Future trends in distributed infrastructures - Nordic Tier-1 example

GRID 2016, Dubna
July 5, 2016

Oxana Smirnova

Lund University/NeIC

norden

NordForsk

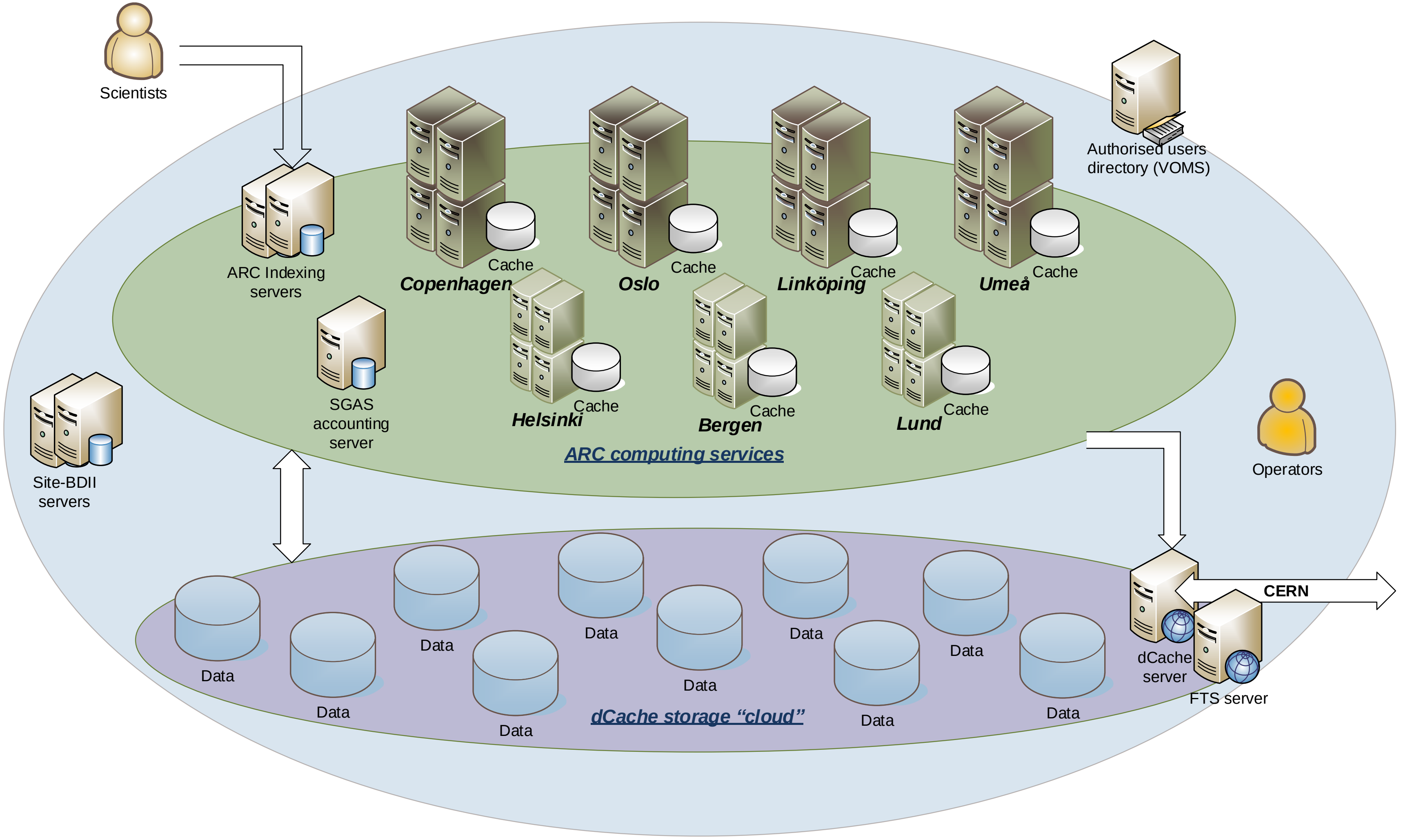
neic Nordic e-Infrastructure Collaboration

Lund University
Faculty of Science

Outlook

- Context: WLCG infrastructure
- Specific use case: the Nordic Tier-1 centre (NDGF-T1)
  - NDGF-T1 is a distributed infrastructure itself
- Computing and storage requirements of LHC will grow very significantly
- How can we meet these requirements?

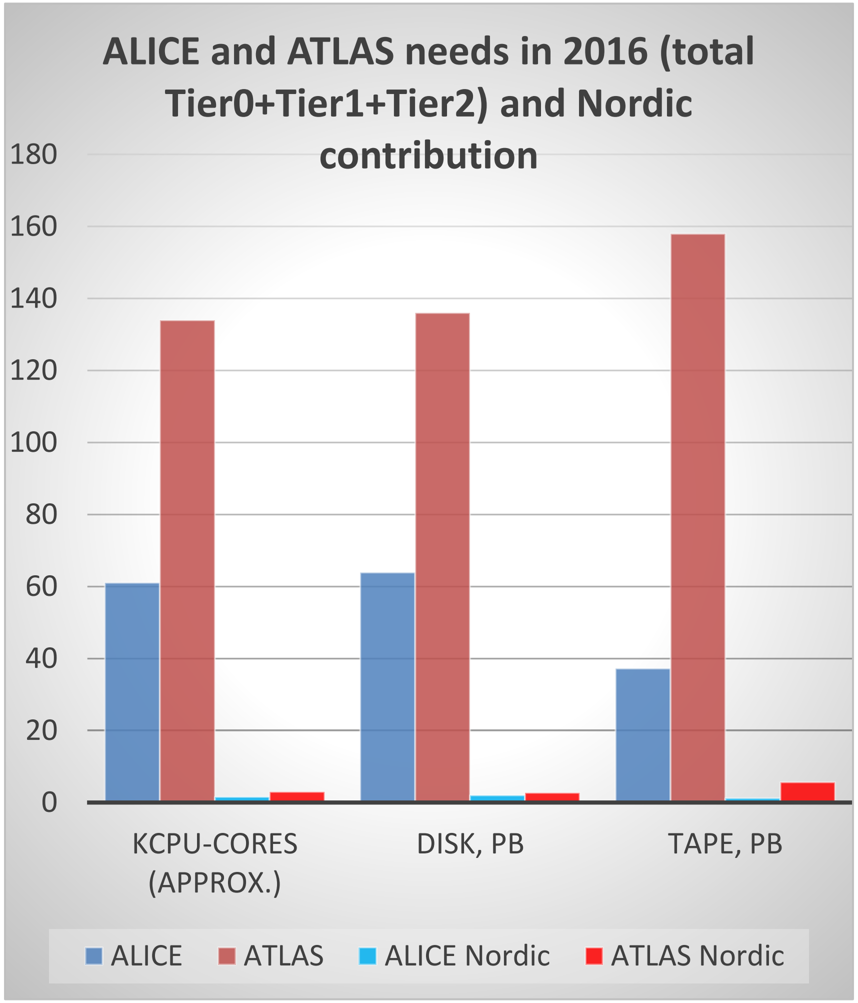
Nordic Tier-1 (NDGF-T1): a distributed center



