IEEE eScience Conference

Graeme Stewart and Radu Popescu

2018-11-19

# 14th eScience IEEE INTERNATIONAL CONFERENCE

29 October - 1 November in Amsterdam

Conference Webpage

Purpose of this Conference

- ## eScience
  - Impact of computer technology that opens up computationally intensive research in science domains
    - Both big data and intensive processing
      - "Big" depends a lot on your science domain
  - Described as the "fourth paradigm" of science by Jim Gray
- ## Audience
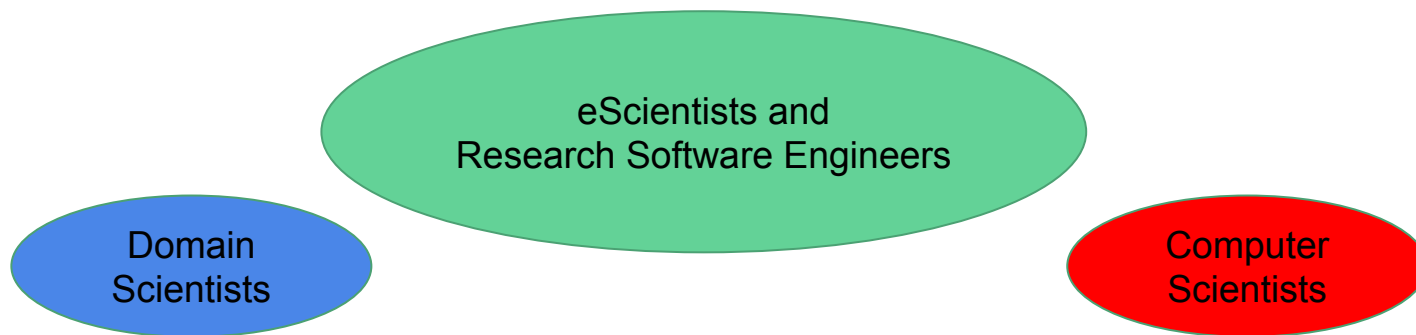
~250 overall, many first time participants

eScientists and
Research Software Engineers

Domain
Scientists

Computer
Scientists

Fomat

- Day 1 - Workshop day
  - Research Objects, **International Workshop on Sustainable Software for Science**, Benchmarking Algorithm Performance for Research, eScience-FAIR Science, Contemporary Peer Code Review in Scientific Software Development, **Generic components of the eScience Infrastructure Ecosystem**
- Day 2 - Plenaries
  - Invited keynote on Digital Humanities, Selection of eScience papers
- Day 3 - Domain and Application Science Days
  - Weather & Climate Science in the Digital Era, Data Handling and Analytics for Health, Advances in eScience for the Humanities and Social Sciences, **Exascale Computing for High Energy Physics**, General eScience
- Day 4 - Plenaries
  - Invited keynote on Real-time Analysis in Astronomy, Feedback Session, Selelection of eScience Papers

Conference dinner in the Maritime Museum

But wait, what? No Indico?

- Conference had no place for speakers to upload slides
  - So no way to check some details before asking a question
  - No reference for the talks
- Conference Papers hidden behind an IEEE paywall for a year
- Made me really appreciate what we do with Indico in HEP

9:00 10:00	**Keynote: Data Science or Data Humanities? Opportunities for Digitally Enabled Analysis of History, Culture and Society** Melissa Terras (University of Edinburgh) **Less information** This keynote will look at the opportunities, issues, and rewards possible for researchers developing data-led approaches to answer research questions in the Arts and Humanities. How can we build and utilise appropriate computational methods for the analysis of our past and present societies? What possibilities and barriers are there in working in this crossover point from data science to the humanities? From multispectral analysis of Egyptian Mummy cartonnage, to the development of Handwritten Text Recognition for archival material, and the mining of millions of words of historical newspaper archives, this talk will showcase a range of innovative international research projects, whilst also giving pointers on how others can approach this interdisciplinary space successfully.
10:00 10:30	**Digital Methods in Holocaust Studies: The European Holocaust Research Infrastructure** Daan de Leeuw et al. **More information**
10:30 11:00	**Coffee break** Location: Atrium Foyer

Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE6.1)

