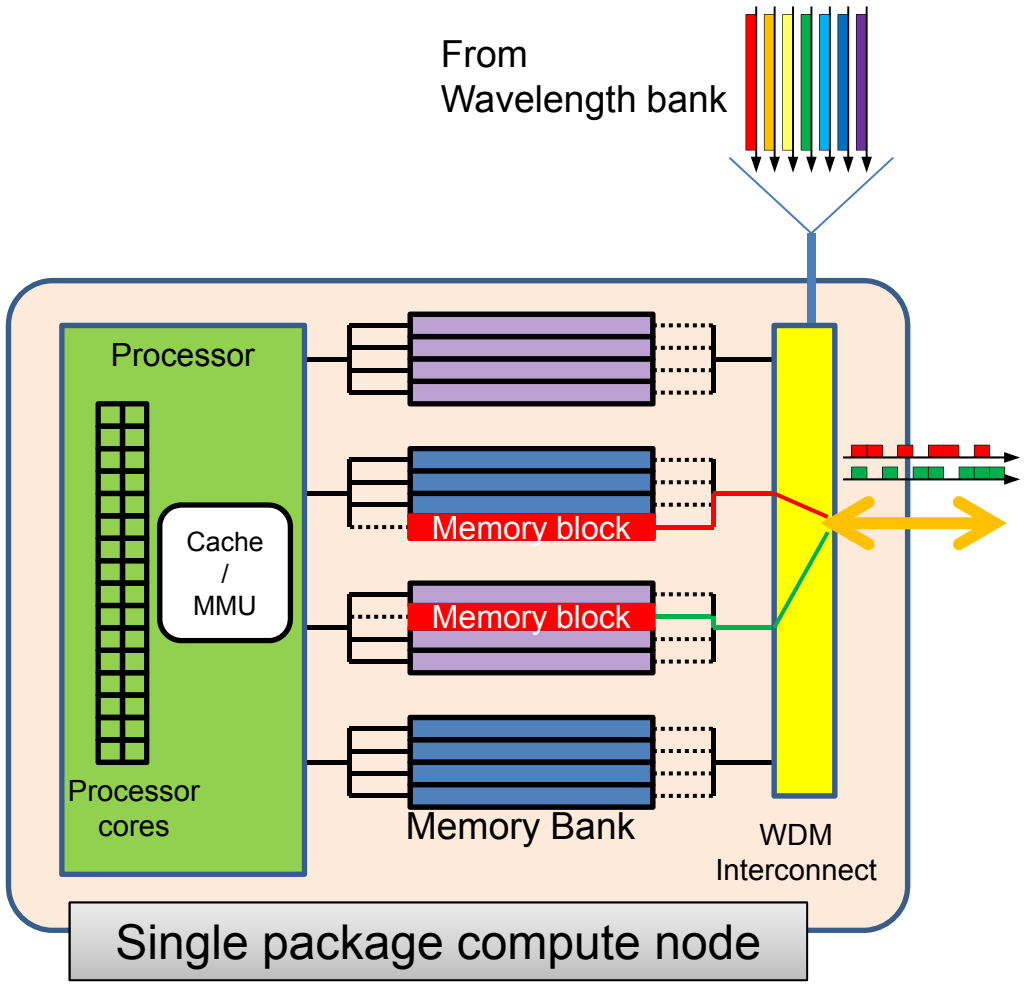
Wavelength Bank and Silicon Photonics

- Wavelength bank (WB) or optical comb source is a centralized generator of wavelengths for DWDM. One WB in a BDEC system.
- Light waves are distributed to computing nodes using optical amplifiers (loss compensation)
  - No light sources are required for each computing node.
  - Distributed light is de-multiplexed to each wavelength, modulated, multiplexed again, and transmitted from each computing node.
- Silicon photonics optical circuits can be used for the whole light wave processing at a node. Low cost and low power consumption. Hybrid implementation with electronics.
- DWDM signals can be switched in one bundle by fiber cross-connect switches, or can be switched separately by wavelength selective switches.



