First Ideas for a Software Institute for Data Intensive Sciences

(SIDIS)

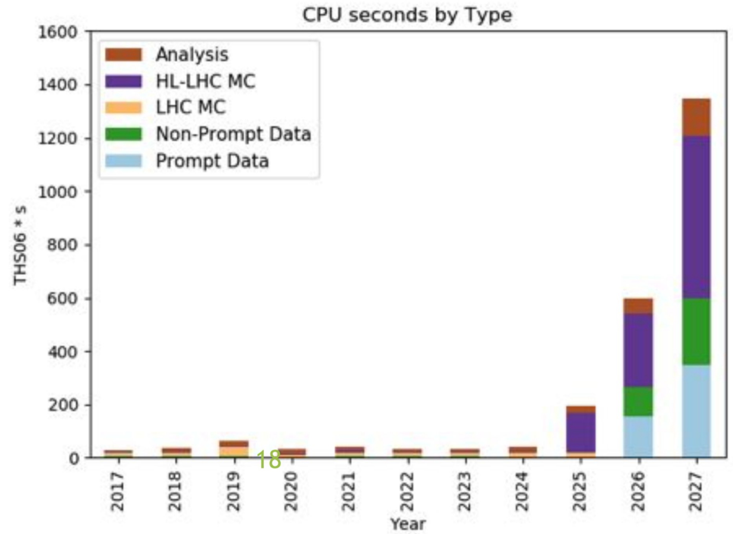
Ian Bird, Simone Campana, Pere Mato, Stefan Roiser, Markus Schulz, Graeme Stewart
CERN EP/SFT & IT/DI

SIDIS
SOUND FOUNDATIONS

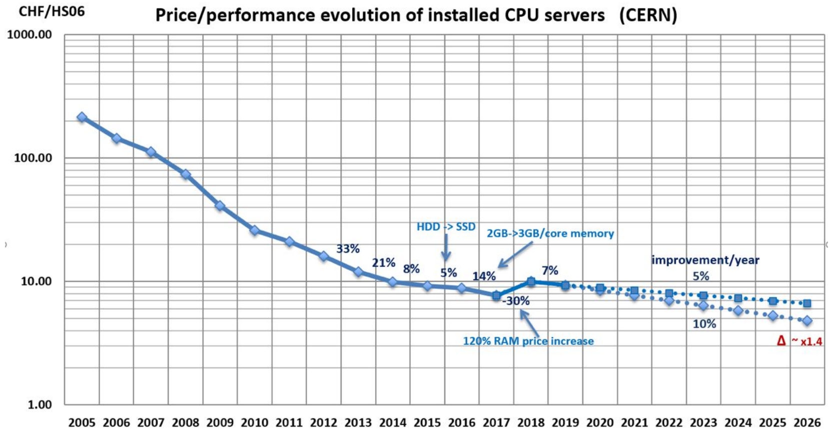
SIDIS
FOUNDATIONS

There are challenges ahead ...

- The HL-LHC, will pose major challenges on the data processing infrastructure of experiments
    - New programming paradigms and techniques are needed to exploit advances and changes in hardware
        - parallelism, vector instruction sets, non x86 architectures, machine learning…
    - The assumed ~20 % yearly increase in performance isn’t certain anymore
        - Streamlining can compensate only some of this
        - Need to invest into software development to gain the needed factors in efficiency



CMS CPU needs projection



CERN CPU price / performance evolution

Other challenges (incomplete)

- Complexity of the code base
  - several million lines of code, O(1000) developers
    - integrated over lifetime
  - high algorithmic complexity
  - long evolution >10 years
- Career prospects for Physicists with a focus on software
  - no clear path within academia
  - often discourages post graduates to work on software
    - very different to the situation in detector development!
- Retention of knowledge
- Current educational focus on managing complexity
  - less on efficiency

The Who and the What

- Propose a strong collaboration among European research institutions & labs, European Universities and scientific collaborations on software R&D, engineering and sustainability to tackle those future challenges
    - Initiative in addition to already ongoing efforts in different countries
- R&D areas to concentrate on are
    - Application software (data intensive and algorithmic complex software)
    - Distributed computing (data management and data processing frameworks, data analysis facilities, exploiting HPC architectures)
    - Focus on fundamental, transferable aspects
        - data structures, methods, best practice,.....