- "It is impossible to conduct research without software" (70% in UK-SSI survey)
  - How to stop software being a disposable, unreproducible part of science?
- Key role of eScientists and RSEs is to encourage good practices in scientific software
  - New roles being identified, like Data Stewards (involves librarians and archivists, as we do in HEP)
    - Other people will care about your data, but you have to also help them
  - Analysis of collaborations and the impact of 'distance' on effectiveness and communication
    - Scales of 100m seen to have in impact
    - I think in HEP we do much better - there's a phase change in large collaborations
  - Software directory has been developed by the Netherlands eScience Centre, storing software metadata alongside the code itself (slides on Zenodo)
    - Could be a good solution for HSF, where our current Knowledge Base didn't really catch on

Speed Blogging

- Interactive session idea
  - Choose from a menu of possible topics, generate a group of 4 or 5 people
  - Work for 90 minutes with the aim of writing a blog article that summarises your thinking on the idea
- I worked on Sustainable Domain-Specific Training
  - How to generate training material that goes beyond the basic skills?
    - Experts are precious and already busy
      - Try to make their input re-usable (adopt software carpentry templates)
    - Share with close disciplines if you can
      - In HEP, e.g., a lot of knowledge is cross experiment
    - Emphasise generic principles and any transferable knowledge
      - Helps ground people in why, not just how
      - Helps convince reluctant supervisors
    - Build a training community

Generic components of the eScience Infrastructure Ecosystem

Used NIKHEF Indico: https://indico.nikhef.nl/event/1316/

Overview:

- Oxana Smirnova (Lund) - Sustainable software as an infrastructure
  - "Data ages like wine, software ages like fish"
  - Infrastructure software is a research tool that needs to be funded appropriately
- Radu Popescu (CERN) - An overview of CernVM-FS for data distribution
  - CernVM-FS can complement dedicated data distributions systems in certain cases
- Dennis van Dok (NIKHEF) - SoftDrive.nl - user areas backed by CernVM-FS
 - CernVM-FS now has support for this workflow

Generic components of the eScience Infrastructure Ecosystem

Overview (contd.):

- Massimo Lamana (CERN) - SWAN: Jupyter-as-a-service
    - Web-based notebooks lower the barrier for collaboration
- Tibor Simko (CERN) - REANA: Reusable data, reproducible analyses
- Jaroslava Schovancova (CERN) - Data lakes
    - Combine geographically distributed storage pool to reduce costs, hide latency and average out bandwidth
- Martin Brandt (SURFsara) - Building next decade's infrastructure: yourself, federated, or public, in the HNSciCloud
    - Public session with conclusions of project at CERN, Nov 29th, www.hnscicloud.eu

Plenary Sessions

- Great Keynote from Melissa Terras (Edinburgh) about digital humanities projects
  - Handwriting recognition for historical documents
    - Crowdsourcing transcription (Transcribe Bentham)
    - Train DNNs to then accelerate the process
  - Multi-wavelength scanning of old texts and drawings
    - Faded and degenerated documents
    - Distinguish human marks from wear and tear
    - Revealed horse drawings from Leonardo da Vinci unseen for centuries
- Audio monitoring of wildlife
  - Try to identify cane toad arrivals on a biologically isolated Australian island
    - "One terabyte a year, we couldn't just upload that..."
    - Open software, open data, open hardware?
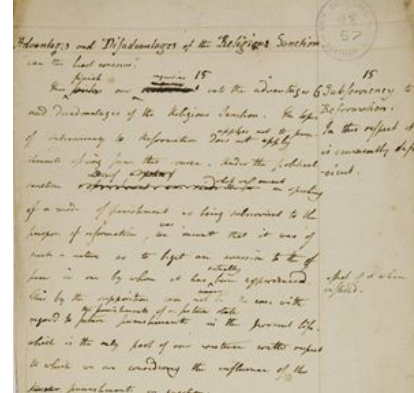
Plenary Sessions