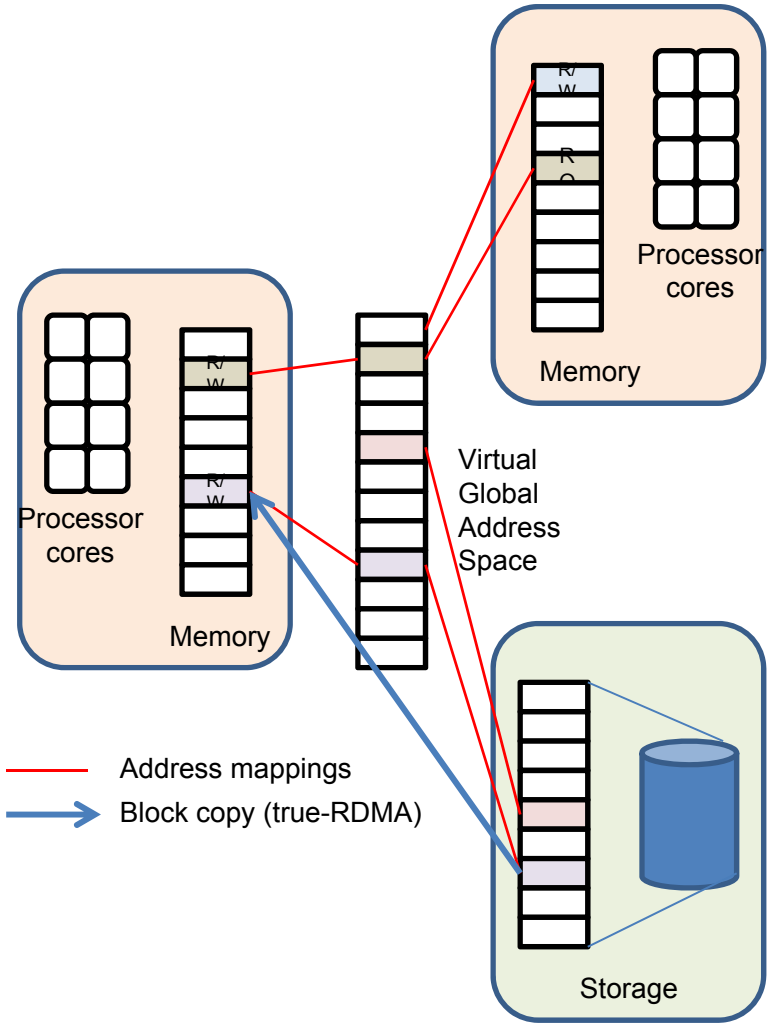
Direct memory copy over DWDM

- Assume processor–memory embedded package with WDM interconnect
- To fully utilize the huge I/O bandwidth realized by DWDM
  - Allow direct access to memory by I/O.
  - Main memory is divided into memory blocks
    - Each memory block can be accessed either from the processor or the I/O at a time.
- Multiple memory blocks can be sent/received simultaneously using multiple wavelengths.
  - For 4MB blocks, a block can be transferred in about 1ms at the rate of 50Gbps
  - Parallel transfer of up to the number of wavelength channels is possible.



Programming model and OS

- Software architecture including operating systems, programming models and memory systems should be re-designed.
- Management of memory blocks
  - Map memory blocks to a global virtual address space.
  - Storage also mapped to the address space.
  - Memory block transfer: RDMA operations
- Impact on the structure of an operating system and runtime systems.
  - Kernel organization (e.g., hybrid of light-weight and general purpose kernels)
  - Data access abstraction on a global virtual address space
  - Fault tolerant/resilience
  - Network resource management (e.g., optimal wavelength scheduling and optical path switching/routing)
- We will conduct a feasible study of the design of both architecture and system software of such system..



HELIX NEBULA
THESCIENCECLOUD

Strategic collaboration between HPC and Cloud to address big data in global scientific challenges

BDEC Workshop, Fukuoka – 28 February 2014

Maryline Lengert (ESA), Bob Jones (CERN), David Foster (CERN), Steven Newhouse (EMBL-EBI)

A European cloud computing partnership: big science teams up with big business

HELIX NEBULA THE SCIENCE CLOUD

Strategic Plan

- Establish multi-tenant, multi-provider cloud infrastructure
- Identify and adopt policies for trust, security and privacy
- Create governance structure
- Define funding schemes

CERN

To support the computing capacity needs for the ATLAS experiment

EMBL

Setting up a new service to simplify analysis of large genomes, for a deeper insight into evolution and biodiversity

esa DLR cnes

To create an Earth Observation platform, focusing on earthquake and volcano research

PIC port d'informació científica

To improve the speed and quality of research for finding surrogate biomarkers based on brain images

Suppliers

Atos CGI CloudSigma CSA cloud security alliance DANTE www.dante.net egi indra interoute SAP sixsq SWITCH T-Systems terradue 2.0 the SERVER LABS the IT architects

