Trusty Research Infrastructure?

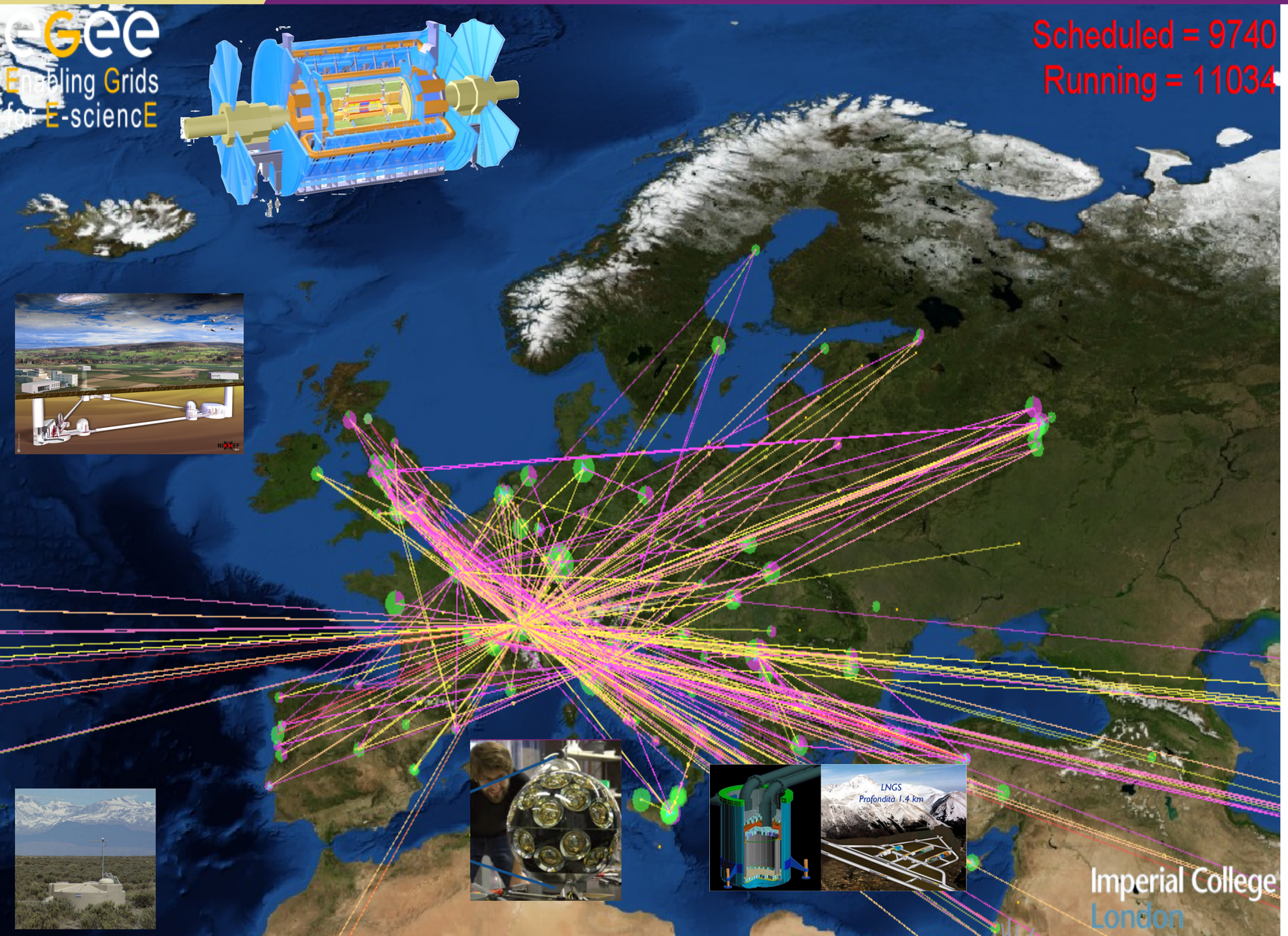
Nikhef

Digital Marketplaces Using Novel Infrastructure Models

I2GS2018 Panel Session

David Groep
davidg@nikhef.nl

Services for (a *lot* of) Research Data



EOSC-hub

EGI EUDAT

GÉANT PRACE

… and that's just the European end!

