PJAS Selection – May-June 2017

Deadlines:

17th May 2017: send the projects to the institutes for publication

12th June 2017: share the candidates CVs with the SLs

19th June 2017: short list compilation and start the applicants interviews

26th June 2017: start the documentation preparation

1st September 2017: PJAS contracts start date (depending on the PJAS availability)

Projects:

Dep/Group	Project name	Project description	Technical skills
IT-CDA-WF	Evolution of GitLab Service and Migration of SVN to GitLab	With the deployment of GitLab as the main code hosting platform, and its importance for both the IT department and the LHC experiments, the selected person will work on ensuring the stability and evolution of GitLab. Regularly new versions of GitLab become available which introduce new features. These new versions need to be tested and prepared for deployment before they are installed on production. In addition, in view of the discontinuation of AFS we expect to discontinue SVN by end of LS2 (Long Shutdown in 2019). Since GitLab works very differently than SVN, significant preparation work is required and we need to provide CERN developers with tools and recipes to migration their code from SVN to GitLab.	- Experience or strong interest in Version Control systems (SVN, Git) and code development lifecycle (code review, tests, pull requests, open source projects hosted on GitHub.com or similar platforms...) - Effective skills for communication with team members and the community of developers at CERN; interest in providing high quality support to the developer community - Goal-oriented personality with ability to resolve issues autonomously - Experience in Linux system administration and Docker containers would be a plus - Knowledge of Ruby and Go would be a plus

IT-CF-FPP	Enabling automated collection and analytics of hardware failure statistics	Develop spare parts stock management in Infor EAM assets inventory and provide a suitable tool for repair contractor to enter the data. This will enable parts tracking, which is a necessary component for calculating annualized failure rates (AFR). Furthermore, the technical details about the failures are recorded in SNOW and that data also needs to be formatted to ease analytics of different failure modes and their evolution over time.	- Linux, in particular RHEL and CentOS - Bash and some "sysadmin" basic knowledge - Python (intermediate/advanced) - SQL (intermediate/advanced) - Basic knowledge of time series DB (Influx) - Basic knowledge of statistics - Familiar with machine learning - Excel (intermediate/advanced) - Knowledge of Grafana is a plus
IT-CM-IS	Accounting Enhancements	CERN's computing centres run more than 200k CPU cores dedicated to physics data processing. With such a large data centre, it's critical to ensure that the resources are being used efficiently and correctly and to ensure the prompt and automated detection of operational problems. The project will enhance CERN IT's accounting and ops analytics activity, with aim of using our new analytics and dashboard machinery (based on Hadoop/Spark/ES/Kibana) to help identify and resolve issues, both individual node issues and global effects, due to, for example, configuration changes or changes in experiment usage patterns (e.g. efficiency problems due to a new release globally using more swap). The project will also extend the current accounting approach used for compute resouces to all IT services, both cloud, services and physical hardware based. This will allow IT to account in an automated way for its services and help stakeholders understand the resource usage in the centre.	The team is organised in a devops style and currently makes use of Python, Go, Spark, Hadoop, ElasticSearch, Grafana and Kibana.
IT-CS-CE	DNS Management code rewriting	The configuration of the CERN IT central DNS servers (based BIND DNS) is generated automatically (from zero) every 10 minutes thanks to a software developed at CERN several years ago and evolved over time. This in-house software based on PERL is reaching its limit in term support and architecture. The proposed project is to redefine the DNS network architecture to introduce a redundant solution for the master DNS and recode, with a modern language (such as Python).	Programming ability, preferably with knowledge of Python. Some knowledge of DNS would be useful but is not essential

IT-ST-DA	Project as part of the Analytics Working Group	The IT analysis working group has, based on data collected by the IT monitoring project, created an analysis repository with computer center metrics and successfully combined IT infrastructure and experiment workflow data to construct an analytical model for CPU-bound workloads. In a second step we plan to run an automated workload classification (currently prototype based on random forests) of user jobs (based on existing metrics) to extend the model from CPU-bound jobs (eg simulation) to other job types with more significant worker node disk and EOS disk access. The candidate should have basic experience with statistical data analysis with (either in python, root, spark or R) and an interest to quantitatively evaluate computer center models. She/he would workin close collaboration with the CERN service experts and developers and would be expected to participate in the evaluation of new data management strategies (eg different balance between archive and direct access storage, delayed data selection steps etc) as preparation for HL-LHC.	- basic experience with statistical data analysis - either in python, C++/root, spark or R - an interest to quantitatively evaluate computer center models
RCS-SIS-AO	Software engineer - INSPIRE digital library, back-office integration and consolidation	INSPIRE (http://inspirehep.net) is the information platform used by the entire High-Energy Physics community to access information about over 1 million scientific artefacts: articles, preprints, theses, conference proceedings and increasingly data and software. INSPIRE is developed and operated at CERN, with contributions from leading High-Energy Physics partner laboratories (DESY, IHEP-Beijing, Fermilab and SLAC) and in cooperation with partners in France and Japan. Beyond its role as a digital library, INSPIRE disambiguates author names and attributes to individual researchers their scientific production, with about 100'000 profiles built, through a mix of machine learning, crowdsourcing and manual curation. INSPIRE also provides community services, index job opportunities and alerting to upcoming conferences. INSPIRE is developed (http://github.com/inspirehep) on top of the Invenio Open Source digital library (http://inveniosoftware.org). The successful operation of INSPIRE service relies on a complex workflow where content from a hundred multiple sources (scientific publishers as well as other	- Web development - Python - Python web framework (preferably Flask, or Django) - SQLAlchemy ORM, - Elasticsearch - Redis - Celery - Optionally, Javascript (preferably AngularJS (1 or 2), and Bootstrap) - Optionally, Scikit-learn

		community services such as arXiv.org) is harvested, promptly enriched and automatically linked with existing content and cross-referenced to external services. Users also can suggest additional content to INSPIRE, which is processed in similar ways. INSPIRE is currently being re-implemented on top of a modern software stack based on Python, Flask, SQLAlchemy, PostgreSQL, AngularJS, Bootstrap, Elasticsearch, and Scikit-Learn. As part of this process there is an opportunity for a software engineer to help integrating and consolidating existing sub-projects in the new stack. These ranges from a machine learning algorithm for author disambiguation, to a web scraper and content aggregator; from several JSON-based manipulation tools to maintain database entries, to a tuning of Elasticsearch mappings for INSPIRE bibliographic data; from bringing into production-level exisitng algorithms based on machine learning, to defining in partnership with librarians rules to identify and merge duplicate database entries. The candidate would join a dynamic and agile team of junior and senior developers, collaborating with content curators both at CERN and in the other participating sites.	