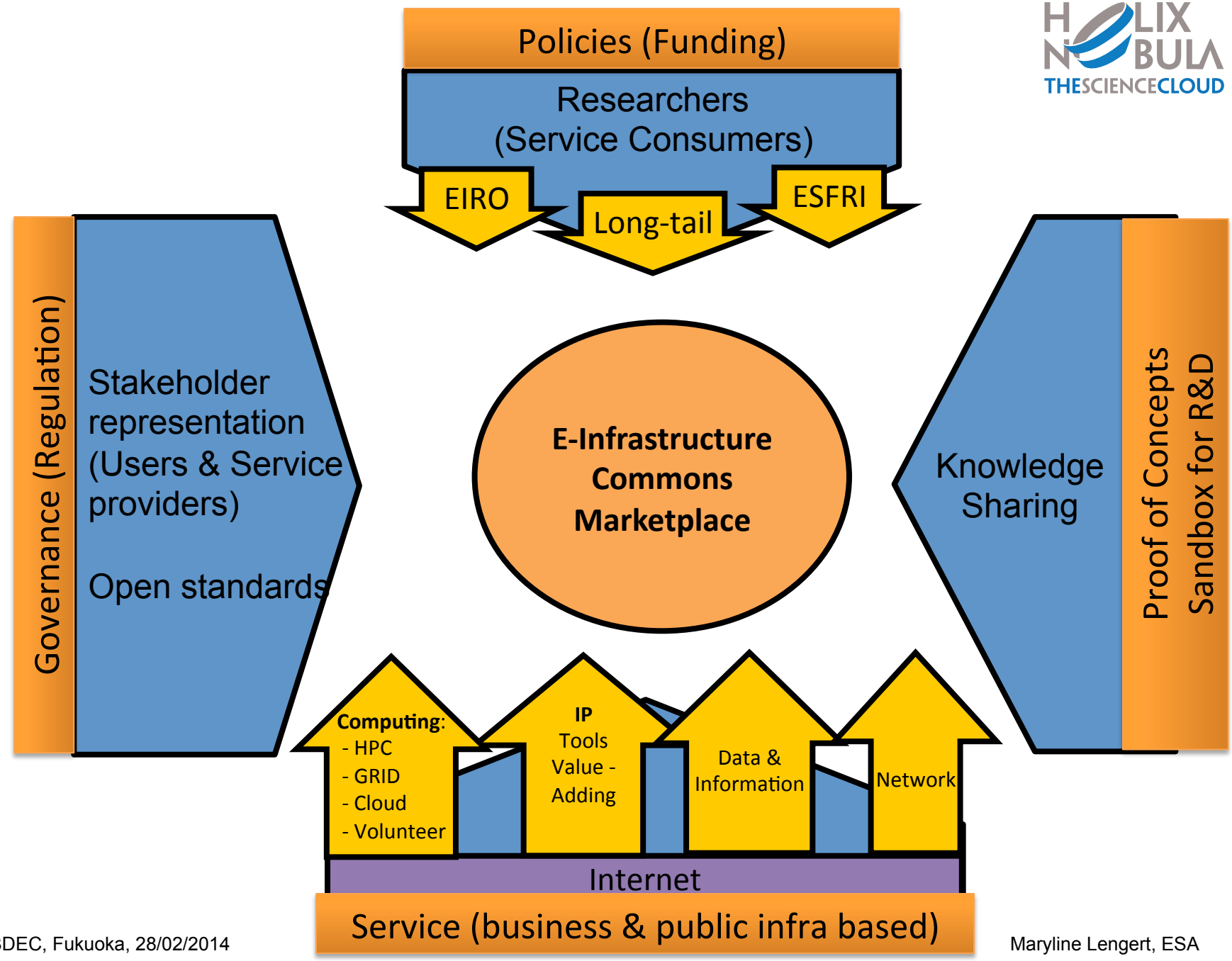
Adopters

AWST Capgemini CONSULTING.TECHNOLOGY.OUTSOURCING cnrs ECMWF Ifremer OpenNebula.org The Open Source Toolkit for Cloud Computing orange Business Services THALES Telefónica Trust-IT Communicating ICT to Markets www.trust-itservices.com

Vision

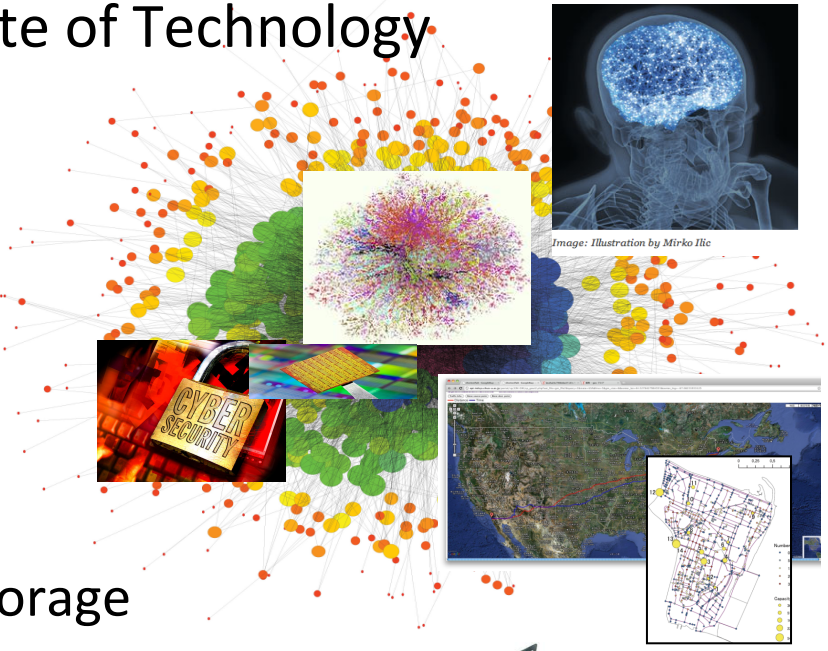
The e-commons infrastructure marketplace

- Will provide access to worldwide and world class resources through a **dynamic and sustainable marketplace**.
- Be built on public and commercial assets, will cover the **entire scientific workflow**
- Will offer the broadest range of services
- Will ensure use of **open standard** and **interoperability** of service providers while adhering to **European policies, norm and requirements.**


Policies (Funding)
Researchers
(Service Consumers)
EIRO
Long-tail
ESFRI
Governance (Regulation)
Stakeholder representation (Users & Service providers)
Open standards
E-Infrastructure Commons Marketplace
Knowledge Sharing
Proof of Concepts Sandbox for R&D
Computing:
- HPC
- GRID
- Cloud
- Volunteer
IP
Tools
Value - Adding
Data & Information
Network
Internet
Service (business & public infra based)

Towards Extreme-scale Graph Processing with Deepening Memory Hierarchy

Hitoshi Sato, Tokyo Institute of Technology



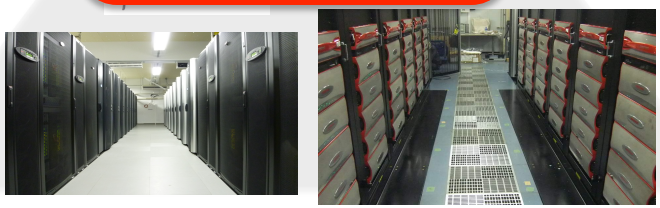
- Large-scale Graphs and HPC
  - Various Applications
    - Traffic network, SNS, Smart Grids, Biology, Cyber-security, etc.
  - Modern supercomputers
    - can accommodate peta-flops class performance w/ peta-byte class storage
  - Important Kernels for Big Data HPC
    - Graph500/Green Graph500

GRAPH 500 Green Graph

NVMs is a key device for I/O systems

cf. TSUBAME2, Catalyst, Gordon, etc.

