exdci
European Extreme Data & Computing Initiative

Pathways to Convergence

Mark Asch - mark.asch@u-picardie.fr

EXDCI – BDVA meeting

Bologna, July 4th 2017

The EXDCI project has received funding from the European Unions Horizon 2020 research and innovation programmed under the grant agreement No 671558.

Outline

- Brief BDEC presentation.
- Why convergence?
- What are the pathways forward?
- How can we synergize EXDCI-BDVA?
- What are future roles in Europe?

BDEC = Big Data & Exascale Computing

- Successor to the IESP (International Exascale Software Project) that played a central role in exascale roadmaps.
- Objective: elaborate international roadmap and guidelines for achieving ***convergence*** between HPC and HDA
- BDEC holds 1-2 meetings per year since 2013 (US, Japan, EU)
- Latest meeting (#5) was hosted in Wuxi (for the 1st time!) by China…

BDEC REPORT

BDEC report: Pathways to Convergence

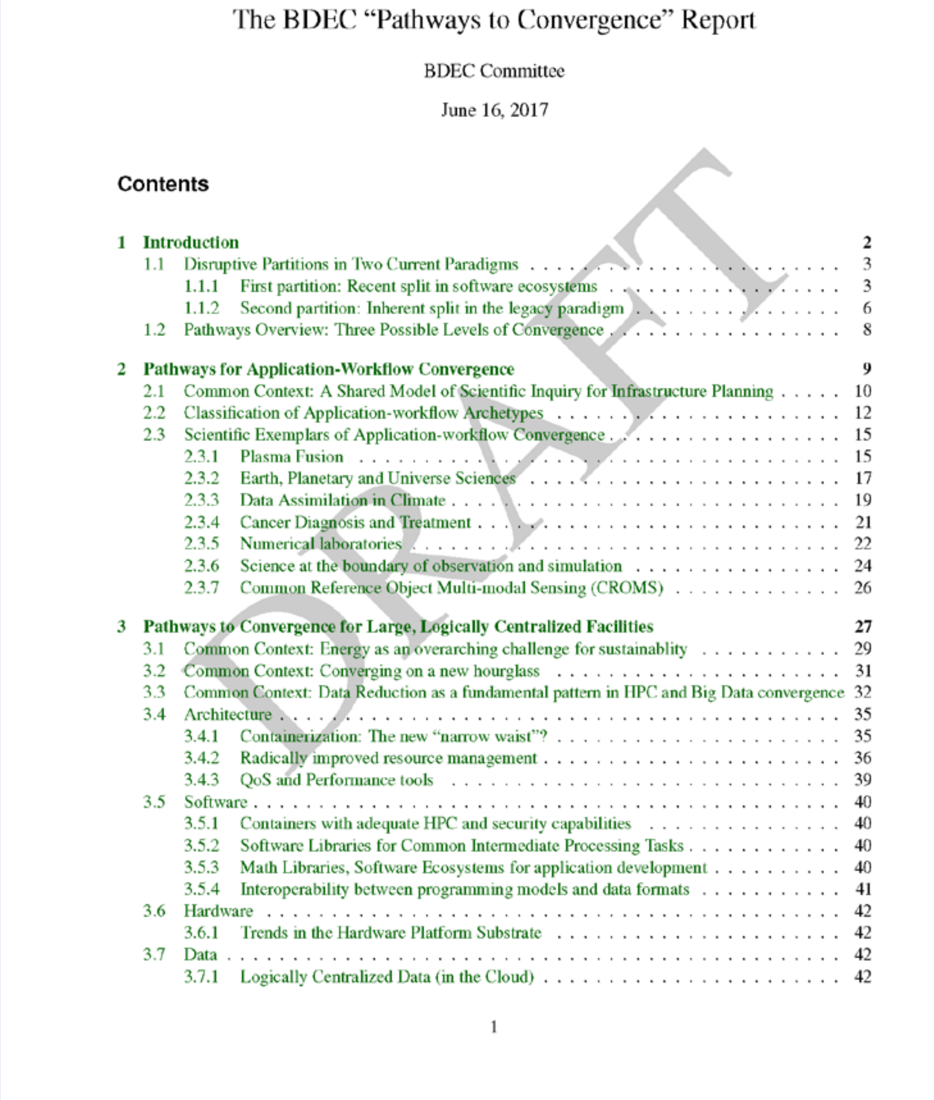
1. Split in 2 paradigms

2. App, workflow convergence

3. Centre convergence

4. Edge convergence

5. Conclusions, recommendations

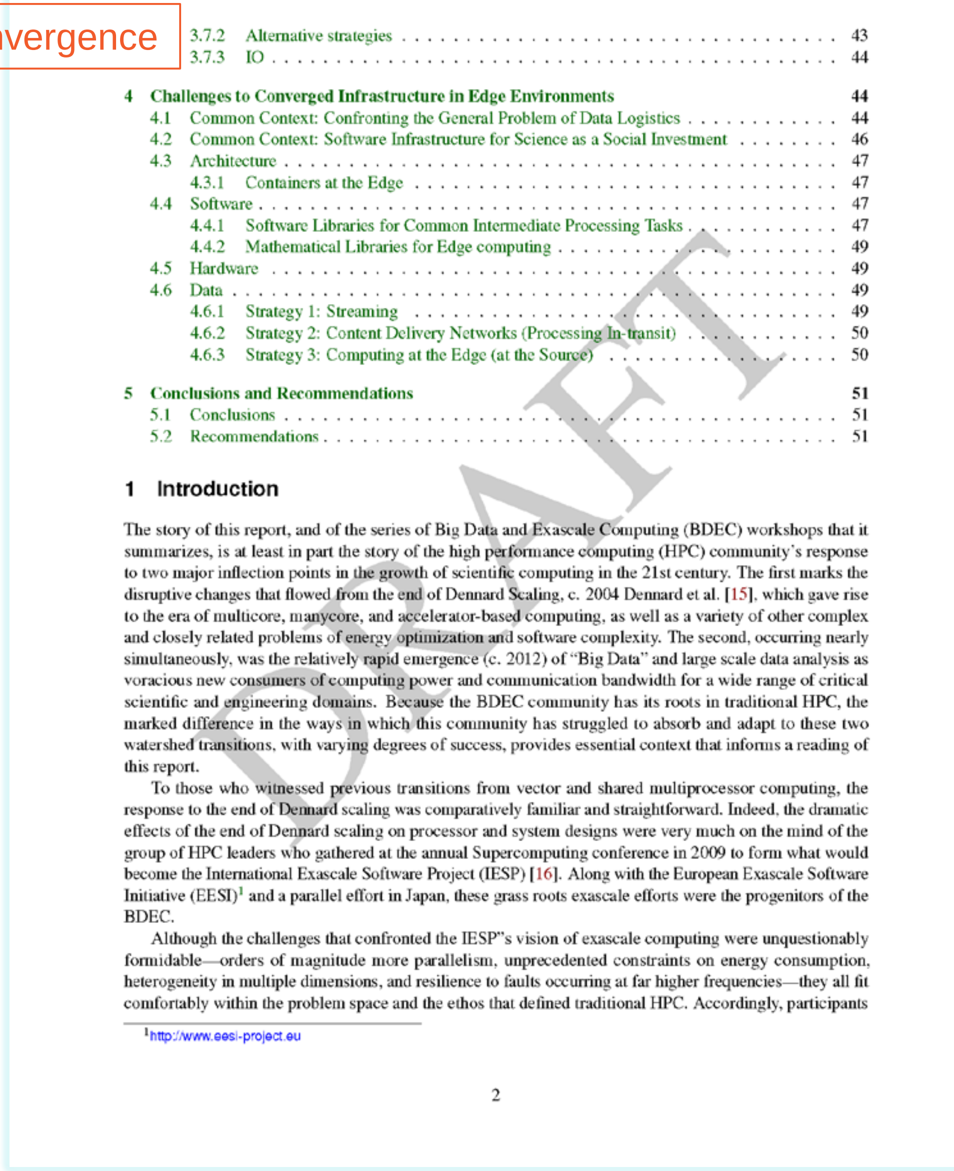


The BDEC “Pathways to Convergence” Report

BDEC Committee

June 16, 2017

Contents



1 Introduction

The story of this report, and of the series of Big Data and Exascale Computing (BDEC) workshops that it summarizes, is at least in part the story of the high performance computing (HPC) community’s response to two major inflection points in the growth of scientific computing in the 21st century. The first