Summary of Blueprint Workshop:

Analysis Systems R&D on Scalable Platforms

June 21–22, 2019
New York University
Meeting URL: https://indico.cern.ch/event/820946/

Workshop Organizers:

Kyle Cranmer (New York University)
Rob Gardner (University of Chicago)
Mark Neubauer (University of Illinois at Urbana-Champaign)

Summary prepared by:

Mark Neubauer (University of Illinois at Urbana-Champaign)

Major Goals

- Review the status of the Analysis Systems (AS) milestones and deliverables.
- Develop the Scalable Systems Laboratory (SSL) scope, architecture and plans, using AS R&D activties as concrete examples.
- Develop requirements on SSL to support the AS area, particularly the prototyping, benchmarking and scaling of AS deliverables toward deployment.
- Increase the visibility of SSL and AS R&D beyond IRIS-HEP to facilitate partnerships with organizations that could potentially provide software and computing resources for SSL.
- Get informed on latest developments in technologies and methods relevant for SSL and AS.

Key Outcomes

- Communication of the AS area plans and preliminary requirements to SSL.
- Kubernetes as a planned *common denominator* for SSL, increasing the capabilities through flexible infrastruture. This idea spawned plans for a multi-site SSL *substrate* project that will federate SSL contributions from multiple resource providers, offerring the AS area a flexible platform for service deployment at scales needed to test viability of system designs.
- Productive engagement of the AS/SSL team with representatives from NCSA, SDSC, NYU Research Computing, industry & cloud providers (Google, Redhat), generating action items.
- A concrete vision for an SSL that serves not as an innovation space for AS developers, but as a testbed to prototype next generation infrastructure for future HEP computing environments.

About the Blueprint Activity: Designed to inform the development and evolution of the Institute's strategic vision. At its core, a series of workshops that bring together IRIS-HEP team members, key stakeholders and domain experts from disciplines of importance to the Institute's mission.

I Overview

Together with the OSG-LHC, the Scalable Systems Laboratory (SSL) is designed to be the primary integration path to deliver the output of IRIS-HEP R&D activities into the distributed and scientific production infrastructure of the experiments. The aim of this workshop is to further develop the IRIS-HEP SSL concept using specific R&D examples from the AS area, including low-latency, query-based data systems and modular, reusable cyberinfrastructure for physics inference and results dissemination. Registered attendees include those from IRIS-HEP (primarily SSL and AS areas), US ATLAS/CMS operations programs, national labs, CERN, supercomputing centers (SDSC, NCSA), university research IT, and industry (RedHat, Google).

The venue for the workshop was the Physics Department at New York University and was hosted by Kyle Cranmer. The blueprint meeting benefited by its proximity to the IRIS-HEP Analysis Systems Topic Workshop which immediately proceeded it.

II Attendees

There were 26 registered participants for workshop, with all but a few attending in person. The workshop attendees were: Andrew Chien (Chicago), Andrew Melo (Vanderbilt), Aravindh Puthiyaparambil (Red Hat), Benjamin Galewsky (Illinois/NCSA), Dan S. Katz (Illinois/NCSA), David Ackerman (NYU), Edgar Fajardo (SDSC), Eric Borenstein (NYU), Gordon Watts (Washington), Ianna Osborne (Fermilab), Jim Pivarski (Princeton), Kyle Cranmer (NYU), Lincoln Bryant (Chicago), Lindsey Gray (Fermilab), Mark Neubauer (Illinois), Mason Proffitt (Washington), Matthew Feickert (SMU), Nils Krumnack (Iowa State), Ricardo Brito Da Rocha (CERN), Rob Gardner (Chicago), Sanjay Arora (Red Hat), Stephen Fang (Google), Stratos Efstathiadis (NYU), Tatiana Polunina (NYU), Tim Boerner (Illinois/NCSA), Wei Yang (SLAC)

III Goals

The primary goals of the workshop were to

- Review the status of AS milestones and deliverables
- Develop the Scalable Systems Laboratory (SSL) scope, architecture and plans, using Analysis Systems (AS) R&D activties as concrete examples.
- Develop requirements on SSL to support the AS area, particularly the prototyping, benchmarking and scaling of AS deliverables toward deployment.
- Increase the visibility of SSL and AS R&D beyond IRIS-HEP to facilitate partnerships with organizations that could potentially provide software and computing resources for SSL.
- Get informed on latest developments in technologies and methods relevant for SSL and AS.

Significant progress was made towards each of these goals.

IV Activites

The meeting was structured as a series of informal presentations which sufficient time for in-depth discussions. In advance of the meeting we identified some specific objectives to guide discussions:

- Collection and curatation of analysis use cases, each with a reference implementation. What patterns and infrastructure are needed?
- Translation of analysis examples into new specifications, providing feedback and iteration.