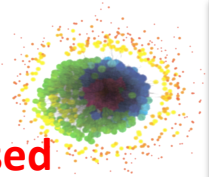
How to utilize deepening memory/ storage hierarchy



Large-Scale BFS Using NVMs for Graph500 [SC13 Poster]

Motivation

- Large scale graph processing in various domains
  DRAM resources has increased



- Spread of Flash Devices

Prof : Price per bit, Energy consumption
Cons: Latency, Throughput

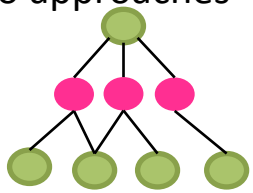
Using NVMs for large scale graph processing has possibilities of minimum performance degradation

NUMA-optimized Hybrid-BFS (Breadth-first Search)

Switching two approaches [Beamer2012] [Yasui2013]

Top-down

$$n_{frontier} < \frac{n_{all}}{\beta}$$



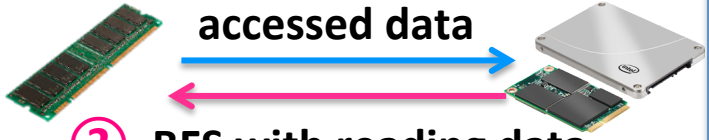
Bottom-up

$$n_{frontier} > \frac{n_{all}}{\alpha}$$

# of frontiers:$n_{frontier}$, # of all vertices:n_{all}, parameter : α, β

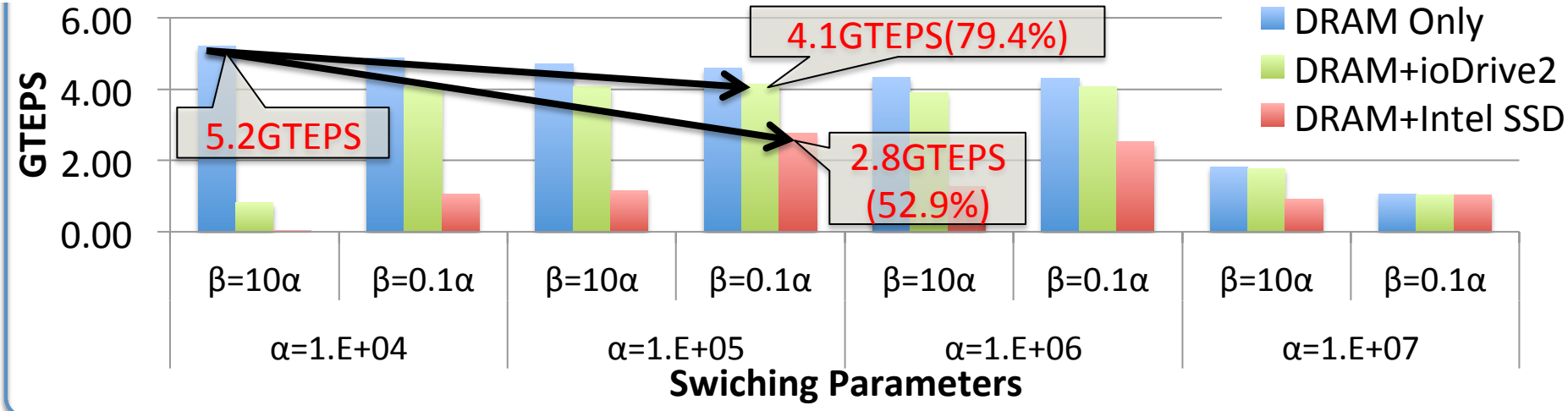
Proporsal

① **Offloading infrequently accessed data**



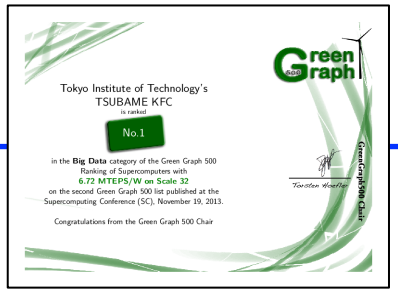
② **BFS with reading data from NVM**

Evaluation (Offload Top-down Graph : We could reduce half the size of DRAM [128GB -> 64 GB] on Scale 27)



The 2nd Green Graph500 list on Nov. 2013

- Measures power-efficient using **TEPS/W** ratio
- Results on various system such as **Huawei's RH5885v2 w/ Tecal ES3000 PCIe SSD 800GB * 2 and 1.2TB * 2**
- **http://green.graph500.org**



In the **Big Data** category:

Rank	MTEPS/W	Site	Machine	G500 rank	Scale	GTEPS	Nodes
1	**6.72**	Tokyo Institute of Technology	TSUBAME KFC	47	32	44.01	32
2	**5.41**	Forschungszentrum Julich (FZJ)	JUQUEEN	3	38	5848	16384
3	**4.42**	Argonne National Laboratory	DOE/SC/ANL Mira	2	40	14328	32768
4	**4.35**	Tokyo Institute of Technology	EBD-RH5885v2	96	30	3.67	1
5	**3.55**	Lawrence Livermore National Laboratory	DOE/NNSA/LLNL Sequoia	1	40	15363	65536
6	**1.89**	Research Center for Advanced Computing Infrastructure	altix	50	30	37.66	1
7	**0.73**	Mayo Clinic	grace	68	31	10.32	64