marks the disruptive changes that flowed from the end of Dennard Scaling, c. 2004 Dennard et al. [15], which gave rise to the era of multicore, manycore, and accelerator-based computing, as well as a variety of other complex and closely related problems of energy optimization and software complexity. The second, occurring nearly simultaneously, was the relatively rapid emergence (c. 2012) of “Big Data” and large scale data analysis as voracious new consumers of computing power and communication bandwidth for a wide range of critical scientific and engineering domains. Because the BDEC community has its roots in traditional HPC, the marked difference in the ways in which this community has struggled to absorb and adapt to these two watershed transitions, with varying degrees of success, provides essential context that informs a reading of this report.

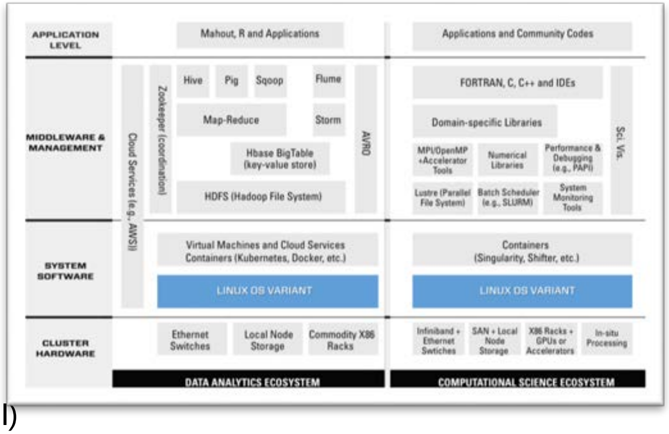
To those who witnessed previous transitions from vector and shared multiprocessor computing, the response to the end of Dennard scaling was comparatively familiar and straightforward. Indeed, the dramatic effects of the end of Dennard scaling on processor and system designs were very much on the mind of the group of HPC leaders who gathered at the annual Supercomputing conference in 2009 to form what would become the International Exascale Software Project (IESP) [16]. Along with the European Exascale Software Initiative (EESI)[1] and a parallel effort in Japan, these grass roots exascale efforts were the progenitors of the BDEC.