- Development of initial specifications for user-facing interfaces to analysis system components.
- Benchmarking of existing analysis components and integrating the benchmarking into SSL.
- Development of accelerator-based fitting & statistical tools (and other relevant components).
- Integrating prototypes of AS components into SSL, followed by benchmarking & assessment.

IV.1 Presentations and Discussion

There were eight presentations chosen to trigger discussion on the main themes and objectives of the workshop.

IV.1.1 The Blueprint Process

Mark Neubauer (University of Illinois) presented *IRIS-HEP Blueprint Concepts and Process* which provided a broad overview of not only the blueprint process but the data and processing challenge to set the context. This was particularly helpful for workshop participants new to computing in HEP and the HL-LHC computing scale.

IV.1.2 The SSL Concept

Robert Gardner (UChicago) presented *Scalable Systems Laboratory (IRIS-HEP) - challenges, opportunities*. The program of work of the SSL activity was described. Given the SSL core has less than an FTE of effort, success will depend on leverage efforts from other areas of IRIS-HEP and engaged partners. The high level purpose of the SSL is to provide the Institute and the HL-LHC experiments with scalable platforms needed for development in context, i.e. the *path to production*:

- Provides access to infrastructure and environments
- Organizes software and resources for scalability testing
- Does foundational systems R&D on accelerated services
- Provides the integration path to the OSG-LHC production infrastructure

The challenges are that it must be a community platform across experiments and institutions; it should support groups/projects with specific organizational membership for access; it must aggregate bespoke resources & configurations; it must do so in *delcarative* fashion such that deployments are reproducible and *mobile*; it must provide services to build & management deployment artifacts; and it must be scalable up and down.

The opportunities presented are that in building out the SSL adhoc collaborations will be formed which cross organizational boundaries; contributions will come from diverse resource providers, broadening participation; new models of infrastructure development will be formed supporting more rapid innocation of new analysis systems; and that artifacts can be redeployed generally will accelerate delivery of R&D systems to the community for use in production.

It was noted that storage and networking become a big issue for the large data sets we're dealing with, and so how does the SSL reach relevant scale for feasibility testing? The possibility of augmenting with cloud resources was discussed (an example of a several million cores provided to an MIT researcher was noted), and comparison costing with on-prem resources. It was agreed this should always be considered and costs periodically assessed. Additionally research partnerships where there is mutual interest with public cloud providers should continue to be exploited. Finally, specifics of the server/instance targets need to be considered, such as memory/core and I/O bandwidth which could pose special challenges and costs.

IV.1.3 Analysis Systems

Kyle Cranmer (NYU) presented *Analysis Systems Perspectives & Goals*. The just finished Analysis Systems Topic Workshop revisited the milestones and deliverables for the rest of IRIS-HEP Year 1 and and Year 2 planning. It was noted there are AS scalbility milestones in Y2Q1 which imply requirements on SSL readiness, The goal is to move testing on whatever resources currently in use to an *SSL-managed* infrastructure. This includes prototypes of analysis systems components to be deployed on the SSL (Aug '20).

There were notable adhoc demos of interest to the workshop. First was the REANA/RECAST demo at KubeCon 2018, focusing on real analysis reproducibility, demonstrated that HEP can engage with modern open source tools and communities. Second was the scalability demo of Higgs rediscovery on Kubernetes (200 GB/s, 70 TB) performed at KubeCon and CloudNativeCon Europe 2019. There are other examples in our field, e.g. CERN's Next Generation Data Analysis Platform with Apache Spark by Enric Tejedor (CERN) at the Spark+AI Summit Europe in London, October 2018.

A number of systems have or will soon emerge from AS that are candidates for deployment and testing on the SSL.

- ServiceX: part of DOMA's efforts to develop intelligent data delivery services (IDDS), the service focuses on reformatting data a the end-stage analysis phase, transforming event data into columnar formats which provide advantages for effciency and Python-based processing frameworks.
- Coffea: a columnar analysis framework being developed at Fermilab. It was noted that soon there are ServiceX + Coffea demonstrators
- MADMINER: containerized workflows with mix of CPU and GPU/TPU (integration of simulation, machine learning, and statistical inference); These workflows are ready for execution on Kubernetes using REANA.
- AMPGEN, pyhf (fitting as a service): have resources setup and available for users to upload information (e.g., pyhf JSON) and then the service performs the fit. Saves the user from being required to set things up on their own (services are simple, but not everyone has a nice GPU cluster ready to go).