Lessons from our Graph500 activities

- We can efficiently process large-scale data that exceeds the DRAM capacity of a compute node by utilizing commodity-based NVM devices
- Convergence of practical algorithms and software implementation techniques is very important
- Basically, BigData consists of a set of sparse data. Converting sparse datasets to dense is also a key for performing BigData processing

www.bsc.es

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC vision on Big Data and Extreme Scale Computing

Jesús Labarta, Eduard Ayguadé, Rosa M. Badia, Yolanda Becerra, David Carrera, Toni Cortés, Adrian Cristal, Fabrizio Gagliardi, Alex Ramírez, Enric Tejedor, Jordi Torres, Osman Unsal and Mateo Valero

BDEC, February 28th 2014

Big Data related projects @ BSC

- Applications
  - Molecular dynamics, docking, genomics
  - Air quality, oil exploration
  - Physiological simulation (heart, brain,), neurology
  - Social graph, smart cities
  - HPC performance analysis

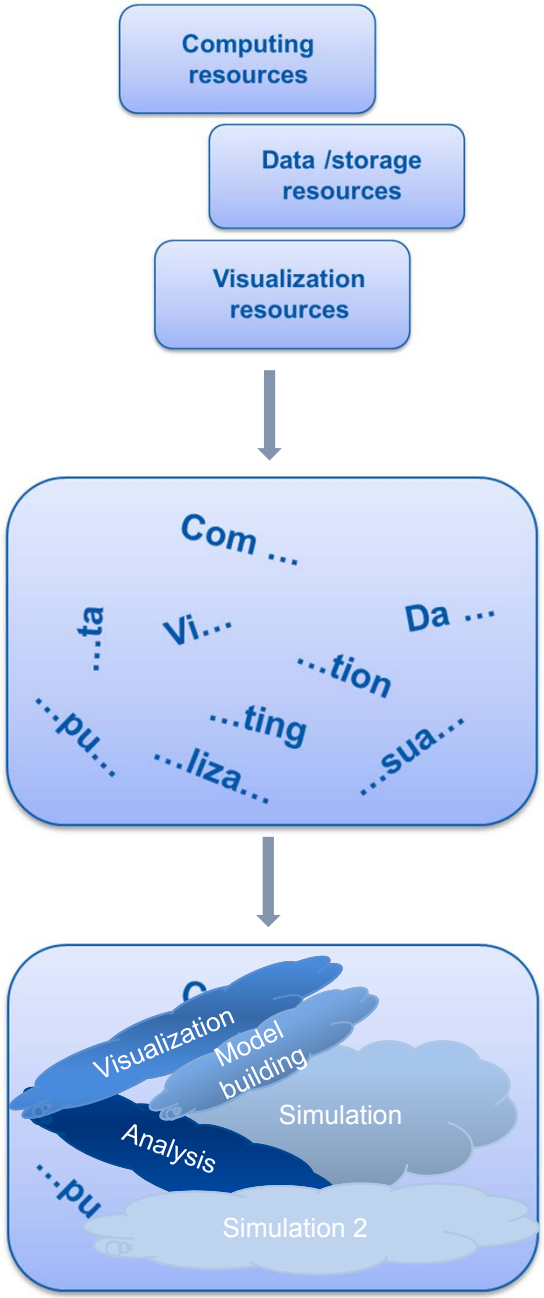
- Observation
  - Broad experience and background **fragmented**

- **Severo Ochoa** project: co-design, unification effort
  - Life science, Earth Science, Engineering, Computer science depts
  - **Unified, productive, easy to use and efficient environment for our broad range of applications**

EXCELENCIA SEVERO OCHOA

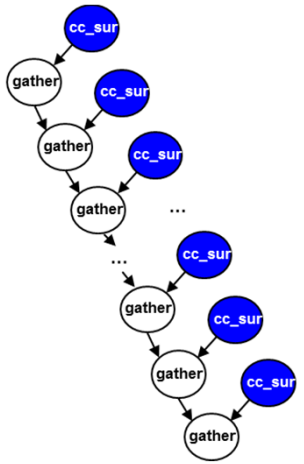
Strategic Considerations

- Architectural support
  - From crystals to plastics
  - Capacity, bandwidth dimensioning = f(technologies, runtime)
- Algorithms: computational and communication complexity
  - Computational and data movement complexity awareness
  - Flexible computational workflows
- Programming models
  - Need more integration between concurrency and data processing
  - Need to close gap between persistent and program data models
  - Simple/minimal extension of existing languages
  - Allow for holistic optimization
  - Clean interface to convey useful information to the runtime
- Usage models and resource management
  - Dynamic, interactive
  - Malleability and dynamic resource management
- Intelligent runtimes
  - Should be given high responsibility to jointly manage data (placement, replications, transfer, query optimization, .) and scheduling



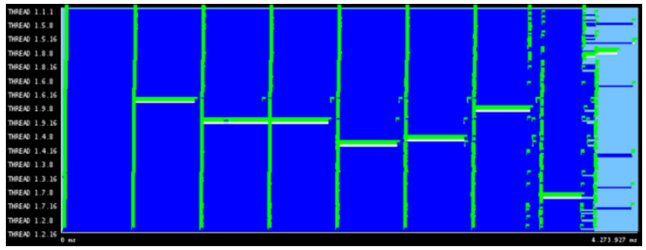
BSC technologies

- StarSs concept (Tasks + directionality annotations)
  - Computational workflow: COMPSs (Java, **PyCOMPSs**)
    - Parallelization of sequential Python code
  - Parallel computing and accelerators: **OmpSs**
    - C, C++, FORTRAN, CUDA, OpenCL