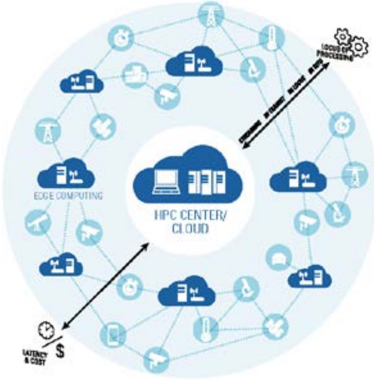
Although the challenges that confronted the IESP”s vision of exascale computing were unquestionably formidable—orders of magnitude more parallelism, unprecedented constraints on energy consumption, heterogeneity in multiple dimensions, and resilience to faults occurring at far higher frequencies—they all fit comfortably within the problem space and the ethos that defined traditional HPC. Accordingly, participants

[1] http://www.eesi-project.eu

Convergence - why and how?

- Why?
  - we are undeniably in the era of Big Data/Analytics /ML/DL
  - but there are 2 major splits:
    - the software stacks
    - the data dispersion (stateless vs. stateful)
  - as scientists, we must consider both!
  - public stakeholders will not continue to finance both!
  - we can advance by sharing experience, competences, resources between the worlds of Big Data and HPC…
  - this has been recognized by others... see HiPEAC 2017 Vision, constructors (HP, Nvidia, IBM, Intel, ...), stakeholders (EC, NSF/DOE, etc.), IDC/HYPERION

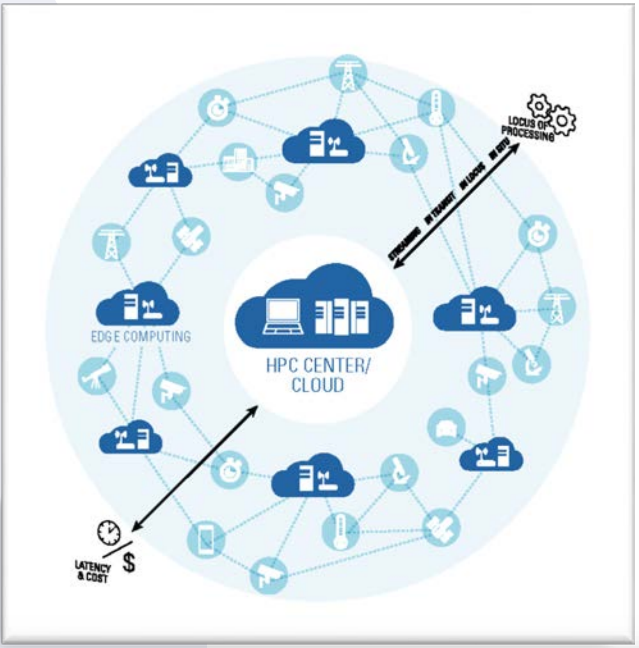

APPLICATION LEVEL
MIDDLEWARE & MANAGEMENT
SYSTEM SOFTWARE
CLUSTER HARDWARE
Mahout, R and Applications
Cloud Services (e.g., AWS)
Zookeeper (coordination)
Hive
Pig
Sqoop
Flume
Map-Reduce
Storm
Hbase BigTable (key-value store)
HDFS (Hadoop File System)
AVRO
Virtual Machines and Cloud Services Containers (Kubernetes, Docker, etc.)
LINUX OS VARIANT
Ethernet Switches
Local Node Storage
Commodity X86 Racks
DATA ANALYTICS ECOSYSTEM
Applications and Community Codes
FORTRAN, C, C++ and IDEs
Domain-specific Libraries
MPI/OpenMP +Accelerator Tools
Numerical Libraries
Performance & Debugging (e.g., PAPI)
Lustre (Parallel File System)
Batch Scheduler (e.g., SLURM)
System Monitoring Tools
Sci. Vis.
Containers (Singularity, Shifter, etc.)
LINUX OS VARIANT
Infiniband + Ethernet Switches
SAN + Local Node Storage
X86 Racks + GPUs or Accelerators
In-situ Processing
COMPUTATIONAL SCIENCE ECOSYSTEM

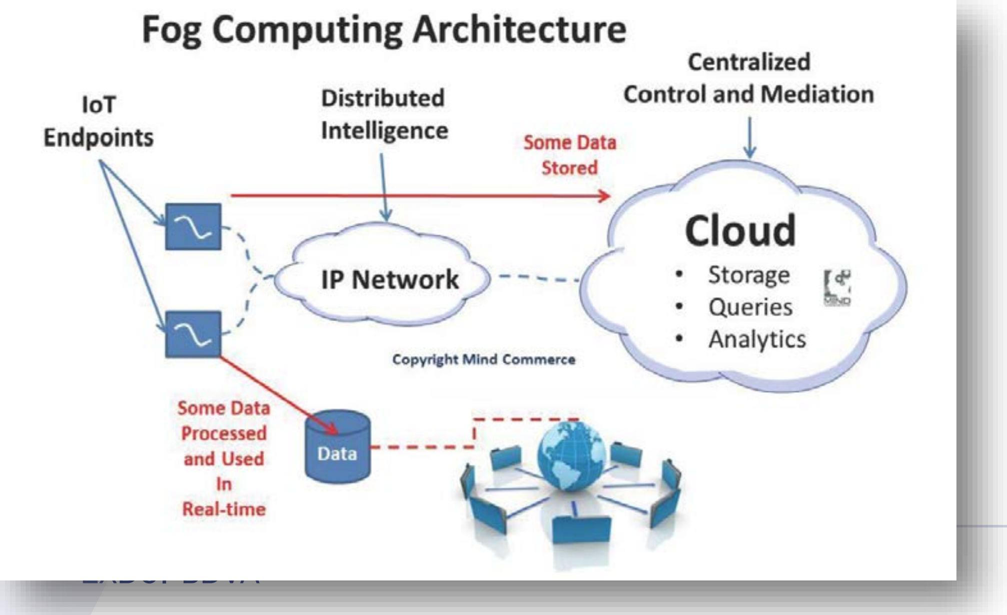


HiPEAC
COMPILATION ARCHITECTURE
HiPEAC Vision 2017
HIGH PERFORMANCE AND EMBEDDED ARCHITECTURE AND COMPILATION
Editorial board:
Marc Duranton, Koen De Bosschere, Christian Gamrat, Jonas Maebe, Harm Munk, Olivier Zendra

