

Scientific Knowledge Discovery Computational Research Challenges for HEP.

March 18th

Ruth Pordes

For the past several decades particle physics has demonstrated the values of collaborative scientific research sharing a common computing environment. The implementations remain, effort intensive and have been affordable only because of the size of the collaborations and the global funding contributions.

The field has demonstrated decreased time from data taking to scientific results, especially with the LHC Higgs reports and rate of physics publications from the LHC collaborations, through development of and commitment to scientific knowledge discovery by small to large analysis groups using globally distributed computing¹. This has been accomplished through large scale a) sharing of data b) sharing of distributed computing c) group software codes and analyses d) team approaches to developing and extracting science from a single facility e) discipline and development and operational efforts to maintain a rigid process of production software releases f) well organized socially based management of analysis teams g) coordinated management and control of the validation and release of results and h) detailed processes, policies and allocation of conference presentations and authorship and publication of scientific papers.

This next generation of particle physics research continues the tradition of collaborative (team) science with collaboration sizes from a few tens of scientists on some of the Intensity Frontier experiments to the more than a thousand on each of ATLAS and CMS (through upgrades over more than the next decade). The field is facing increasing constraints in budgets for personnel (especially engineers and professionals) – which makes the allocation of the more effort intensive activities increasingly less possible.

The following plans for scientific discovery are generating significant computing needs as well as human resource challenges:

- a) The LHC accelerator and detector upgrades will deliver 10x more data (100 Petabytes/year stored from the detectors and an Exabyte of managed and transferred data by 2017) and require more than 10x more computation than currently deployed (due to changes in beam topologies that make the overlays of different detector events more complex to read-out and detangle).

¹ As part of the 2012 press release related to the potential HIGGS particle discovery, the CERN director general gave credit to the huge contribution that the world wide computing grid made to the discovery.

- b) Accelerators in support of neutrino experiments which need increasingly sophisticated and timely real time processing, feedback and control of accelerator components to minimize beam loss, increase the quality of the physics output, and generate worthwhile returns on investments. This requires a large amount of simulation of the detectors – including the generation of events and investigation of the response of the detectors and accelerator elements.
- c) Next generation (2016-2020) dark matter and neutrino experiments that need high throughput data filtering pipelines to manage and synthesize up to 100GByte/sec front end data collection rates, to move to GPUs, multi-core processors and/or easily modified software in pace to allow ongoing revisions of the codes in response to real-time conditions.
- d) Many-pass exploration, increased multi-hypothesis analyses, multi-disciplinary interpretation and longer-lived comparative studies of common (private and public) datasets in the 1-100 Petabyte range. (e.g. astrophysics experiments, LHC upgrades – currently most popular datasets are in the several 100 TB range)
- e) Simulation and modeling of potential new accelerators and their components, requiring many-dimensional system and many-thousand component models, management, and methods (e.g. such as for the Project-X at Fermilab, next generation linear collider.)
- f) Particle-astrophysics collaborations collecting Petabytes of data and Terabytes of supporting meta-data and provenance, which must be efficiently traversed and reused as part of single or multiple analyses.
- g) Increasing scientific drivers towards deployment of remotely managed, accessed and controlled unique facilities, sometimes in harsh environments (e.g. down mines, inclement climates such as deep underground or at the South Pole)

The output of the current “Snowmass Community Summer Study 2013²” preparations will detail the specific requirements, challenges and priorities both in the physics roadmap and the computational programs to support and enable it.

Needs already identified are bringing the following principles as drivers to the goals of computing for the future of the field:

- a) Ability to access and use dynamically changing and regularly changing set of computing resources.
- b) A significant redesign, recoding of million line offline processing, analysis, simulation codes, frameworks and algorithms to use many-core, specialized processors (including powerful graphics processors), and position the field to

² <http://www.snowmass2013.org/tiki-index.php>

take much faster and more effective advantage of new processor hardware architectures and implementations.

- c) Adoption of just in time, real time global data access based on the availability of at least some 100 Gbit network connections, rather than pre-placement of data.
- d) Increased and increasingly complex software analysis of all data acquired from the detector – rather than less flexible hardware triggering systems - to allow real time adaptation and feedback of the algorithms based on the output results.
- e) Planning for increases in sharing and reuse of both data and software.
- f) Increased priority for previously and newly acquired data to be available for long-term reuse and reanalysis, and to be accessible for cross-experiment and cross-discipline correlation and combination
- g) Increased attention for comprehensive set of supporting information and meta-data that describes the complete provenance of the processing activities and algorithms used to obtain results, for validation, combination and re-analysis.
- h) Integration and support for “untethered” and mobile computing devices as part of the end-researcher toolset.
- i) Increased caching of large-scale data in-flight to allow inline, real-time software analysis, filtering and feedback as part of data acquisition, storage and categorization pipelines.
- j) Need to protect very large investments by providing headroom that allows evolution, innovation and serendipity in the scientific goals and methods supported.

The overarching Grand Challenge is to model, design, implement and sustain a system of many stages and complex components providing a seamless pipeline from 100GHz digital data generators through the application of adaptive scientific processes to team based knowledge discovery, wisdom and innovation; in an environment that supports evolution, multi-faceted trust, comprehensible and re-usable multi-tiered heterogeneous computing and software artifacts.

This system must provide for sustained extreme-scale data acquisition, complex in flight analysis of many terabyte size data streams; provisioning of high throughput data transport, data description and data storage services for 100s Petabytes of data; enabling of remote distributed collaborating researcher teams abilities for active and timely access to any and all data to extract knowledge over decade long explorations; and provide robust, trustable, understood and reusable infrastructure and scientific software components and toolkits;

Summary of Areas of Beneficial Research	Support
Advanced models and methods for discovering, communicating, reserving, allocating and using a set of many-types-of resources that make up the distributed computing system	a) c) h)
Frameworks, templates, and libraries that support significant improvements in end-to-end error reporting, handling and diagnosis, which enable more robust and manageable many-	b) c) e)

layer computational systems.	
New methods, models and frameworks for human-computer and human-software interfaces that result in increased ability to extract knowledge (and generate wisdom) and lower the human cost of data and information management and evolution.	b) d) e) f) g)
Rigorously defined but easy for the user to understand, apply, evolve and validate authentication, authorization (trust) and protection models and implementations across many, distributed, dynamic heterogeneous information repositories and resources.	e) f) g)
Advances in real-time and continuous transport, management, caching, storage and access of multi-tier systems of networks and large scale distributed datasets.	c) i)
Parametric models and semantics for hybrid (simulated/modeled and implemented/measured) indeterministic, distributed systems that can predict, optimize and explain the behavior of complex dataflows and workflows.	a) c) j)