The What (ctd)

- Access, preservation and dissemination of knowledge
  - A lot of expertise available within the community, academic and industry partners
  - Many successful initiatives which can benefit the community as a whole
  - Together with existing computing schools development of new curricula

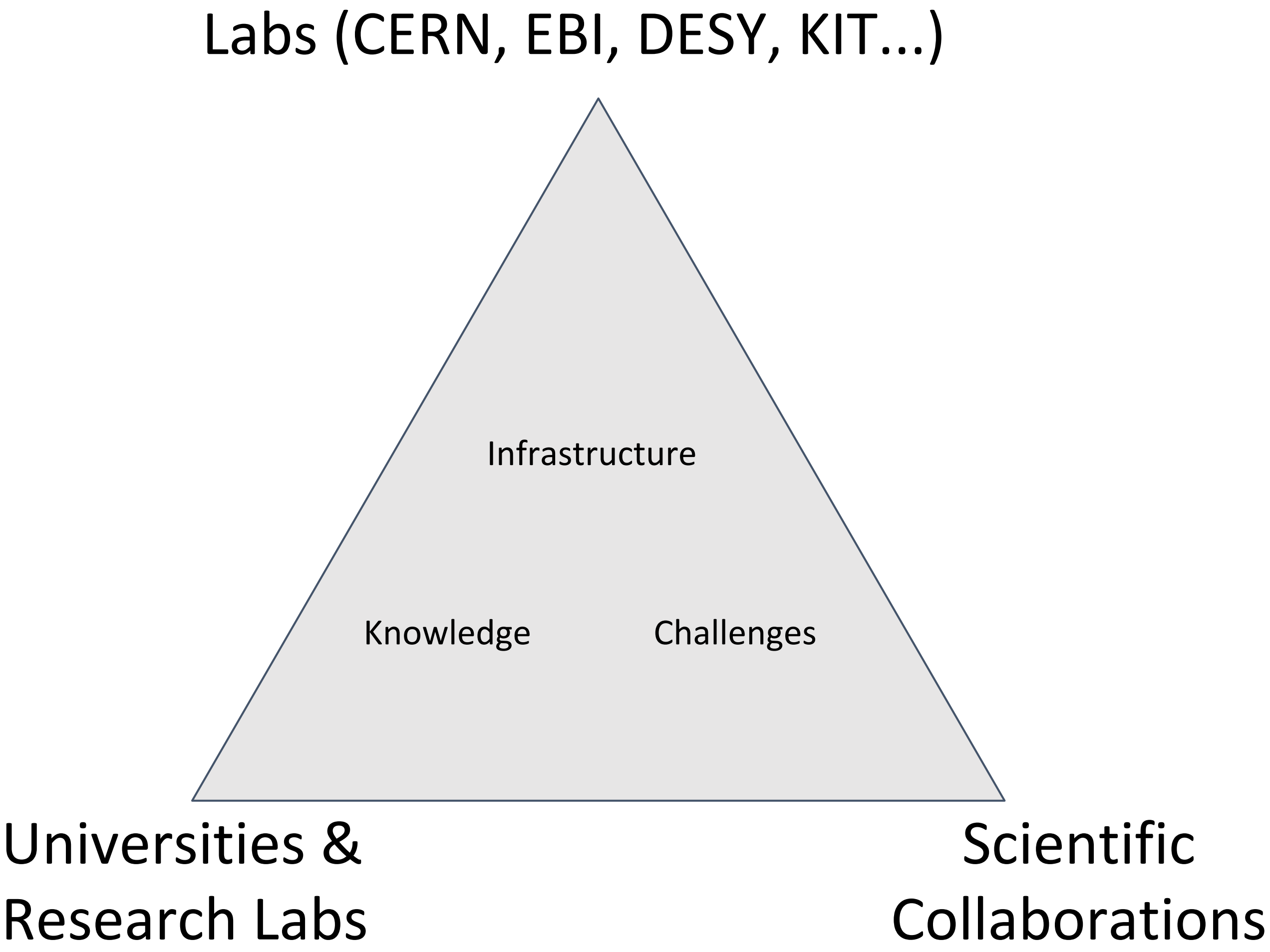
Advanced C++ course in LHCb

1 week hackathon on upgrade R&D

Mandate of the institute

- Establish a group of researchers to lead and engage in R&D and engineering activities
    - Exchange of people between universities, research institutions & labs
- Enable training and qualification for students and young engineers
    - Leverage on already existing activities (UK/”Data Scientist””RSE”, DE/ErUM, …)
- Develop a career path for data science/software engineers in sciences
    - In close collaboration with academic institutions
- Act as a lobbying organisation towards funding agencies
    - EC and national agencies
    - promote real investment in advanced software skills

Interplay of stakeholders



Labs (CERN, EBI, DESY, KIT...)