- Real Time Multi-Wavelength Astronomy (LOFAR and Dwingeloo), Joeri van Leeuwen
  - Fast Radio Bursts (FRBs), incredibly intense millisecond bursts in radio frequencies
    - Use dispersion to measure the 'distance' - these are extragalactic
  - No one knows what thay are
  - GPU cluster analyses Dwingeloo signals in real-time
    - Alert message to LOFAR that 'freezes' the sky in its readout buffers
      - Then takes 24 hours to download the data
  - Aim to greatly enhance the catalogue to shed light on the origin of these objects
- Coupling Exascale Multiphysics Applications
  - Oak Ridge project to couple different plasma physics simulations together
  - "Message passing" between the particle code (accurate, but slow) and the fluid code (faster, but not all physics processes)
  - x10 speed-up compared to using only the particle flow
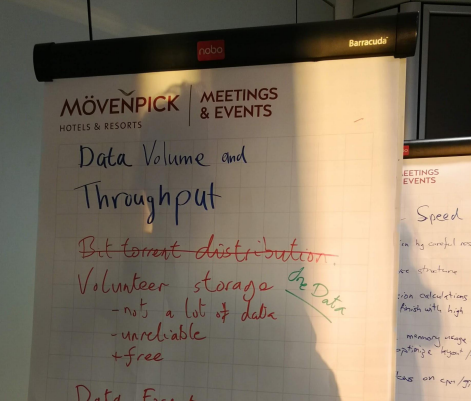   - This is a huge application - weeks on full Oak Ridge machine
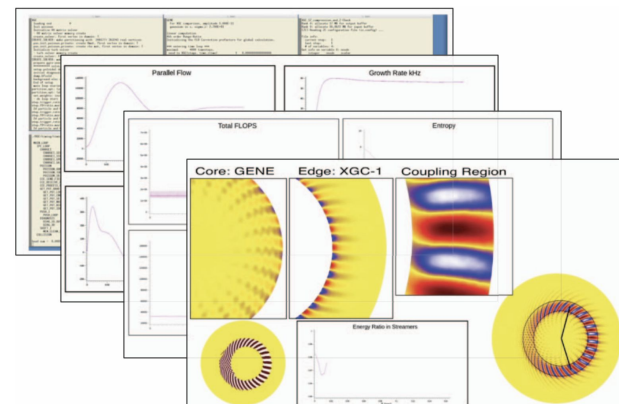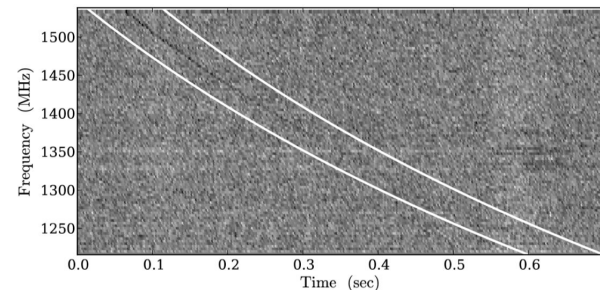
HEP Exascale Workshop

- Aim to attract interest from outside the HEP community in our 'exascale' problems (Jeff Templon, Simone Campana, Graeme Stewart, Yifat Dzigan, Frédéric Suter, Ana Lucia Varbanescu)
  - Produced a challenges document for non-HEP people
  - And we have an Indico too!
- Plenary talk from Miron Livney (project leader for HTCondor)
  - World of computing is changing, we have to adapt and share with other sciences
  - Good designs stand the test of time - stick to your principles (including recognising that perfection is unattainable)
- Overall we had great contributions to the session, but not enough non-HEP people came/stayed
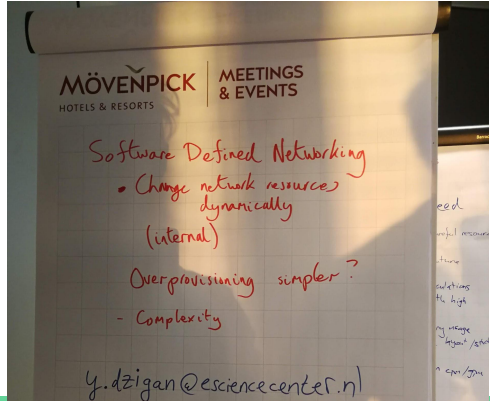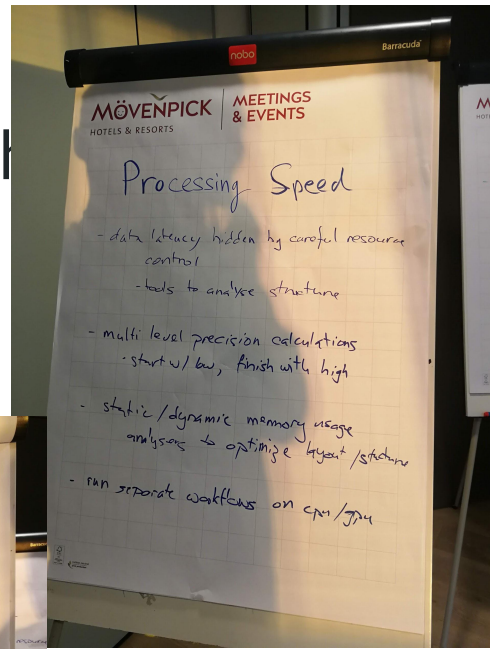 - There were too many parallel applications area sessions