NDGF-T1 logistics

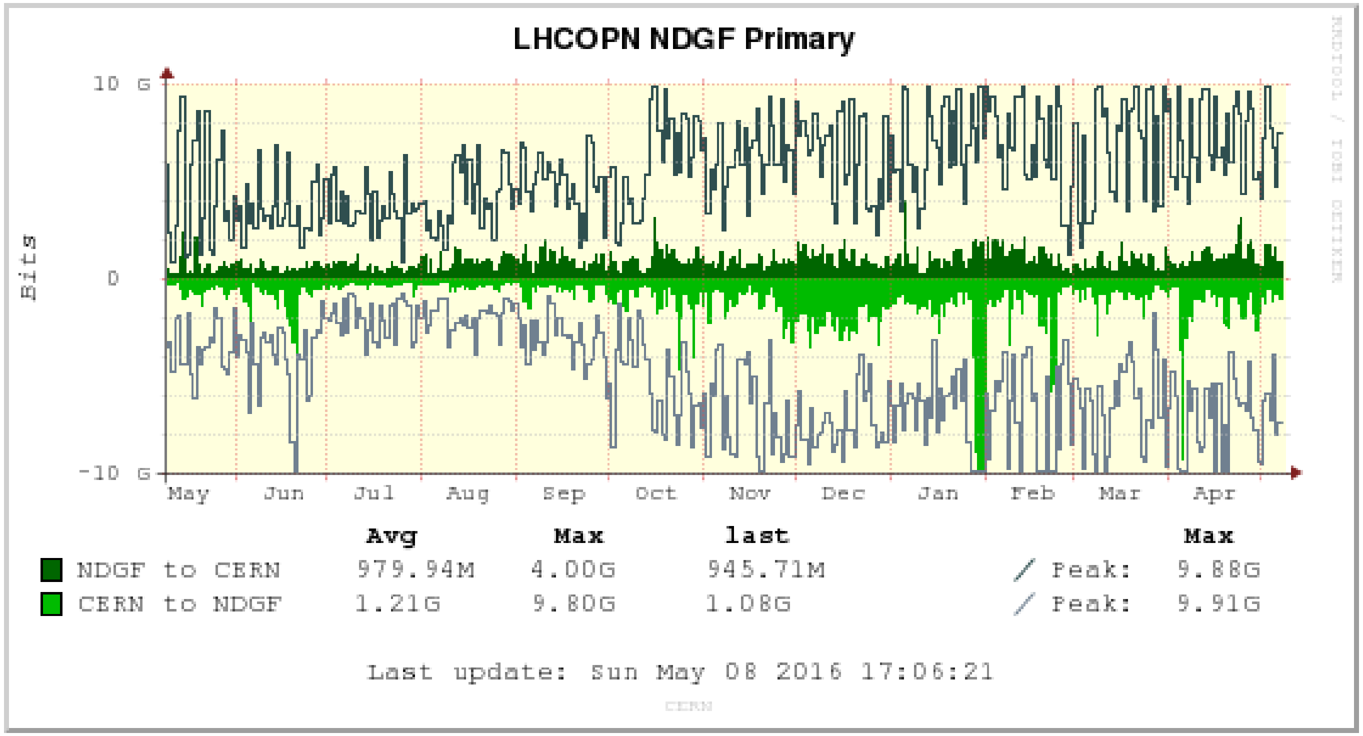
- In production since 2006
  - WLCG MoU signed by 4 Nordic countries
- Operated by the Nordic e-Infrastructure Collaboration (NeIC)
  - Funding from the Nordic Council of Ministers via NordForsk
- Staff: ~5 FTE (funded by NeIC), external experts: ~7 FTE (funded by national research programmes)

NDGF-T1 is a Tier-1 for ALICE and ATLAS



- NDGF-T1 targets:
  - **ALICE**: 9% of all Tier-1 needs
  - **ATLAS**: 6% of all Tier-1 needs
- Internally the targets are split between Denmark, Finland, Norway and Sweden proportionally to the authors count
  - Thus no ATLAS resources in Finland

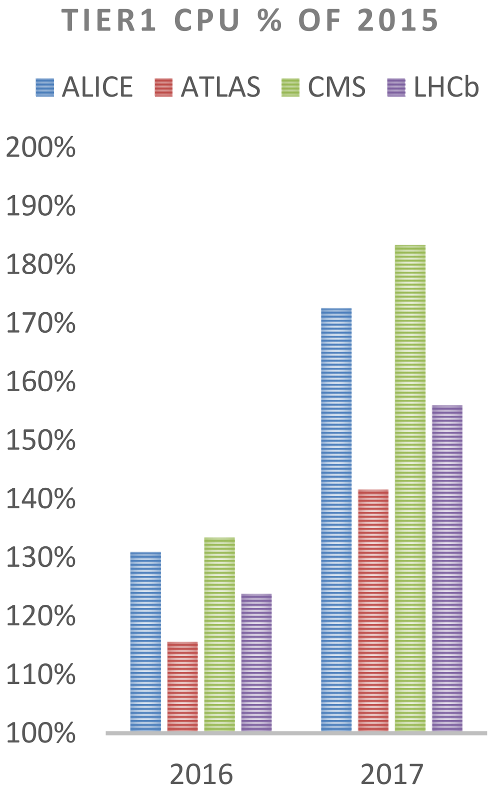
NDGF-T1 network load (LHCOPN)