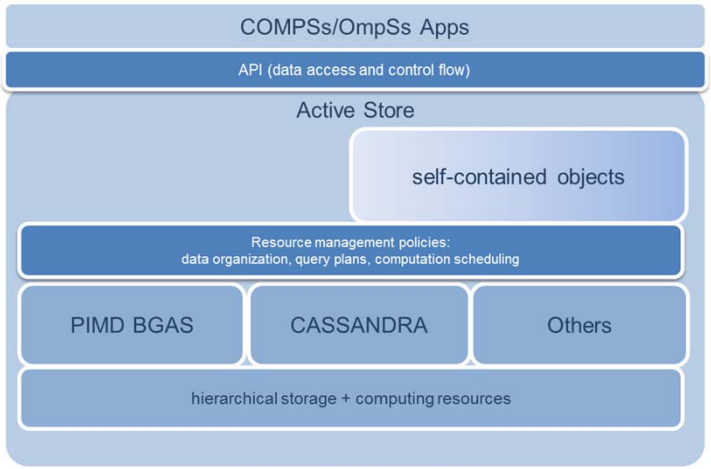


- Persistent share object model
  - Integrate experience in Cassandra, BGAS, in PyCOMPSs
  - Enrichment

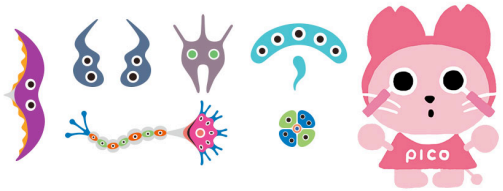
- Intelligent runtime & resource management
  - COMPSs, NANOS, interposed libs,
  - Concurrency
    - Within node, across nodes and cloud
  - Locality
    - Data placement transfer and query optimization
  - Dynamic load balance (DLB)



- Algorithms
  - Genomics
  - Performance analytics
  - Graph analytics



世界一小さいものが見えるX線レーザー　ピコスコープ　SACLA



SACLA and the K Computer

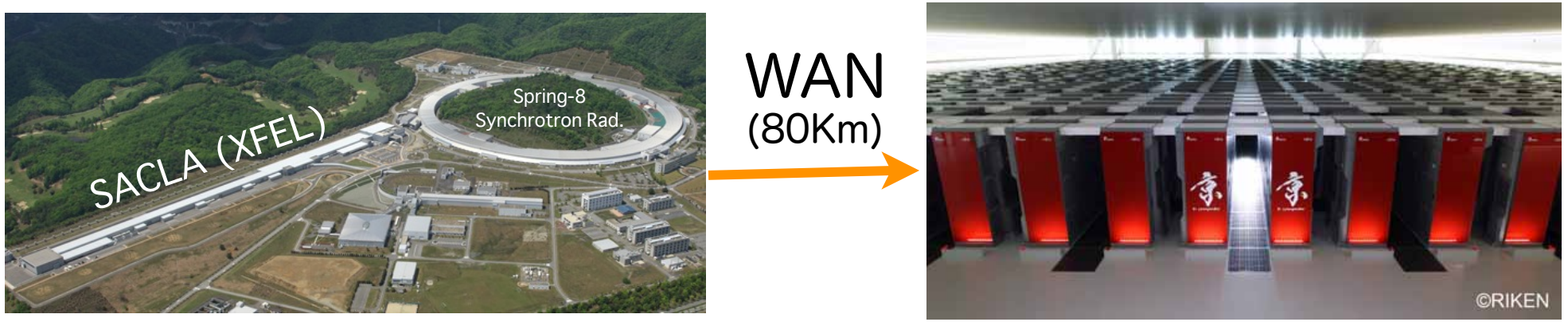
A. Hori (RIKEN AICS)

A. Tokuhisa (RIKEN AICS)
T. Kameyama (RIKEN AICS)
K. Okada (JASRI/RIKEN RSC)
M. Yamaga (JASRI/RIKEN RSC)
Y. Joti (JASRI/RIKEN RSC)
M. Yabashi (RIKEN RSC)
Y. Ishikawa (Univ. Tokyo / RIKEN AICS)
K. Yoshinaga (RIKEN AICS)
J. ARAI (Univ. Tokyo -> NTT)
T. Sugimoto (JASRI)
R. Tanaka (JASRI/RIKEN RSC)
T. Hatsui (RIKEN RSC)
Y. Sugita (RIKEN AICS)
N. Go (JAEA)