SPLIT #1

The 2 worlds: edge/fog and centre/cloud

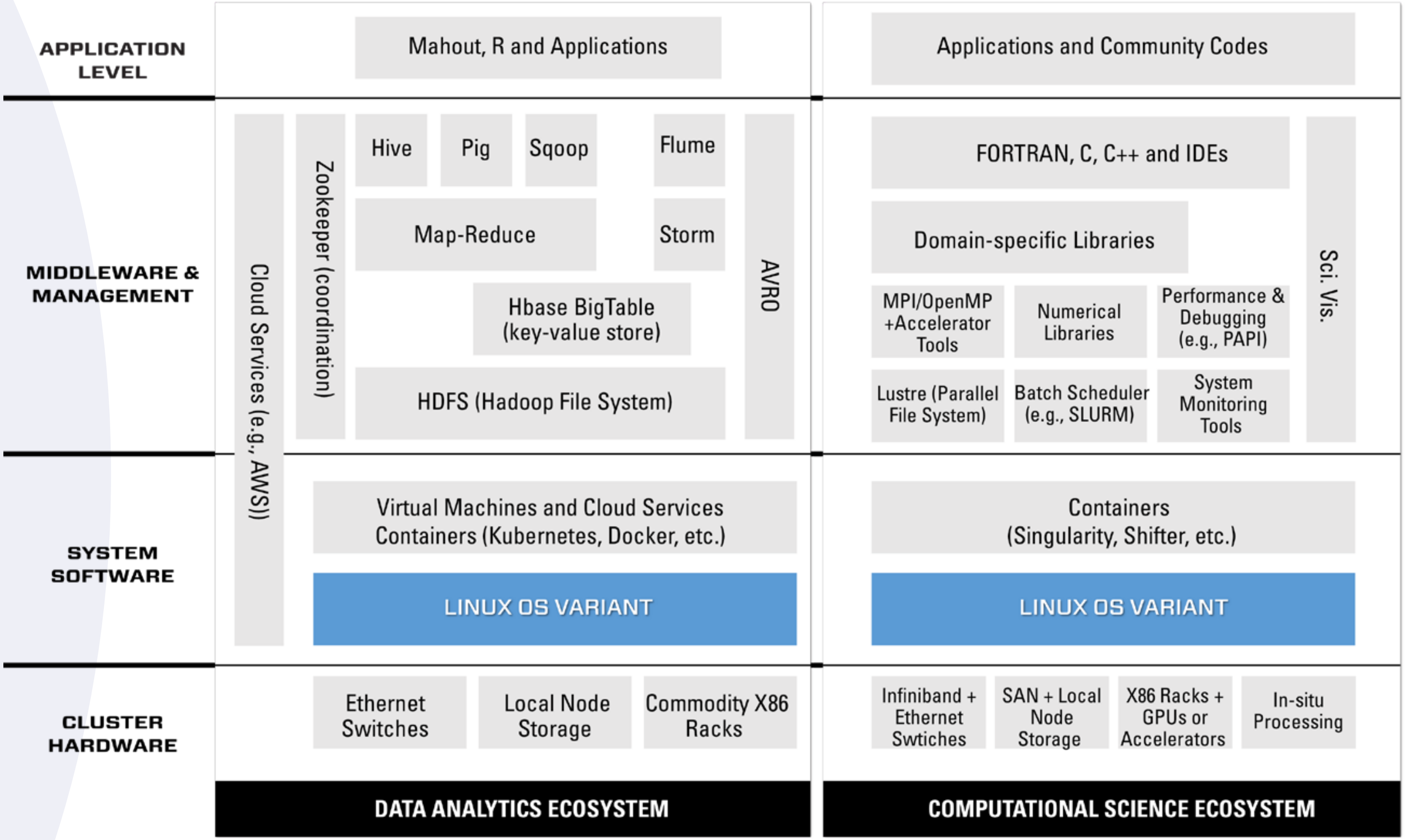




- (big) instruments and (small) sensors are measuring more and more data (IoT, Smart Cities)
- but these are out on the “edge”
- HPC and Cloud infrastructures are in the centre
- how and where are we going to do the processing?
- how can we limit data transfer/communication costs?
- what standards and protocols are needed?
- At the moment: networks only forward datagrams while all other storage and computation is performed outside the network.