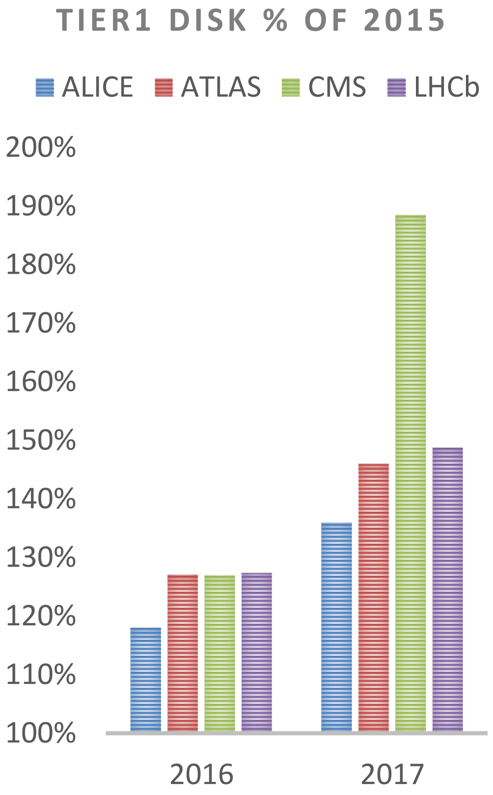


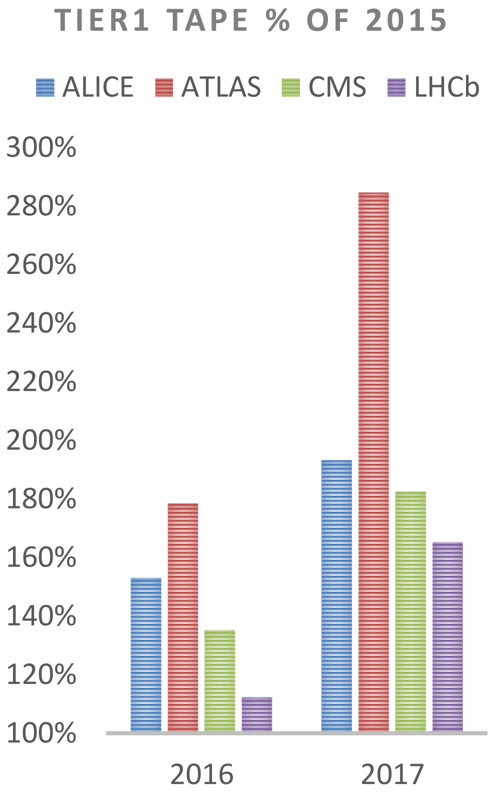
	Avg	Max	last		Max
🟩 NDGF to CERN	979.94M	4.00G	945.71M	/ Peak:	9.88G
🟩 CERN to NDGF	1.21G	9.80G	1.08G	/ Peak:	9.91G

Last update: Sun May 08 2016 17:06:21

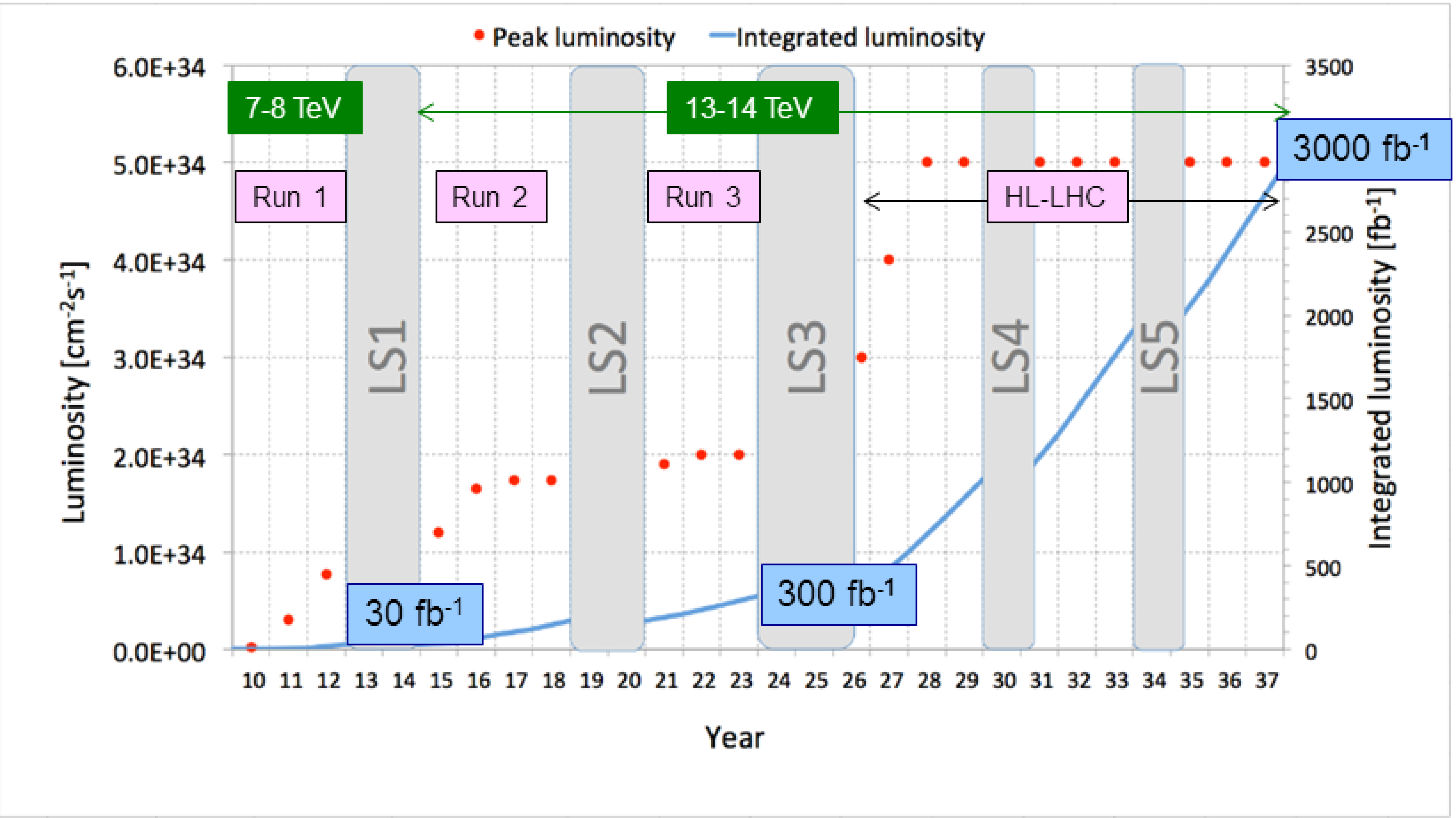
Growing LHC experiments requirements, % of 2015