The types of systems the AS team has considered for development and testing were discussed. These incluced public cloud (speed of startup, additional services), university resources (on-prem costs, data storage), existing grid infrastructure (e.g. the Open Science Grid) which has a dedicated integration team (OSG-LHC) in IRIS-HEP, DOE and NSF leadership class HPC systems. The importance of having the ability to move service deployments and workloads between these resource categories was noted. The role of container usage on the grid was discussed, including early applications in distributed training (hyperparameter tuning). Much existing work can be leveraged here, with previous efforts reported at ACAT 2019, and talks on machine learning in ATLAS using Docker images. The possiblity of providing HPC "backends" to REANA, including HPC, was discussed and considered a worthy goal. An interesting side topic was the emerging market place of resources for machine learning outside the public cloud providers and HPC centers; in particular vast.ai provides a cloud computing, matchmaking and aggregation service focused on lowering the price of compute-intensive workloads.

IV.1.4 SSL Architectural Principles

Lincoln Bryant (UChicago) presented *SSL patterns: hybrid models, developer support, deployments*. There are a number of desirable features that have been identified for the SSL. These include community access - open to all working on software infrastructure in HEP - which can be implemented with federation tools based on CI-Logon, for example, providing a single sign-on capability; a lightweight group (project) management system; infrastructure itself should be *composable and reusable*; being able to accommodate/aggregate a diverse resource pool and user community; a container-based service orchestration on dedicated resources; VC3-like technology to connect to HPC/HTC resources for batch scale-out; facilitate integration of commercial cloud resources when needed;

Regarding declarative & reproducible deployments, the goal is to have infrastructure built under the SSL to be easily reusable and deployable to other sites. The declarative nature of Kubernetes is a good fit and gets us a long way down that road.

SSL itself should not become a production center; rather it should serve as an incubator for projects which then *graduate* to become full-fledged infrastructures that run on production resources. Services to build & manage artifacts, tools that provide SSL to be scaled up and then back down are part of reducing cognitive load for developers and deployers.

The SSL team is currently using Google Cloud Platform (its Kubernetes Engine) to test ServiceX deployment; this will soon be pulled that off into in-house resources.

From the university point of view, groups which would like to reproduce AS systems locally should have resources to have simple versions of what they need to provide, e.g., a base Kubernetes cluster, a functional REANA instance, as a start. From campus IT/research technology point of view, easy to deploy systems offer the ability to make a compelling case to the Dean/Provost that they are providing resources that enable good science.

Suggested were some light-weight mechanisms for discovery of resources. The value of reporting and showing science that is happening on the contributed resources to incentivize resource providers at the universities was noted. Capturing success stories, so university community understands how they can benefit from investments to shared campus resources, including staff, were noted. Having a dashboard may help communicate the contributions. This would be important for products that can be used outside HEP, giving them higher visibility.

There were questions as to whether the SSL provisioning method, dashboard panels, and other tools developed to materialize and manage the service infrasructure would be open sourced and productized? These were interesting possibilities and would depend on the level of effort and other priorities.

IV.1.5 Experience with Google Cloud Platform

Lukas Heinrich (CERN) presented *Ecosystems I: Google Cloud Platform*. The KubeCon 2019 keynote Reperforming a Nobel Prize Discovery on Kubernetes was illustrative of the power and flexibility of Kubernetes and its relevance the HEP computing. The CMS open data sample (70 TB, 25000 files) was reprocess on stage using legacy software from the CMS scientific software stack using Kubernetes at a large scale. The main lessons learned were that the Google Network can serve extreme data rates into compute nodes (2 Gbps/core) once handled appropriately. Incoming data could be staged using Google tools but disks that can handle the required rates are scarce (local SSD drives). A write-to-memory scheme was therefore developed. At highly parallel workloads, scheduling become very important and these systems are still undeveloped in Kubernetes.

IV.1.6 Easing Kubernetes Deployment

Sanjay Arora (RedHat) presented *Ecosystems II: RedHat OpenShift*. OpenShift is distribution of Kubernetes that makes on-prem clusters easier to deploy and maintain. The model is similar to

RedHat release of Linux, and there is an open source equivalent to CentOS for OpenShift: OKD. If Kubernetes is to play a central role in a re-engineered WLCG computing infrastructure, its distribution and management could benefit from solutions such as these and similar.

IV.1.7 Accelerated Systems

Andrew Chien (UChicago) presented *Accelerated Systems and Optimization*. System scalability research offers an opportunity to optimize use of resources in specific areas in an end-to-end computing system. In a local database, it is about query optimization with predicates and filters; by exploiting selectivity one can increase the scalability and performance of a system. In a public cloud context, services such as AWS S3 Select offer the ability to accelerate services through partial selection of objects before delivery to clients. Opportunities in HEP include (optionally hardware) filtering in strategic locations to reduce traffic on the wide area network. Previous findings on studies of data transformation with recoding accelerators, with programmable accelerators allowing the right representation choice (and format) performed in the *right place* have potential for cheaper computation, less data movement and higher performance.