SPLIT #2

The 2 worlds: Big Data and HPC software stacks



	DATA ANALYTICS ECOSYSTEM	COMPUTATIONAL SCIENCE ECOSYSTEM
APPLICATION LEVEL	Mahout, R and Applications	Applications and Community Codes
MIDDLEWARE & MANAGEMENT	Cloud Services (e.g., AWS) Zookeeper (coordination) Hive Pig Sqoop Flume Map-Reduce Storm Hbase BigTable (key-value store) HDFS (Hadoop File System) AVRO	FORTRAN, C, C++ and IDEs Domain-specific Libraries MPI/OpenMP +Accelerator Tools Numerical Libraries Performance & Debugging (e.g., PAPI) Lustre (Parallel File System) Batch Scheduler (e.g., SLURM) System Monitoring Tools Sci. Vis.
SYSTEM SOFTWARE	Virtual Machines and Cloud Services Containers (Kubernetes, Docker, etc.) LINUX OS VARIANT	Containers (Singularity, Shifter, etc.) LINUX OS VARIANT
CLUSTER HARDWARE	Ethernet Switches Local Node Storage Commodity X86 Racks	Infiniband + Ethernet Swtiches SAN + Local Node Storage X86 Racks + GPUs or Accelerators In-situ Processing

Pathways to Convergence: new realities

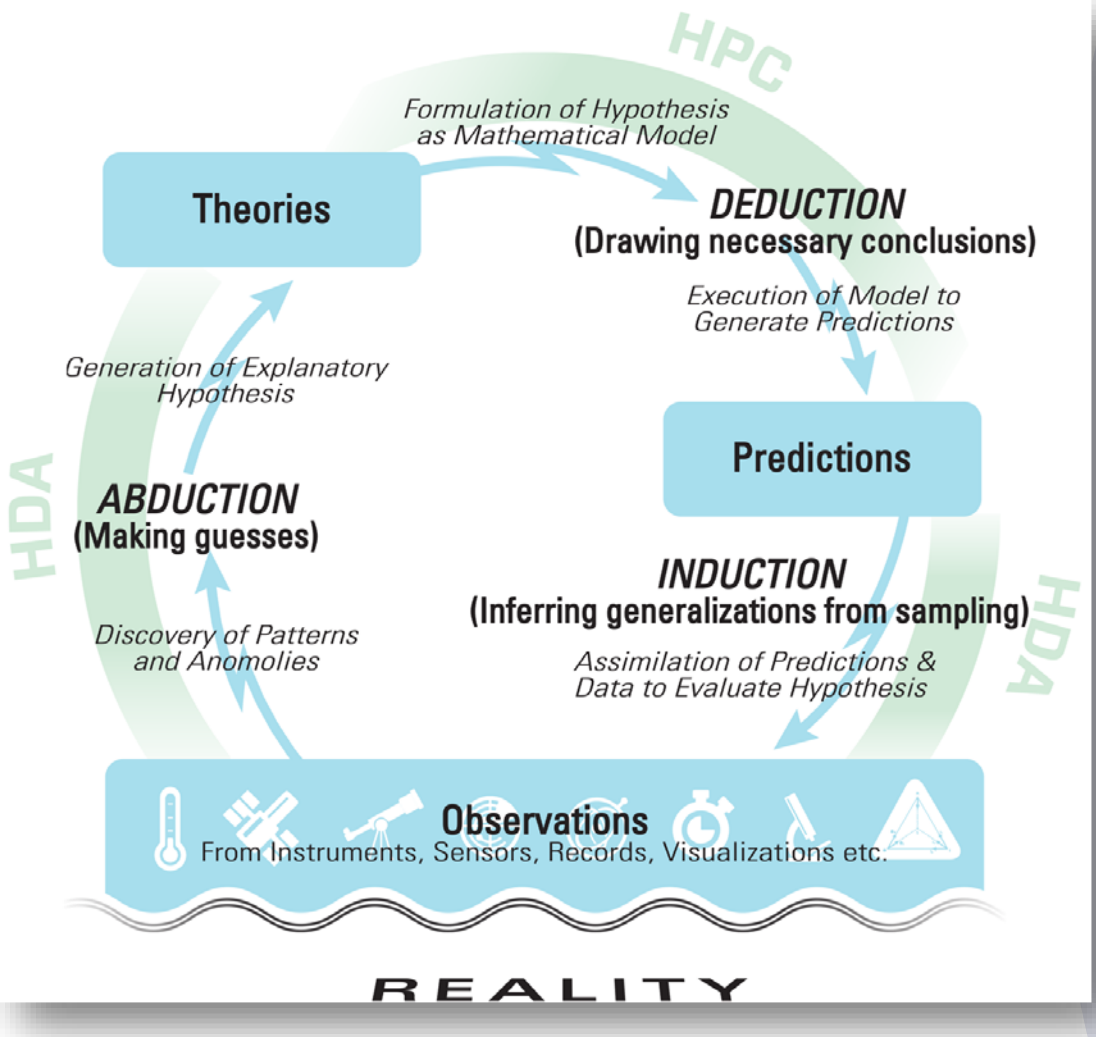
- The rising importance of AI and ML.
- Edge (IoT) is the next (very) big thing…
- CDN and other in-transit processing must be considered;
- What does this mean for future:
  - hardware,
  - software,
  - architecture,
  - and applications?

Pathways to Convergence – report structure.

- (1) Applications and Workflows are the drivers.
- (2) HPC/Cloud Infrastructures (Physically or logically centralized resources)
- (3) Edge/Cloud Infrastructures.
- (4) What does this mean for future hardware, software, architecture and applications?