TIER1 CPU % OF 2015


TIER1 DISK % OF 2015


TIER1 TAPE % OF 2015

Future LHC upgrades: much more data



Source: CERN

Moore’s law no more

- LHC experiments so far relied on Moore’s law
  - Assumed that

$$N_{data}(t) \times \$_{proc}(t) = Const$$

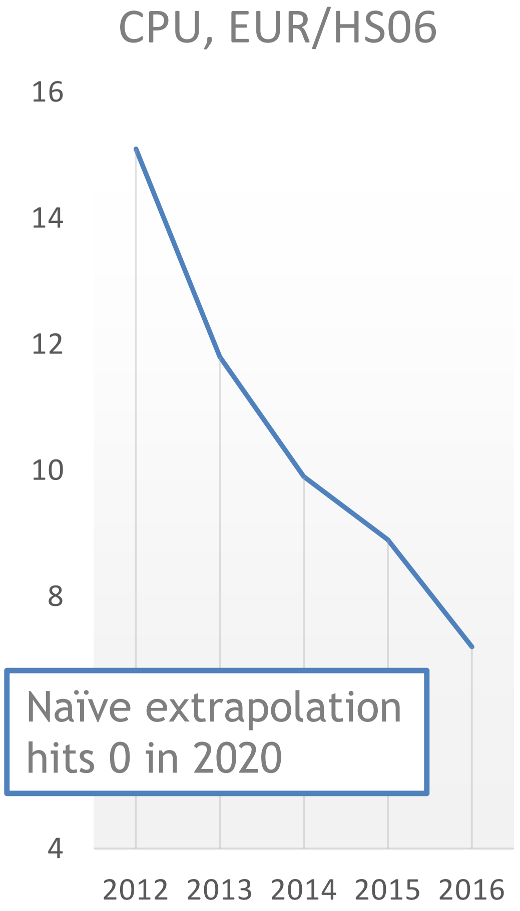
where N_{data} is data volume in bytes (growing), and $\$_{proc}$ is the cost of data storage and processing, in USD/byte (decreasing)

- In reality,

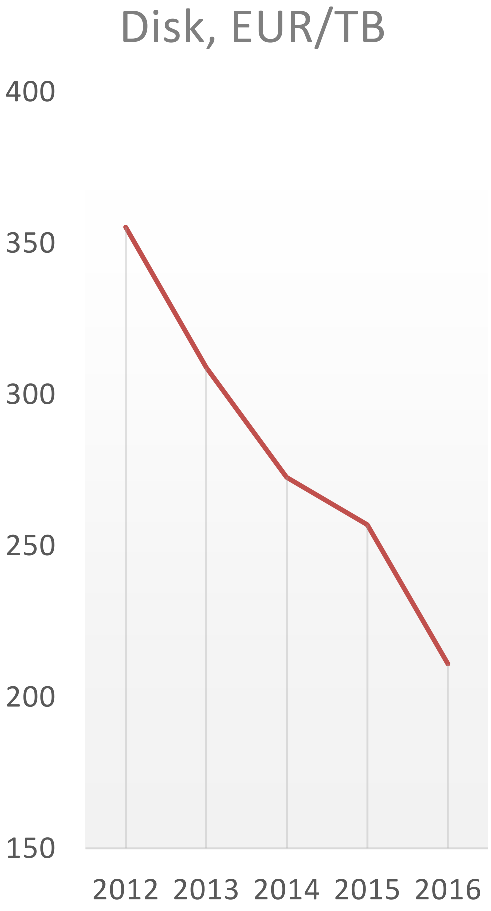
$$N_{data}(t) \times \$_{proc}(t) = F(t)$$

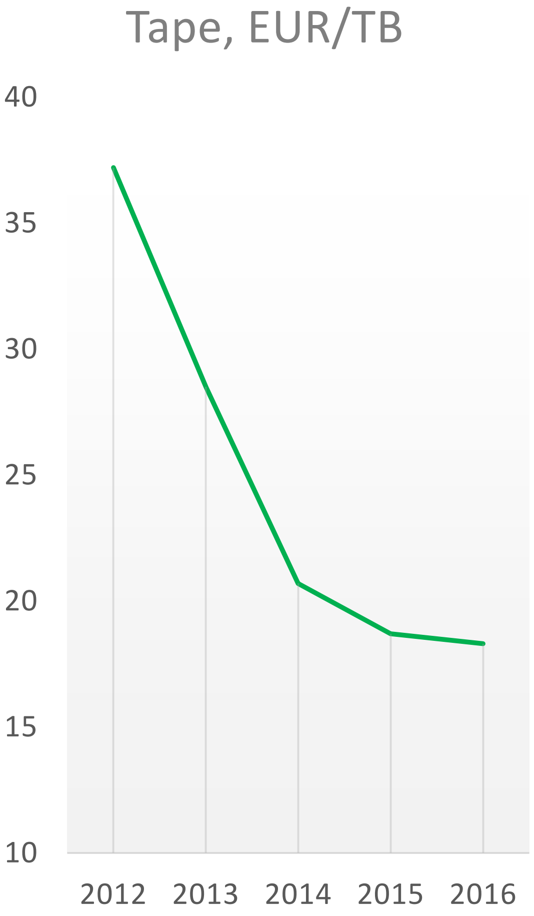
where F(t) is a growing function

Hardware costs evolution: CERN assessment (B. Panzer-Steindel)