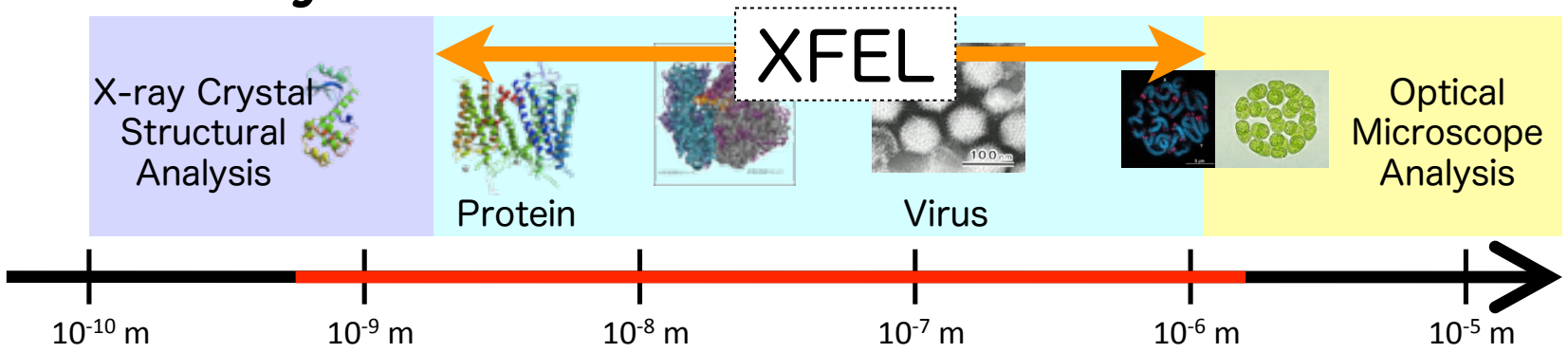
BDEC'14@Fukuoka

SACLA and the K

- Data Acquisition: SACLA (XFEL Facility)
  XFEL: X-ray Free Electron Laser
- Data Processing: the K computer



- To Analyze 3D Structure of Particles



XFEL Project and ...

- Few years ago,
  - SACLA can produce lots of data
  - The K computer is used to process and analyze the data
  - We have succeeded to optimize the program
    - Minimizing I/O, Balancing Load
- Few months ago,
  - RIKEN Center for Life Science and Technologies contacted me
  - New Electron Microscope yields lots of data
  - Their computing environment is too weak ...
- Scientists are flooded by BIG DATA !!

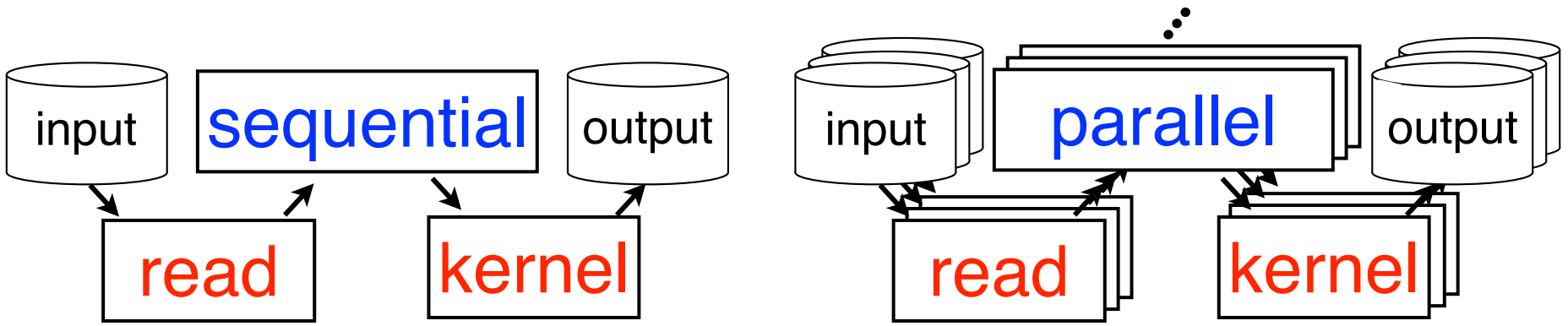
Rescue Mission

- Generalization to handle *any* all-to-all data processing
  (e.g. Transmission Electron Microscope (TEM))

➡ Decoupling kernel code and parallelizing (MPI) code

- User develops sequential programs
  - file read and sequential kernel code
- Easy-to-develop, easy-to-debug,
  easy-to-port, and programing language free

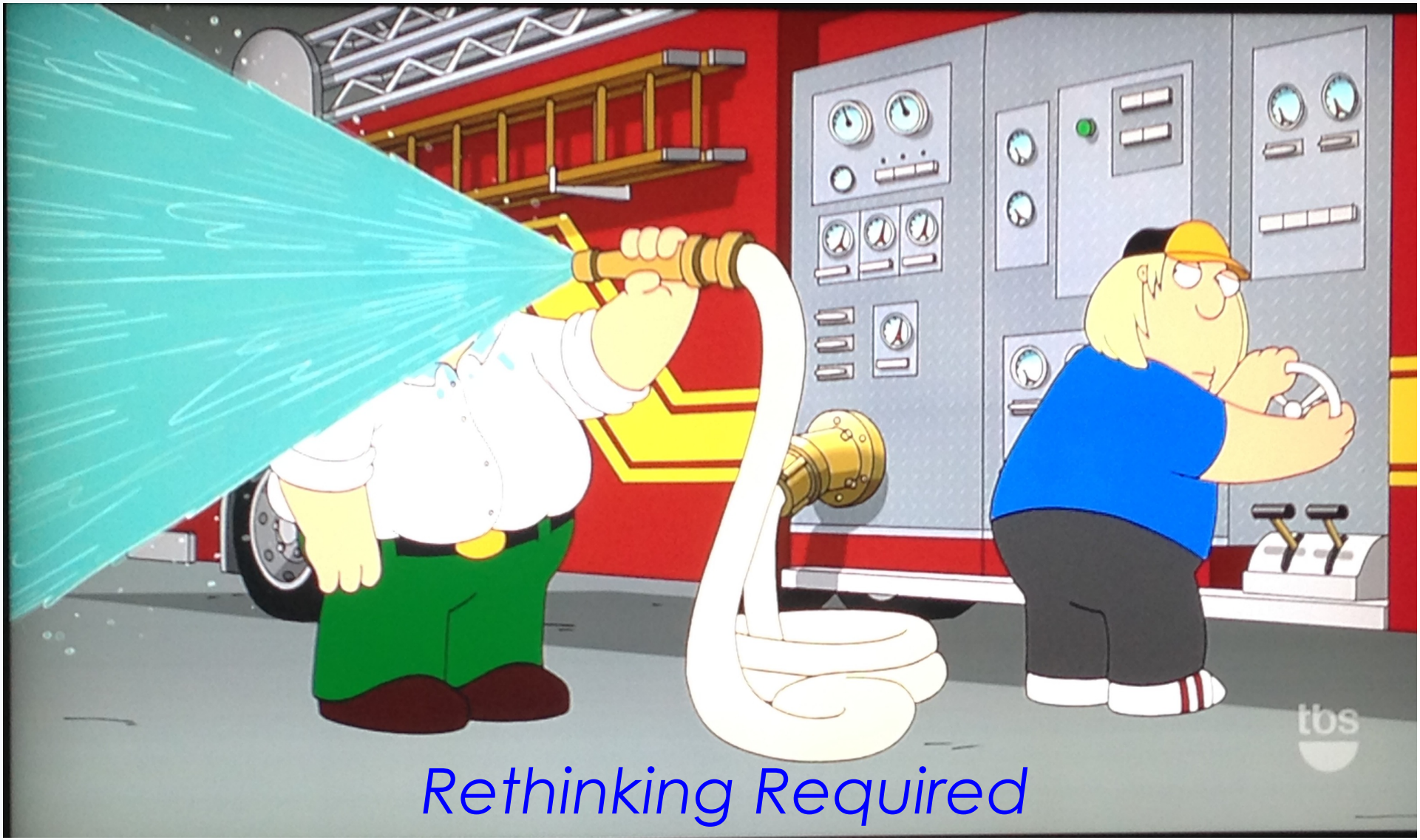
```
$ ./nonameyet ./read-prg params ./kernel-prg params
```