---

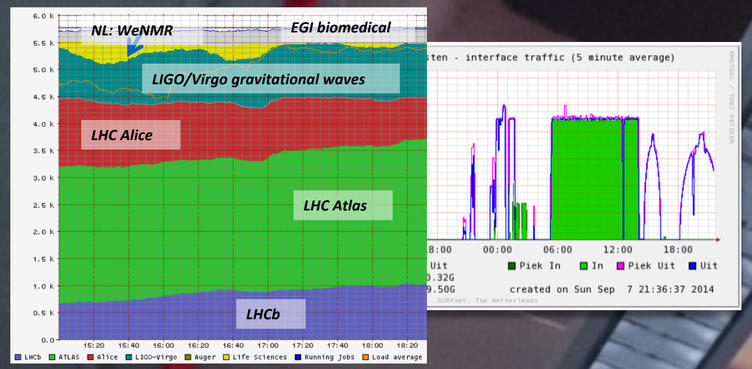
*Sharing common **infrastructure** trust*
and much (and converging) joint AAI services

IGTF Interoperable Global Trust Federation AP|EU|TAG

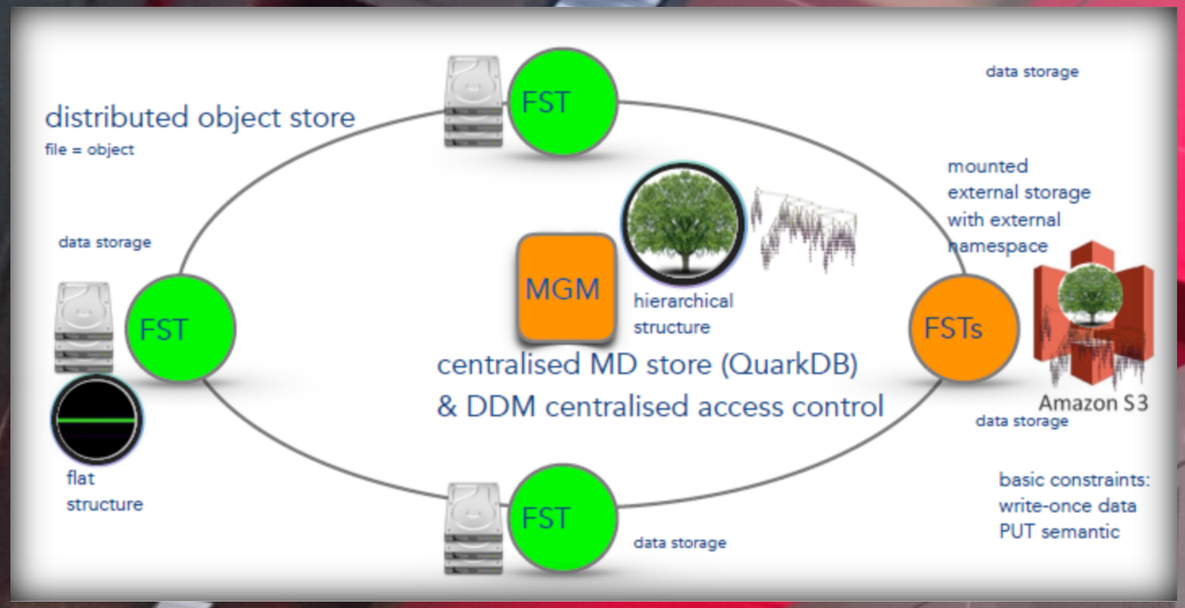
eduGAIN

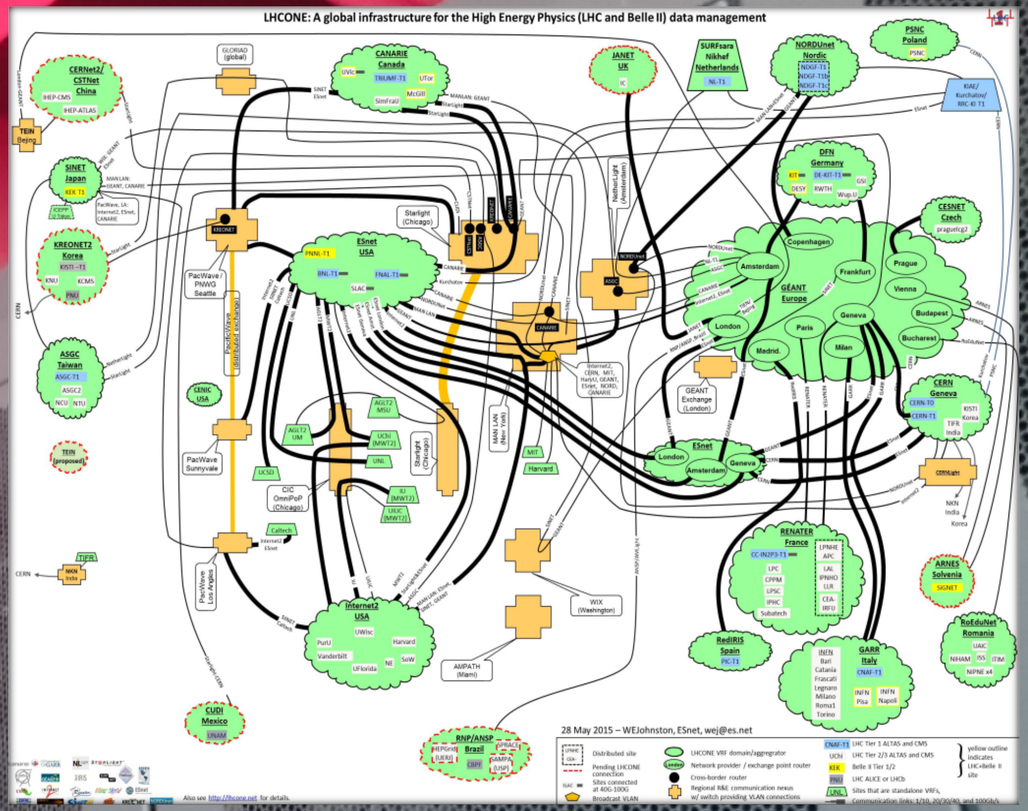
AARC

Built on ‘balanced global systems’



Full mesh with dynamic data placement interconnected with sufficient bandwidth for ‘opportunistic’ movement of data to compute