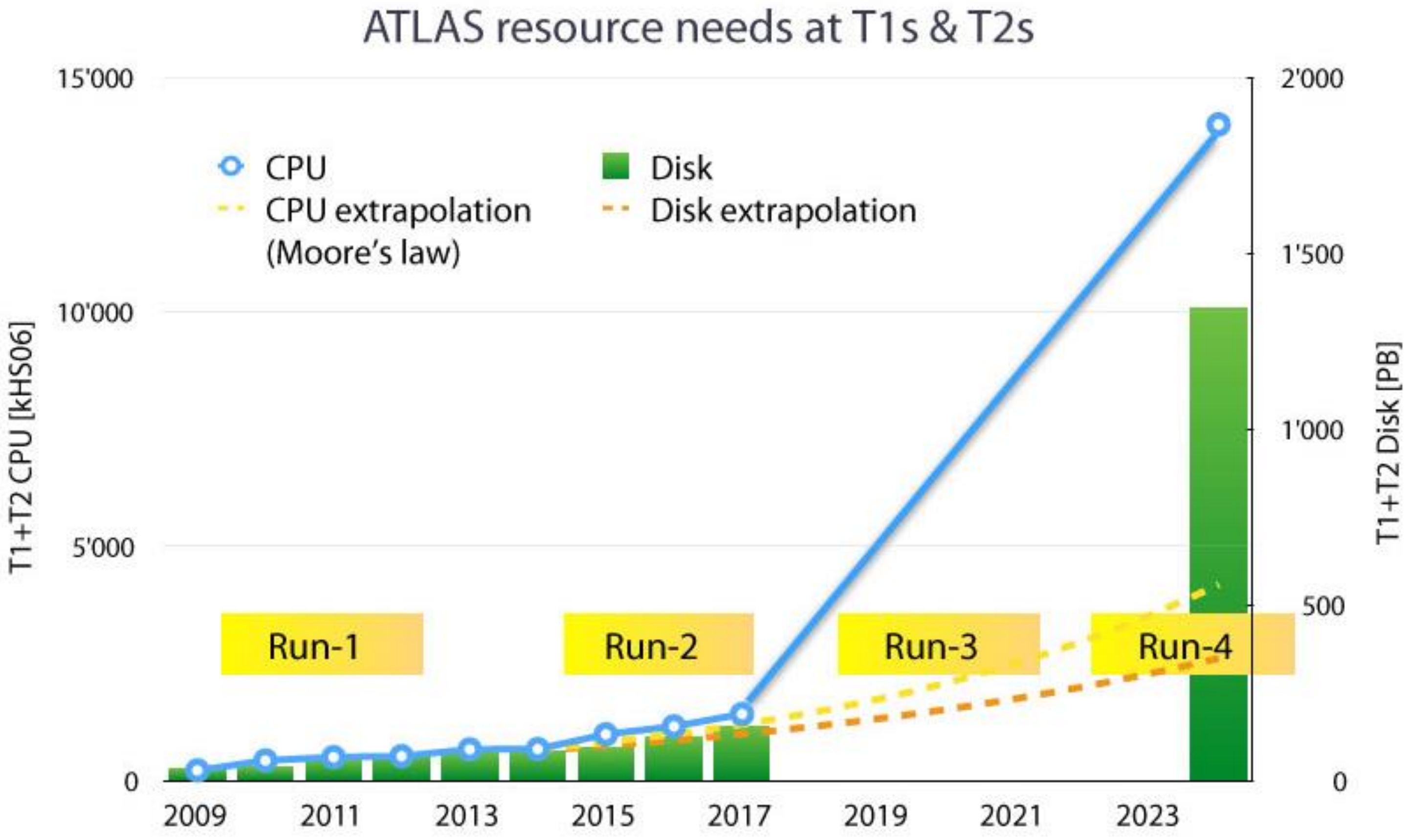

CPU, EUR/HS06

Naïve extrapolation hits 0 in 2020


Disk, EUR/TB


Tape, EUR/TB

ATLAS requirements vs Moore’s law



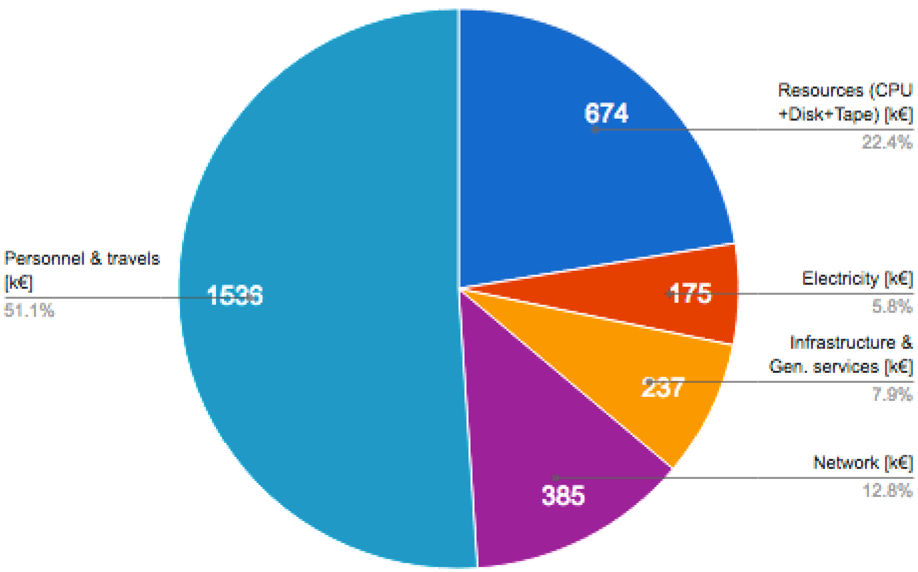
Source: E. Lancon/ATLAS

Data volumes are growing much faster than Moore’s law accommodates!

Other costs: NeIC investigation

- NeIC commissioned a study of NDGF-T1 costs and possible savings
  - Full report will soon be made public
  - Many thanks to Josep Flix for the great study!

NDGF-T1 costs: categories (2015)