HEP Exascale Workshop

- Interactive session - <u>Brain Writing</u>
  - Idea is to generate new ideas in groups, but avoiding dominance of extroverts and reducing the scope too early to the first idea
  - In groups everyone writes their best idea (2 minutes)
    - Pass ideas to the left, comment/observe/clarify the idea you just got
    - Continue until you get your original idea back, now with comments
  - Then have a group discussion of each idea + comments
  - Then present the best ideas at 'challenge stations', discuss with experts
    - Data throughput
    - Processing speed
    - Machine learning

e Worksh

Data Volume and Throughput

~~Bit torrent distribution~~

Volunteer storage _the Data_
- not a lot of data
- unreliable
+ free

Data Format:
* Jagged Arrays
  Interesting for others?

Processing Speed

- data latency hidden by careful resource control
  - tools to analyse structure

- multi level precision calculations
  - start w/ low, finish with high

- static/dynamic memory usage analysis to optimize layout/structure

- run separate workflows on cpu/gpu

Machine Learning

- Convince industry mgmt to give sabbatical their engineers /DS/ML to work on HEP
  - PR
  - HR

- ML-ify data processing chain (counters → decay topology)
  - debug, systematics?
  - borrow ML community tools/approaches
  - systematic

Bit torrent Distribution

⅋ - seeding / availability
+ fast optimising
+ sharding

Put it all into RAM.
- mixed workflow / sharing

CPU/GPU
+ run mixed workloads
  cache utilisation ?

Software Defined Networking

* Change network resource dynamically
  (internal)

Overprovisioning simpler?

- Complexity

y.dzigan@esciencecenter.nl

No 'world beating' ideas, but excellent for interaction and good discussion of solutions
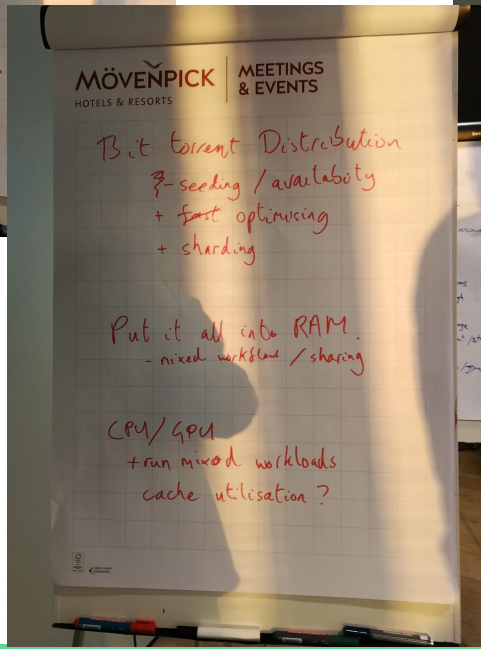
Session got really positive feedback from participants

15
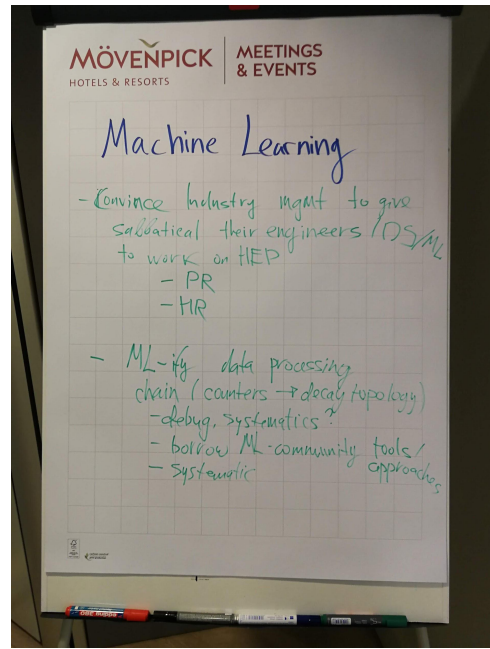
Thoughts and Observations

- Positioning of the conference in research space still seems awkward
  - Still quite distant from domain specific meetings (e.g., for us we have ACAT, CHEP and all the physics conferences)
    - People lack time and money
  - eScience and RSE communities do seem better established and more comfortable in their roles
  - Not enough computer science people as far as I could tell
- Cool tool used for interactive feedback session, https://www.mentimeter.com
- Next ideas?
  - Find thematic and cross-cutting ideas that link communities
  - E.g., I would organise a next application area meeting on 'Exascale Science Challenges'
    - Have abstracts from different domains
    - Wider audience with more diverse ideas and solutions
    - Then hopefully more attractive to computer science community