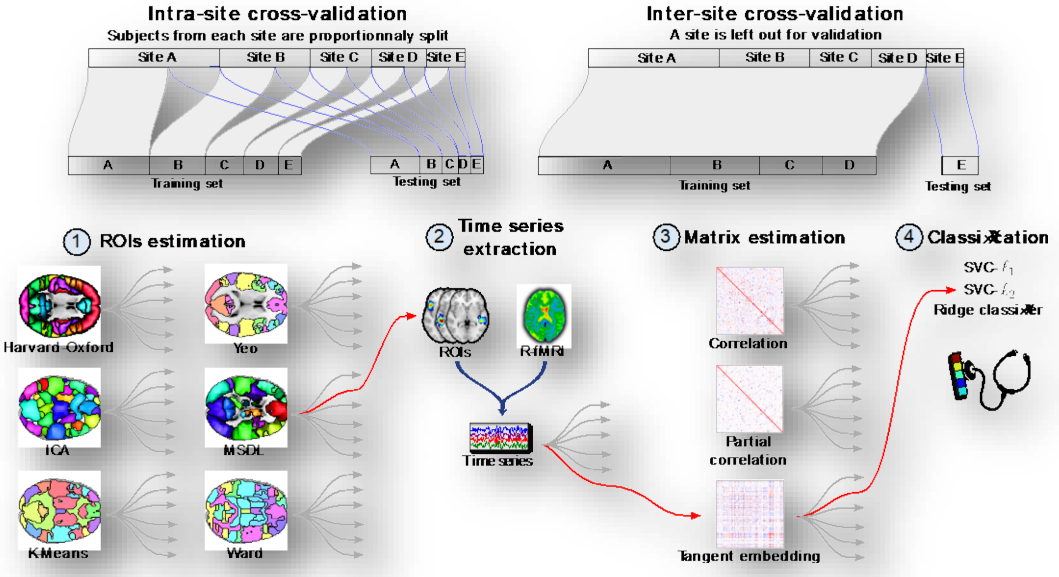
Pathways for Application-Workflow Convergence

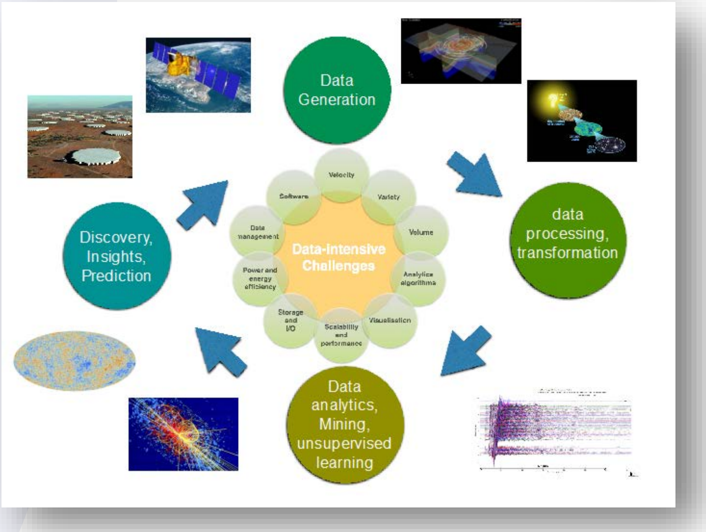
- Common context: a shared model of scientific inquiry
- Archetype classification.
- Exemplars…

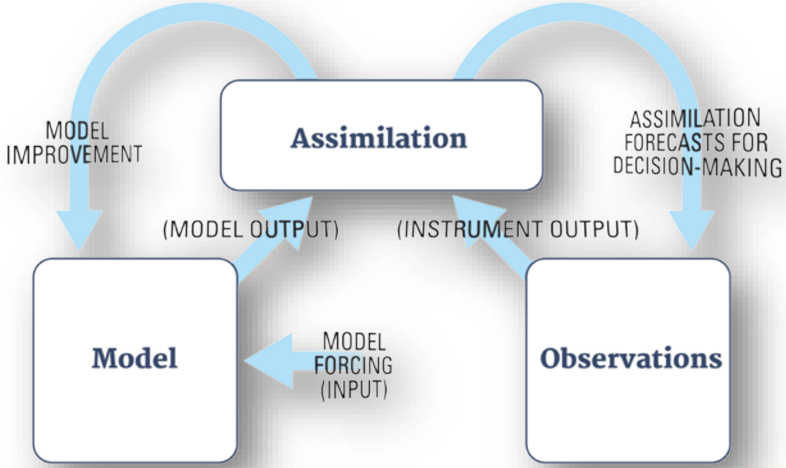

HPC
Formulation of Hypothesis as Mathematical Model
Theories
DEDUCTION
(Drawing necessary conclusions)
Execution of Model to Generate Predictions
Predictions
HDA
INDUCTION
(Inferring generalizations from sampling)
Assimilation of Predictions & Data to Evaluate Hypothesis
HDA
Generation of Explanatory Hypothesis
ABDUCTION
(Making guesses)
Discovery of Patterns and Anomolies
Observations
From Instruments, Sensors, Records, Visualizations etc.
REALITY

Exemplars for Application-Workflow Convergence

- fMRI for Autism
- DA for climate
- EPU…
- Fusion for energy




Data Generation
data processing, transformation
Data analytics, Mining, unsupervised learning
Discovery, Insights, Prediction
Data-Intensive Challenges
Velocity
Variety
Volume
Analytics algorithms
Visualisation
Scalability and performance
Storage and I/O
Power and energy efficiency
Data management
Software



MODEL IMPROVEMENT

Assimilation

ASSIMILATION FORECASTS FOR DECISION-MAKING

(MODEL OUTPUT)

(INSTRUMENT OUTPUT)