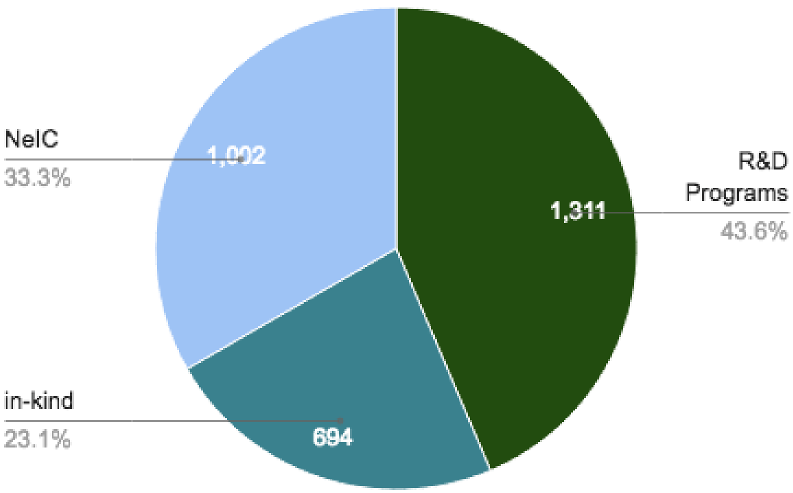


Hardware is just 22% of the Tier1 costs

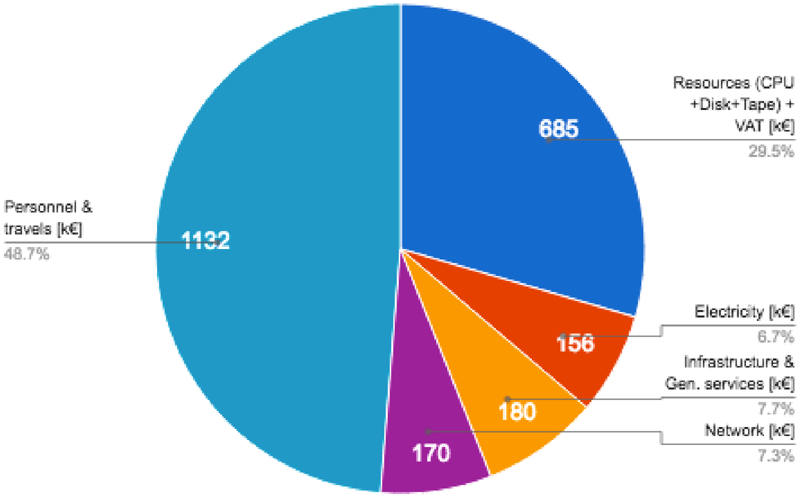
Single-site model (from J. Flix report)

- NDGF-T1 is distributed, which creates some logistical overheads
- Single-site model reduces overheads, but also reduces in-kind contributions
  - The net change is negligible; manpower is still the dominant cost

NDGF-T1 costs: funding origins (2015)



NDGF-T1 single-site costs: categories (2015)



If we can not buy (much) more hardware, what can we do?

- In computing:
  - Make use of all possible kind of resources: grid, cloud, HPC, volunteer computing, anything
  - Re-consider the role of a Compute Element (lost most functionality to pilot jobs)
  - Re-consider the role of traditional batch systems
  - Consider volunteer computing as a technology for small-size resources, reduce number of small Tier-2s
  - Seriously consider provisioning of resources via clouds, including commercial ones
  - etc

… what can we do

- In data handling:
  - Limit lifetime of replicated data
  - Reduce disk copies, increase tape storage
  - Re-consider roles of Tiers in data hierarchy, consolidate functions
  - Data streaming, stage on-demand
  - Federated storage
  - Learn from Big Data technologies (cloud storage, file systems, databases etc)
  - etc

… what can we do

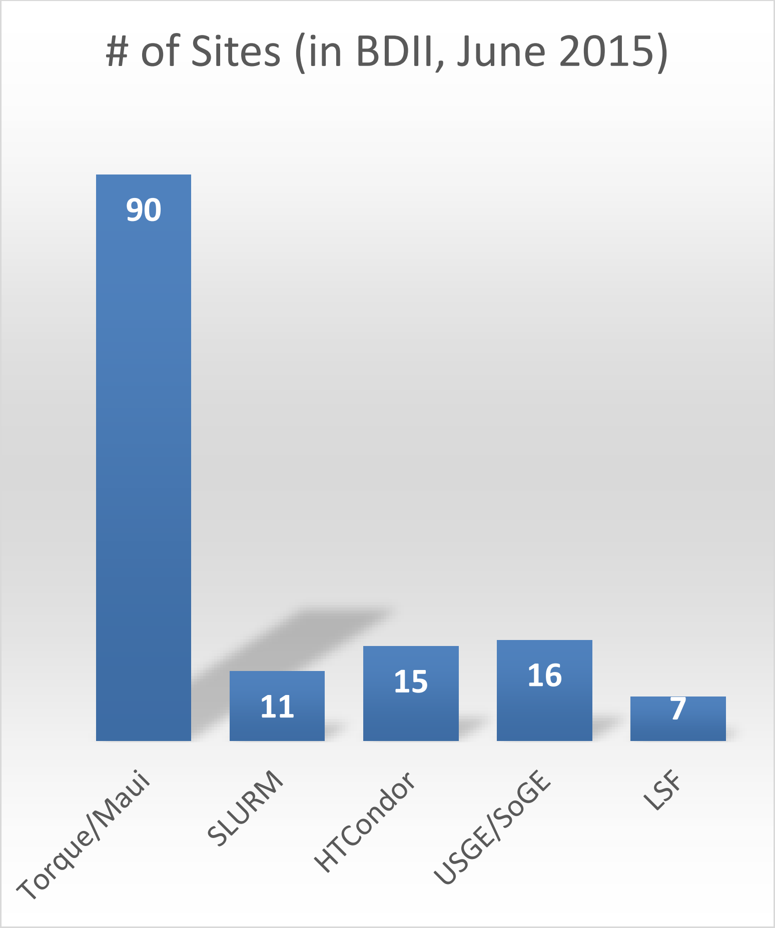
- In algorithms:
  - Optimize all steps of data processing, including simulation, reconstruction, analysis etc
  - Make maximum usage of parallelisation (for modern processor architectures)
  - Optimize I/O
  - Make use of new processing methods, e.g., Machine Learning
  - etc