Throughout the presentation a number of *dimensions of benefit* involving system optimization choices with acceleration were identified:

- Reduction in parsing & filtering costs (through acceleration)
- Performance through more aggressive query optimization
- Representation of encoding in Query Execution Plan
- Beyond tuple to block/transpose, special recoding (ex. ML inference and analysis for systems), etc.
- Expose new optimizations, eliminate transformation overhead
- Reduces the CPU load, offloaded computation, reduce total data processed
- Shift of Filtering computation to Storage node
- Reduces Data Center network load

IV.2 Afternoon Discussion

In thinking of concrete progress for AS-SSL activities, REANA was identified as a likely good first deployment target for the SSL. In addition, ServiceX as deployed on Google Cloud Platform via Helm would make it an interesting use case if the SSL API is Kubernetes.

We discussed an REANA deployed on an SSL cluster provisioned with OpenShift. Success of the deployment would be to have two independent sets of people deploy. Templated python scripts, HELM charts, and YAML would be the ingredients. What would be needed from the SSL?

- Storage
- Internet ingress
- Load balancing
- Pod capacity

Other questions arose: *Is there one SSL (as a service) or a standard?* As a pattern, we'd like not to have a site administrator having to follow notes off a twiki to stand up Kubernetes. We'd like the admin to connect their nodes to an SSL console and everything is automatically deployed. The process to join must be lightweight. If one observes a deployed pattern, they should be able to deploy the pattern to a local environment.

If you use SSL then you should at least publish deployment instructions so others can reproduce it. An R&D Hybrid Cloud provider. Try to make the substrate as compatible as possible with CERN-IT and FermiLab.

Question: *if a federated set of Kubernetes clusters is the substrated approach for SSL, do we lose anything there? Are there any blockers to this approach for the AS R&D plans on SSL?* The answer for AS was that there does not seem to be, but maybe HPC integration would need further consideration. For scaling test, federating SSL clusters will be desirable.

Metric data would be useful to have from SSL. Dashboards and retention of results, the ability to mine metadata indexed by ElasticSearch, use of Kubernetes monitors such as Prometheus would provide some of this. Developing the complete suite for logging and metrics collection is out of scope of SSL, but maybe provide a few standard tools and a repository to collect notes on best practice.

There were comments that not everyone in our group knows how to build a Helm chart, the (current) defacto standard for Kubernetes deployments. While that is true, often the stuff we are all working on is going to fit into a Helm chart eventually - so there is something to be said for being able to at least use it. Docker/desktop can run kubernetes and I don't know how hard it is to make a flexible chart that can scale from a single node cluster to many, but this means that the person working on the component can basically run it in the environment they will eventually have to run in. Generally it was agreed the burden would be to the R&D areas to define the metrics and guide SSL on what data to retain.

We discussed SSL **service Level agreements** between research teams and the SSL; their role and utility for setting developer and resource provider expectations. Some ideas included agreement to publish deployment artifacts; ability to request time and scheduling for scalability tests; agreement the needed information is agreed upon before devoting significant resources.

Opportunities & planning for resources was discussed. Should there be regions of the SSL where clusters nearby are logically group or technically joined via a federation or mesh software. It was agreed that successful contributins would rely on tools that allow operators the ability to re-create SSL environments at another location. Specific initial sites discussed included:

- NCSA: the ISL (Integrated Systems Lab), the Openstack cluster, Blue Waters (a short term allocation), the Illinois Campus Cluster (opportunistic use of GPUs), and an NSF MRI Deep Learning research platform
- Redhat (Openshift) - there is the Massachusetts Open Cloud
- CERN
- NYU
- Fermilab
- BNL
- SDSC and the Pacific Research Platform

IV.3 Day 2 Breakout Session

On day 2 we broke into two groups for more detailed discussion. The first examined specific issues related requirements from AS year 2 plan. The second on SSL infrastructure and identifying the impactful path moving forward.

IV.3.1 Analysis Systems Breakout

IV.3.2 SSL Infrastructure Breakout

V Action Items

The following action items were identified:

1. Deploy an AS validator application on the initial SSL cluster at UChicago
2. Identification of institutes with interest in building the SSL substrate
3. Organization of regular SSL technical meetings

VI Feedback from Attendees

- Some participants expressed interest in having more preparatory documents available in advance of the meeting. This was seen as particularly important for participants new to high energy physics computing or IRIS-HEP, including potential industry partners.
- In the preceeding Analysis Systems Topical Workshop, much discussion focused on establishing community development patterns and management infrastructure more in aligned with best practices found in professsional engineering settings.
- The planning for the meeting should have begun sooner to ensure the needed representation from participants and subject focus.

VII Summary

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

A Revision History

- Version 0.0
  - Initial version
- Version 0.1
  - Version for IRIS-HEP Executive Board review