Model

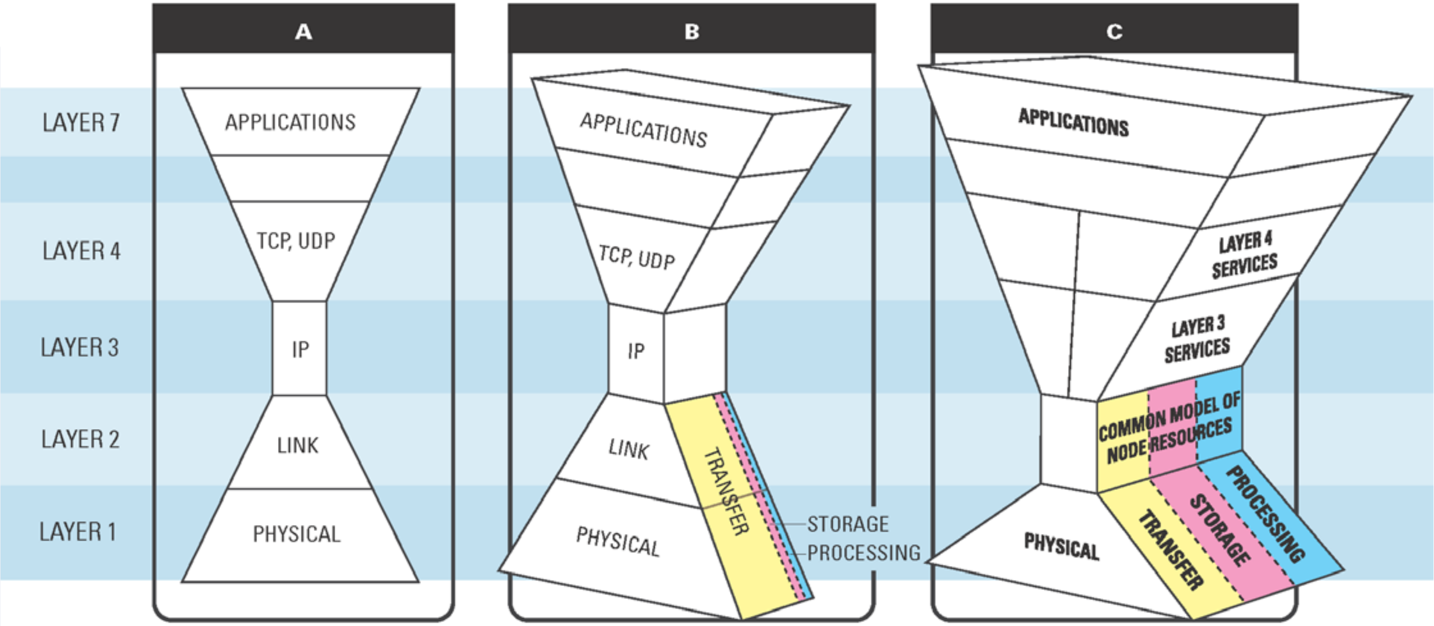
MODEL FORCING (INPUT)

Observations

Pathways to Convergence – some answers

- Answer **#1**: China and (especially) Japan are building “converged” machines with both Big Data/AI and HPC capacities--- but this does not adequately address the data split...
- Answer **#2**: Look for a new narrow waist? (see below)
- Answer **#3**: Develop and concentrate on new workflows and dataflows
- Answer **#4**: Borrow ideas from BD world and implement them in HPC
  - New file systems.
  - New data and communication procedures, resources and protocols.
- Answer **#5**: Develop new software for data reduction, data analytics, dataflows, Function as a Service (FaaS)