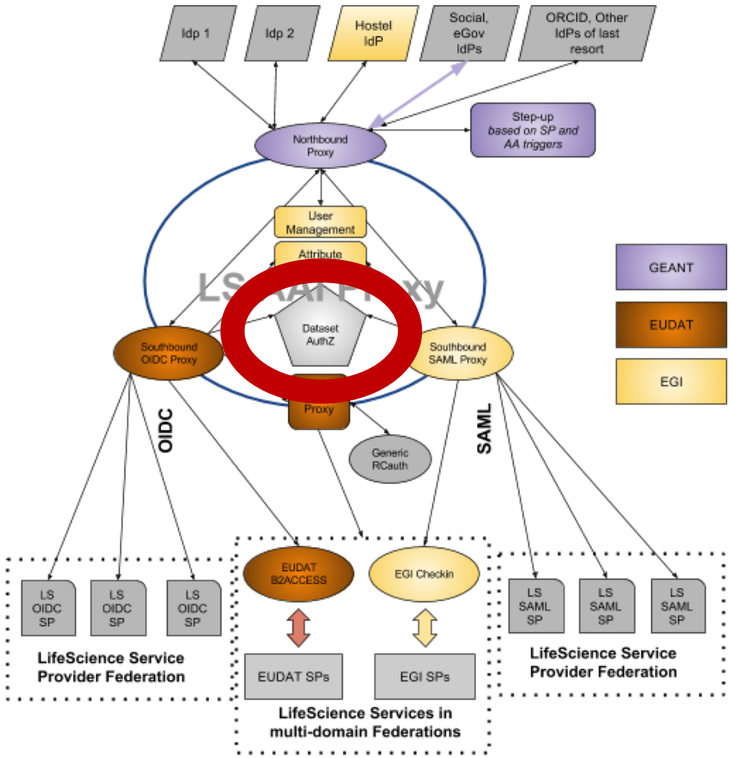


Distributed Object Store (and caching edges not shown) using credential delegation through the caches

Image credit Data Lake: Andreas-Joachim Peters, CERN EOS team,

Image: LHC One

But: use cases requiring more trust?



Life Sciences AAI:
need for a "Dataset Authorization Service"
… but different requirements for, say,
genome infrastructures and biobanks