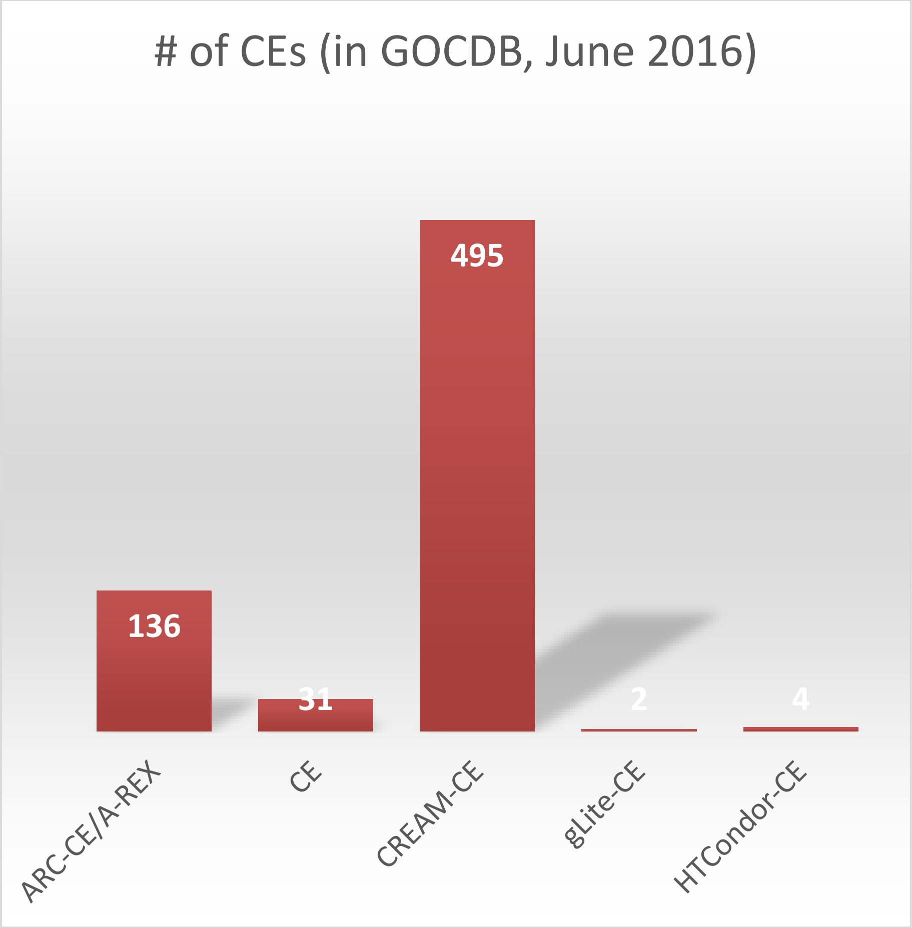
- While these tasks lie mostly with the LHC experiments, Tier-1 experts can provide valuable input and test environment

Grid or not?

- Usage of HPC resources in WLCG is increasing (TITAN, Tianhe, Archer, SuperMUC etc)
- There are also many successful examples of cloud resource integration
  - For example, Finnish CMS Tier-2 is entirely in a cloud
- Volunteer computing take-up is truly impressive (e.g. ATLAS@Home via BOINC)
- And yet, ~90% of LHC computing is still on the Grid
  - Actually, most “opportunistic” resources above are used via Grid interfaces

Grid Clusters and Compute Elements in WLCG





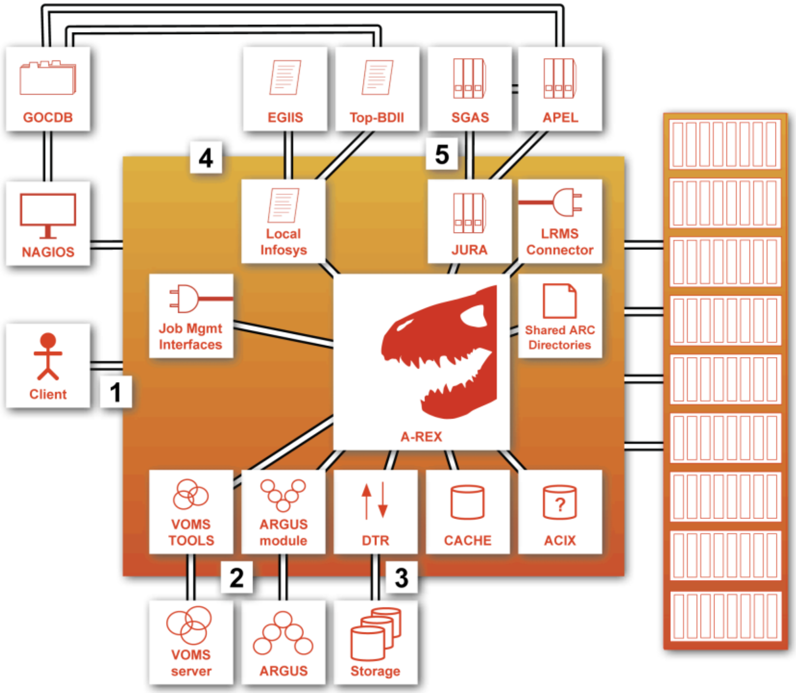
* ~15% of all CEs are not in production

** There are more HTCondor-CEs in the USA

NDGF-T1 uses ARC-CE and dCache

- ARC-CE:
  - Designed for time-shared HPC resources
  - Interfaces not just to Grid and HPC resources, but also to BOINC for volunteer computing
  - Also used as a front-end to an HTCondor-based virtual Tier2 in a cloud
  - Works with Amazon S3
  - Supports all relevant batch systems
  - Supports data caching – can thus be used at diskless sites
- dCache:
  - Scalable solution supporting disk and tape storage
  - Supports all the necessary protocols and IpV6
  - Implements a storage federation
  - Allows for transparent consolidation of storage resources