Pathways to Convergence – narrow waist

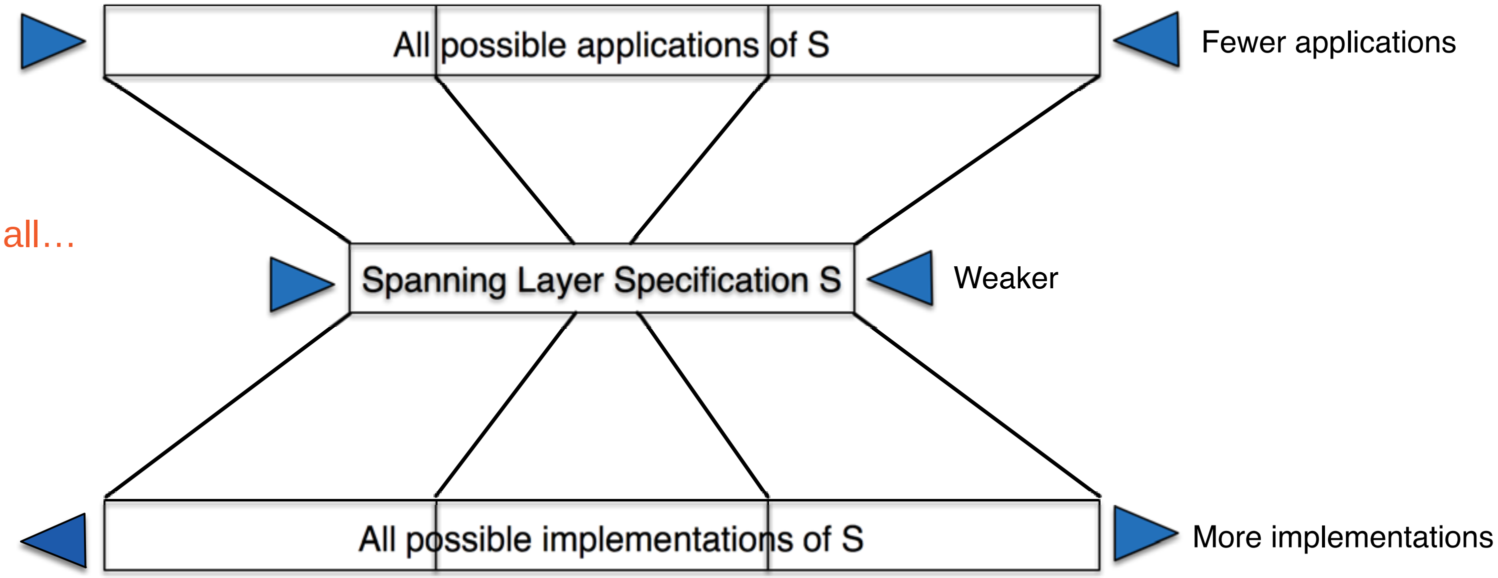


Classical Internet Hourglass

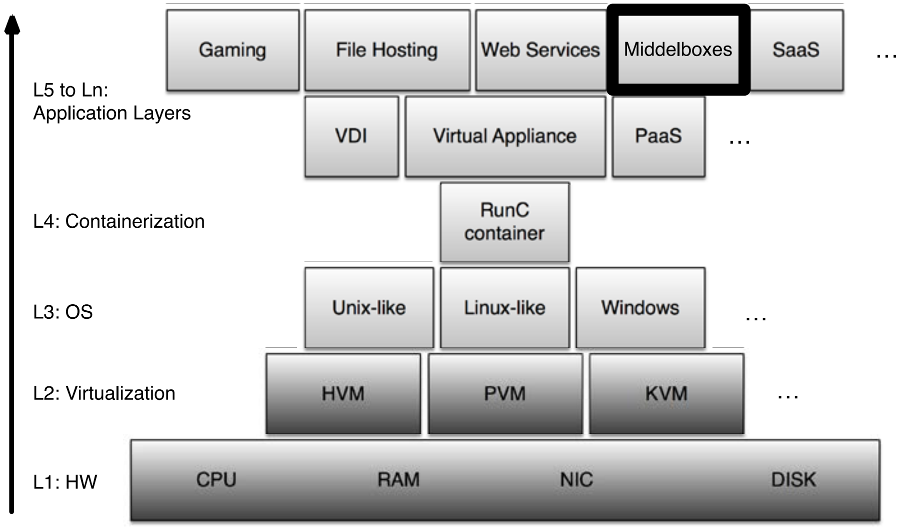
- waist must be minimal, weak, restricted

More general Hourglass that exposes all resources at lowest possible level

Pathways to Convergence – narrow waist


All possible applications of S
Fewer applications
Spanning Layer Specification S
Weaker
All possible implementations of S
More implementations

You can't win 'em all…



Containers as narrow waist?

containers have emerged as a viable delivery platform for HPC software, and are already used widely in HDA…

General problem of Data Logistics

- *How to manage the time-sensitive positioning and encoding/layout of data relative to its intended users and the resources they can use?*

- 3 categories that "follow the data": (from an Applications viewpoint)
  - (1) data arriving from the edge (often in real time), never centralized;
  - (2) federated multi-source archived data;
  - (3) combinations of data stored from observational archives with a dynamic simulation.

Data Logistics solutions/approaches/strategies

- Data is physically distributed, but its control is logically centralized – eg. Google based on large compact data centres (50,000 servers each with 64 cores) but data is physically distributed with software defined networks enabling over 0.5 petabits per second bandwidth into each data centre - advanced architectures with NVM having 10 µsec latency replacing disks with 10ms latency. Google Spanner gives the illusion of a consistent database on top of this infrastructure.
  - This commercial architecture ensures that data and computing are logically co-located (no need to "bring the compute to the data")
- Streaming: multiple sources and sinks of data, not at the same location (eg. SKA); data may need to move from location-to-location in the system, and there may need to be processing that occurs at locations as the data moves through the system; data reduction is often necessary at the intermediate stages; Apache Spark and Flink provide libraries and runtimes in Cloud context

Data Logistics solutions/approaches/strategies

- The use of data streaming has three benefits: a) the low overheads, particularly for local coupling, means that it is efficient for scientists to compose low-cost steps, called ``fine-grained workflow'', b) the direct handling of streams means that scientists can develop methods for live, continuously flowing observations, and c) the load placed on the increasingly limiting bottleneck of disk I/O is minimised.
- Content Delivery Networks (Processing In-transit): an answer to the relentless increase in demand for rich multimedia content; but expensive and difficult to operate, and are practical only for implementing Web and media streaming sites that generate enough income to pay for their service; out of reach of (most) scientific communities (ESGF has a peer-to-peer network…); CDNs use a proprietary resolution protocol that provides the network interface of an uncacheable DNS server; need: scalable implementation, organizational will of scientific communities.

Other issues (that are addressed in the report)

- Software and mathematical libraries.
- Energy and sustainability.
- Improved resource management at system level.
- Interoperability between program models and data formats.
- Scientific exemplars exhibiting different types of data use.
- Science at the boundary of observation and simulation.
- Numerical laboratories.

Where do we stand in Europe?

- In the (pure) exascale race, we're 2-3 years behind… (IMHO)
- We need to concentrate on domains where we have (consensual) strengths and move forward on these:
  - federated platforms for convergence,
  - algorithm and software development for convergence,
  - analytical methods and tools for convergence,
  - applications using convergence.
- The ETP4HPC-SRA coupled with BDVA is the best vehicule for this.
- EXDCI is playing a central role in coordinating this, and the BDEC initiatives..

Thank you!

- Please consult www.exascale.org for reports and presentations from all the BDEC meetings, where you can download an early draft of the document “Pathways to Convergence”.
- “Pathways to Convergence” report will be officially presented at Supercomputing 2017 in Denver USA.
- You can contact us at:
  - mark.asch@u-picardie.fr
  - tmoore@icl.utk.edu
- **Questions and/or comments?**