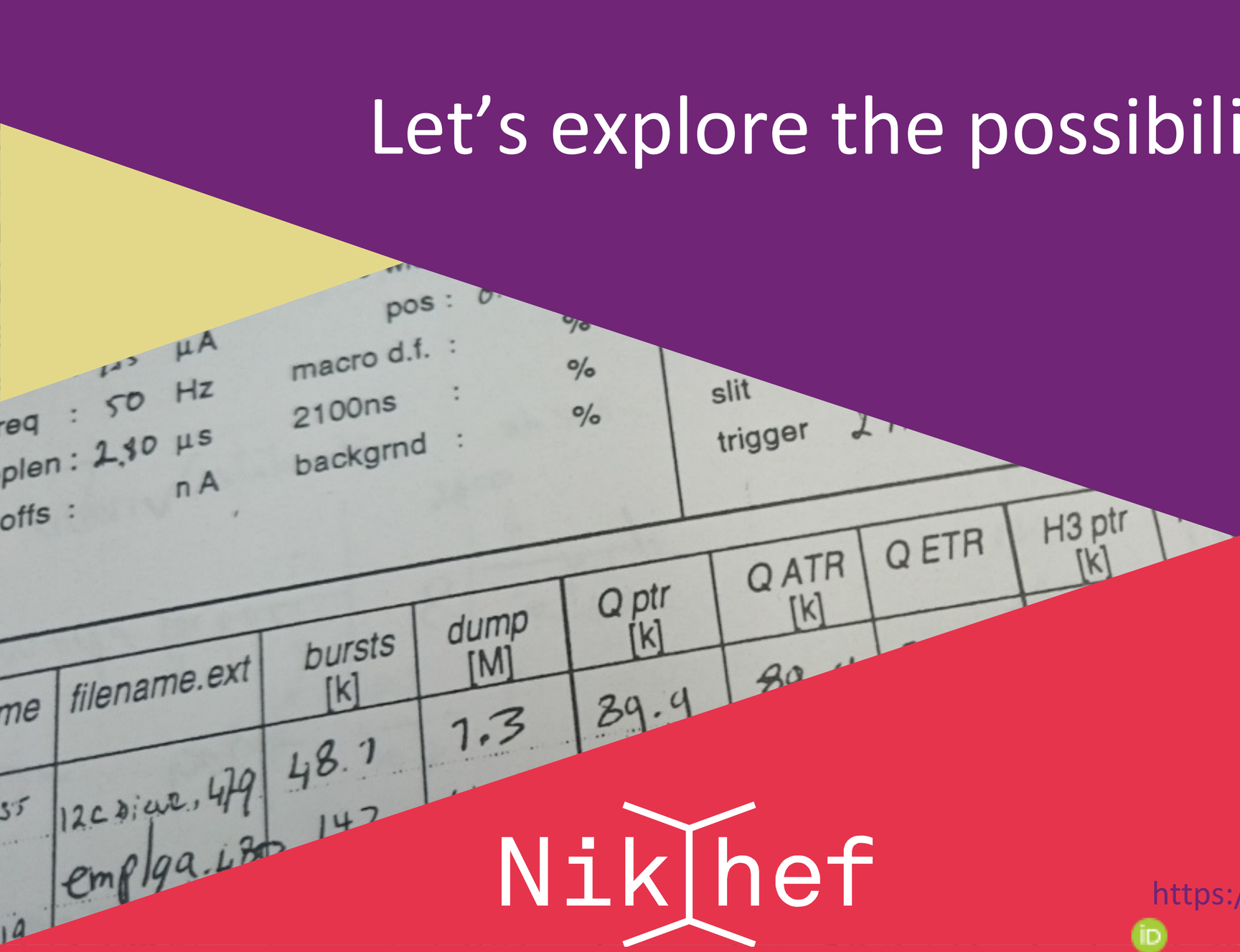
Same researcher, different data sets –
and ethics does not allow them to be combined
we need technical role separation for the same user?

…

Can that be done in an ecosystem of shared services?

Can we retain the efficient balance that today allows us to use any resource anywhere?
Is certification of data centres the solution that will last? Or neutral places?
Will it result in a contraction to only ISO 27k and ISM audited data centres?
Are we about to loose opportunistic resource usage that helped exciting science, from Higgs to GW?
Or can secure overlays or other novel infrastructures help us remain efficient AND trustworthy?

Let’s explore the possibilities!



Nikhef

David Groep
davidg@nikhef.nl
https://www.nikhef.nl/~davidg/presentations/
https://orcid.org/0000-0003-1026-6606

Abstract

Research and e-Infrastructures, and science in general, have for long been generating large amounts of data, and built a distributed infrastructure to cope with it. The premise has been that all components, compute, storage, and the network in-between, are sufficiently balanced that technology limitations in general no longer play a role in data distribution. The prolific use of ‘opportunistic compute’ resources by some of the LHC and other experiments bears that out: moving the data is no longer a ‘real’ issue. But now for the first time policy considerations and the need for higher trust start to change that premise: there are real access control needs on research data relating to people, really large data sets start appearing that have embargo controls on publication, and even a single researcher may have data that is illegal or unethical to be combined – and role separation is needed even here. And the rise of ‘distributed caching’ models for data require access rights to travel with the data – either physically or by cryptographic means. Dataset authorization systems are fast becoming necessary parts of a research AAI solution, and linking compute resources to such research data very much an open question. Yet only through collaborative analysis does the data actually result in real research outcomes.

Will pushing compute to neutral and extra-territorial facilities alleviate the trust issue? Is the ever-increasing push for ISO audits and ISM trust marks for service providers actually reducing risk? Should we abandon the opportunistic resource use and have our analysis models driven by data placement? Do we have to abandon opportunistic compute and move everything to where the data is? Or can we here find a data compute model that allows both data owners to manage access as well as federated service providers to trust that what they the code or container they are about to execute can in itself be trusted?