Infrastructure

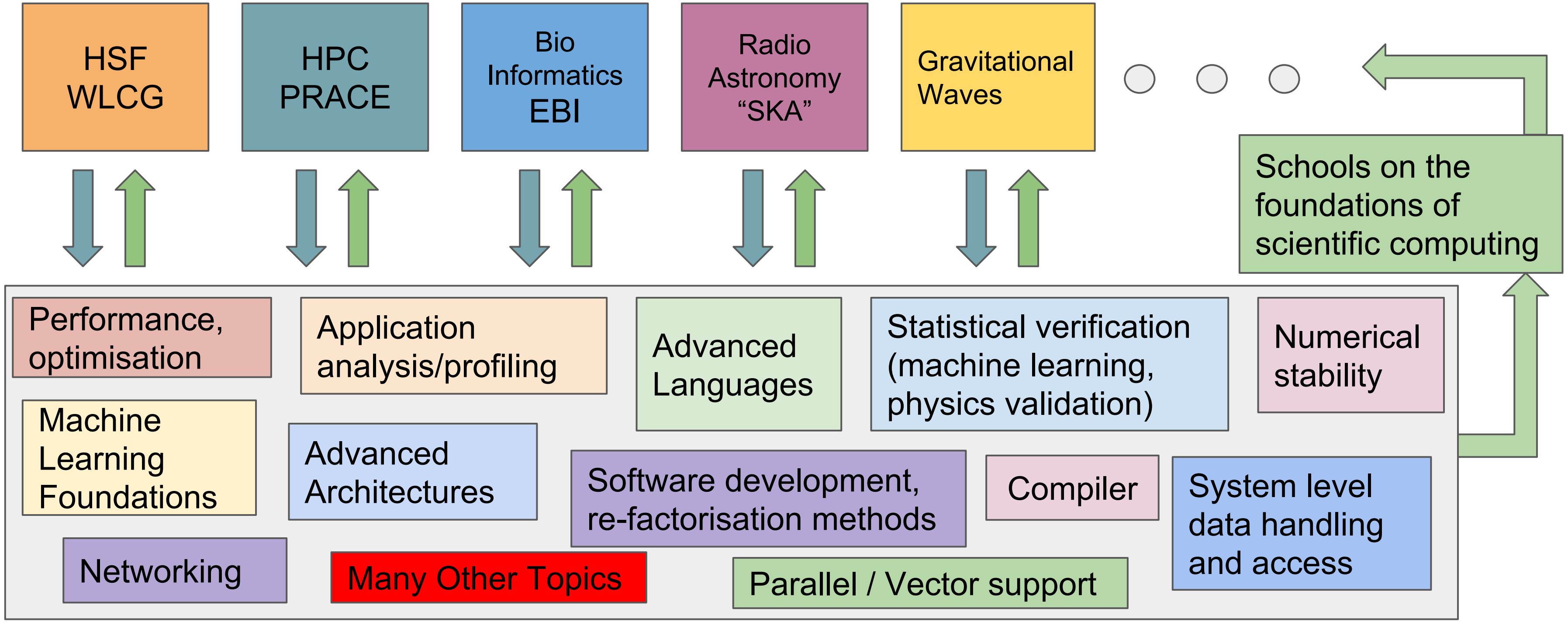
Knowledge

Challenges

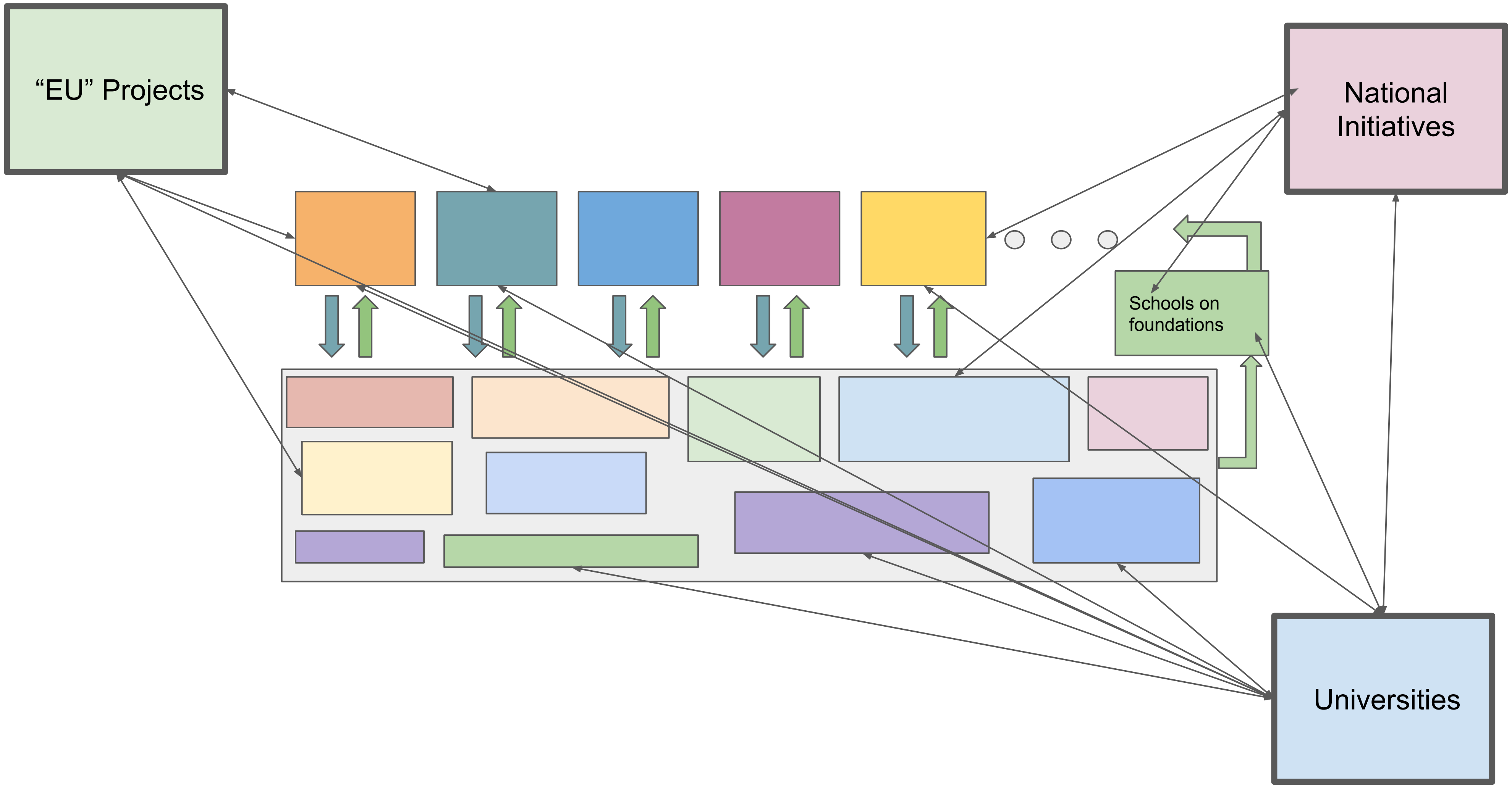
Universities &
Research Labs

Scientific
Collaborations

Topics and interactions with computing intense sciences


HSF
WLCG
HPC
PRACE
Bio
Informatics
EBI
Radio
Astronomy
“SKA”
Gravitational
Waves
Schools on the foundations of scientific computing
Performance, optimisation
Application analysis/profiling
Advanced Languages
Statistical verification (machine learning, physics validation)
Numerical stability
Machine Learning Foundations
Advanced Architectures
Software development, re-factorisation methods
Compiler
System level data handling and access
Networking
Many Other Topics
Parallel / Vector support

Biotope


"EU" Projects
National Initiatives
Schools on foundations
Universities

Some common benefits

- A hub for R&D and exchange of ideas and people
- Preservation and access of engineering knowledge and experience
- Students receive additional valuable training and career opportunities
- Teaching opportunities and external visibility for lab employees
- Dual career opportunities for computing oriented physicists
- Experiments get solutions for their future challenges
- Additional international collaborations for universities
- Framework for leveraging national funding opportunities
- Recruitment pool for future hirings via working with students

Interplay with other parties?

- HSF, WLCG and others are key players to establish discussions among the experiments
  - Can those fora collect relevant input for common R&D work?

- Collaboration with IRIS-HEP
  - Can this institute act as the European part of the HEP software initiative?

- Openlab, EU projects (BEST4HEP, …)