In-Situ Big Data Analysis at Scale for Extreme Scale Systems

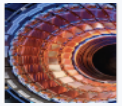
Alok Choudhary - Northwestern University

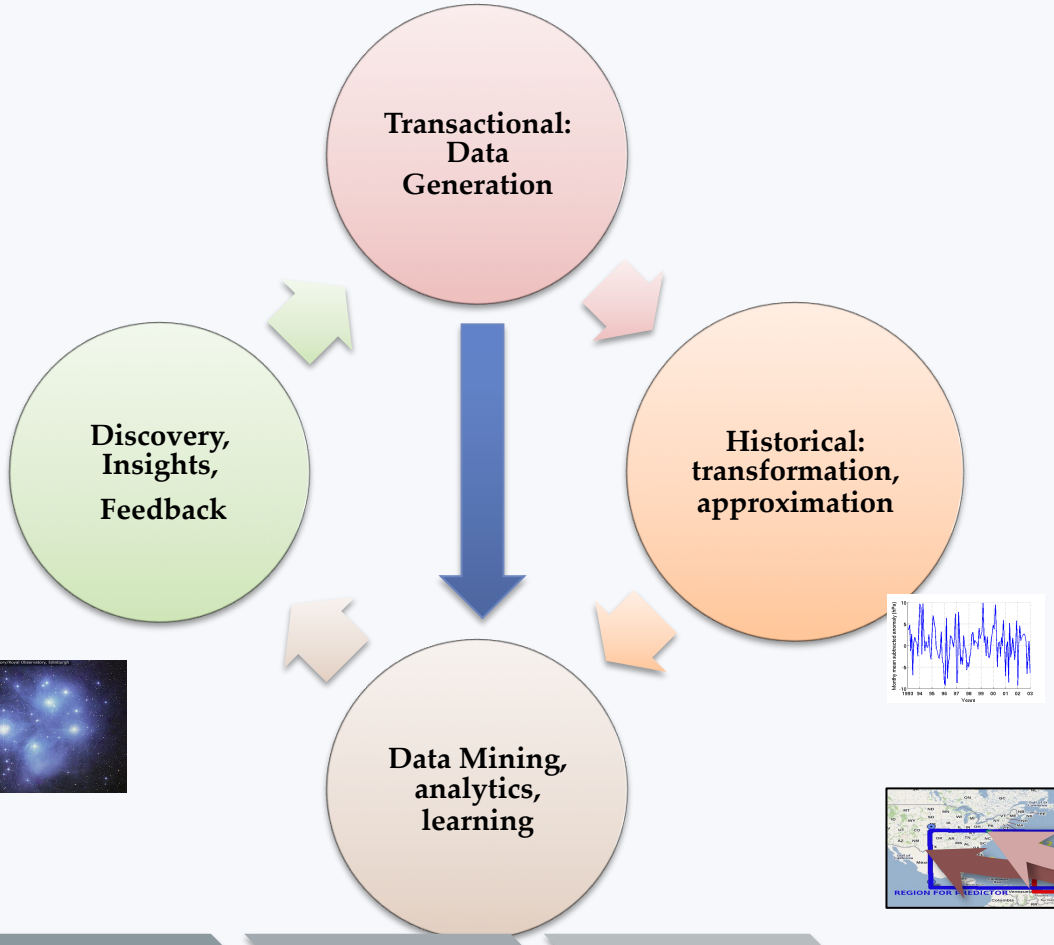
Drinking from a Fire-hose..



Rethinking Required

Strategy - Integrating Data Driven Science

Instruments, sensors   **supercomputers**



- Data Management
- Data Reduction, Query
- Visualization

Historical data → Learning Models → Trigger/ questions → Predict

In-Situ Analysis

Contributes to

In-Situ Learning

Historical Learning

Enhanced by

Transactional Learning

Operational

+ integrate Learning from historical data

+ integrate instrument data