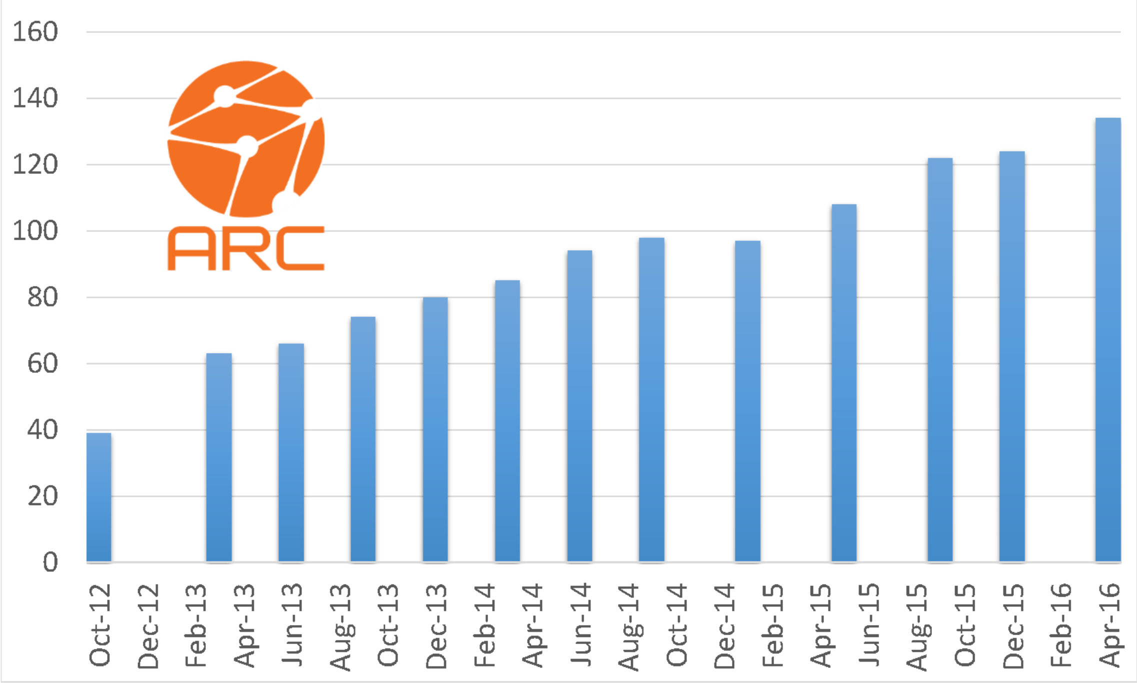
ARC Compute Element



1. Job submission (brokering based on info from GIIS, Local Infosys and ACIX)
2. Check credentials (VOMS, ARGUS, etc.)
3. Data staging from/to external storage
4. Registration to information indices (EGIIS); serving information requests of global aggregators (Top-BDII)
5. JURA parses job logs, prepares and sends job usage records to either SGAS or APEL accounting databases

- Development driven by NDGF-T1 needs
  - Common interface to shared supercomputer-class facilities
- First release: April 2004
- Current release: 15.03u8 (tagged on Friday)
  - 43 releases so far
- 100 000+ lines of code
  - Free, Open Source Apache v2.0 license
- 44 past and present contributors from all over the world
- Coordinated by the **NorduGrid Collaboration**

ARC usage: a growing trend



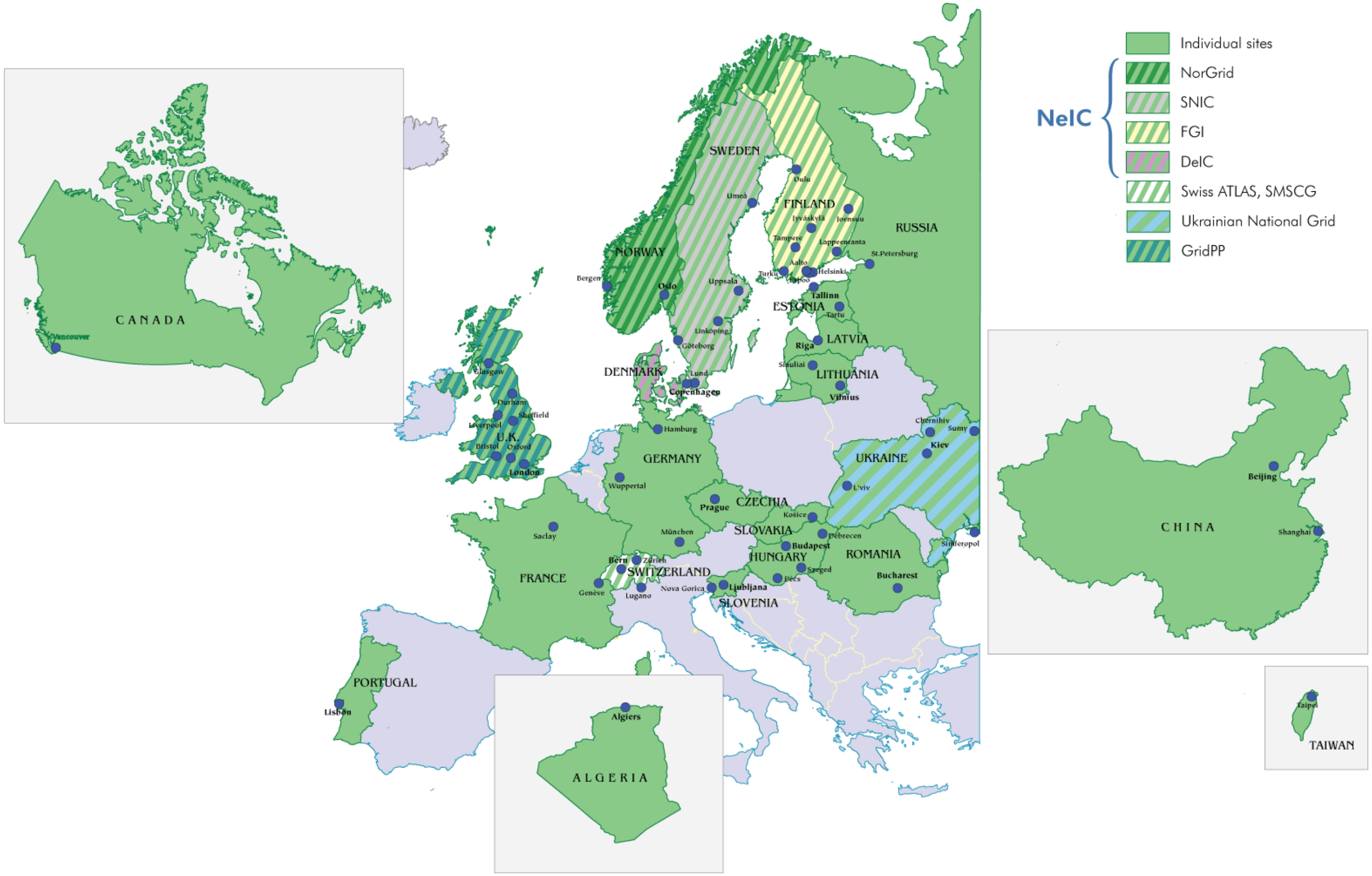
- ARC usage in WLCG:
  - In 2010: 10 sites, all in Nordic countries
  - In 2016: 136 sites around the world (~20%)

LUND UNIVERSITY
Faculty of Science

norden
NordForsk

Nordic e-Infrastructure
Collaboration

ARC worldwide (as of May 2016)



Conclusion: Future tendencies

- Data distribution *a la MONARC* is all but gone
- ”Jobs to data” paradigm is being replaced by data streaming and caching
- Data federations increase network load even more
- Virtual centres using cloud resources become a reality (saves manpower)
- Consolidation of resources in large HPC centres is inevitable
- Grid technologies stay as interfaces to various resource types
- From experience, storage support requires much more effort than all other services, thus if we want to save costs, we need to find more robust storage solutions (learn from Google or Facebook?)