  - Strong coordination needed. How to interact?

How to reach out further?

- How to engage applied computer science institutes?
  - Get in contact with software engineers
    - Within the scope of Master and Phd thesis work this has been done successfully in the past, but no systematic approach has been established
  - How can this be done?
    - A tangible entity can help in connection interested parties
- How and at what level should we engage with other data/compute intensive sciences?
  - From the start, in a second step?
  - With those that are somewhat similar to us (HEP) like SKA?
  - Balance between scope and complexity …. (see early grid projects >30)

Summary

We propose a scientific software institute which aims at:

- R&D on future software challenges in natural sciences
- Training, qualification and recognition for the next generation working on software
- Foster collaboration between universities, research labs, sciences
- Promote and establish a career path for applied software engineering for science
  - a bit like the UK RSE program (Research Software Engineer)

Discussions are at a very early stage!

Next step is to use the many existing point to point contacts between HEP and Computer Science people to investigate further how to proceed best

So far the consensus is that software will become more important

- and that we have to do something….

Many thanks !

- For a lot of feedback and input from
    - Tommaso Boccali
    - Concezio Bozzi
    - Pete Clarke
    - Davide Costanzo
    - Michel Jouvin
    - Thomas Kuhr
    - Gonzalo Merino
    - Andrea Valassi

First ideas on a name

Software Institute for Data Intensive Sciences

(SIDIS)