WLCG workshop 2017

Summary points (Early draft!)

See talks at https://indico.cern.ch/event/609911/timetable/#20170619.detailed

Introduction (I Bird)

- Run 2 in 2016 delivered 50 PB of new data, following exceptional performance of the LHC
  - Continued to set new performance records in all areas
- WLCG infrastructure continued to be even more active in the EYETS
- 2017/18 look to be challenging in terms of resource availability, esp if LHC meets expected luminosities, availability
- Scientific Computing Forum https://indico.cern.ch/category/9249 (2 meetings)
- Activity (& engagement) is ramping up to look at evolution of the computing models for the future
- Goal to have a Community White Paper (CWP) on overall strategy & roadmap for software/computing for HL-LHC

Strategy Document in 2017

- Describe the HL-LHC computing challenge given what we currently understand
  - Running conditions, trigger rates, event complexity, based on reasonable extrapolations of today's computing models
  - This will be a snapshot of a (yearly?) update of these numbers
- Describe the potential computing models and how they could change the cost and/or physics output
  - Necessarily at a high level
- Cost models
  - Appropriate metrics, balance/trade-off between CPU, storage, network etc
- State-of-the-art understanding of evolution of technology
  - 2-3 years is already difficult to predict; 10 years is impossible (even for the technology companies)
- Set out what we see as R&D areas, and potential prototyping activities or demonstrators:
  - Goals, metrics, resources, plans
- The HSF CWP will provide the basis of this

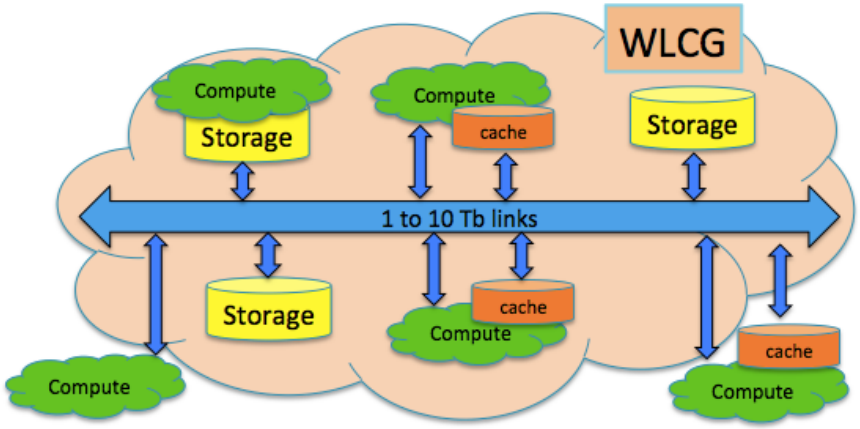
European SKA RC Compute Model (A Scaife)

- Background
- Example workflows
- Opportunities for collaboration:
  - Frameworks for incorporating opportunistic & heterogeneous resources;
  - Middleware;
  - Work flow management;
  - Persistence management for secondary data products;
  - Data transfer services

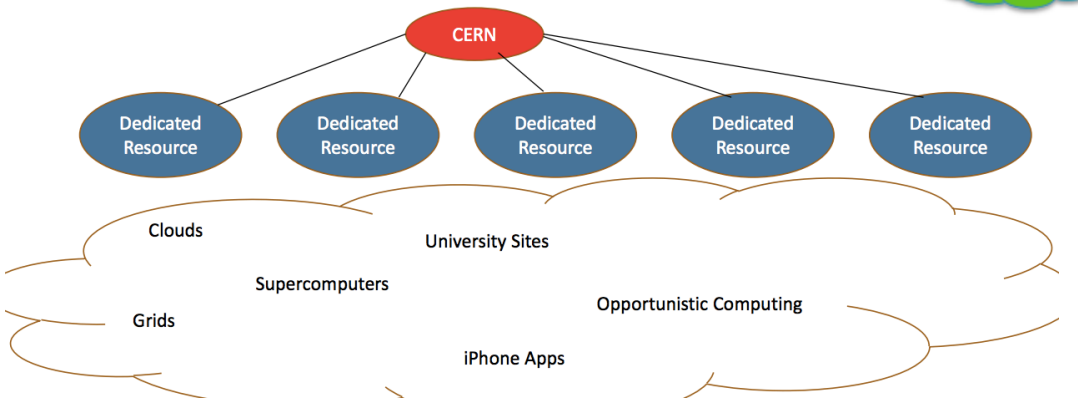
Evolution of the WLCG infrastructure (S Campana)

Evolution in the direction of

- Network centric model
- Consolidation of storage
- Diversification of facilities
- …



WLCG at HL-LHC



WLCG at HL-LHC (I. Fisk's representation)

- … diversification of compute resources

No need to wait 2026 for this

Lightweight sites compute – possible models (M Litmaath)

- T2 vs T3
- Classic view (storage / computing)
- Alternate: No need for CE+batch.
- Distributed site operations
- Volunteer
- SLAs – make resources lightweight.

Singularity introduction and use in OSG (B Bockelman)

- Singularity is another container technology in our toolbox.
- •
- Different set of tradeoffs than Docker:
- •
- I.e.,
- setuid
- binary but no system service.
- •
- Currently, most popular where HTCondor runs as non-root.
- •
- Interface will be a work-in-progress during 2017.
- Currently,
- completely managed/implemented by sysadmin
- .
- •
- CMS and OSG utilize Singularity as a mechanism for
- isolation
- and
- OS portability
- .

Use of containers at RAL Tier-1 (A Lahiff)

- Containers are being used a lot at RAL in production
- –
- migrated our HTCondor batch system to run all jobs in
- Docker containers
- –
- have started rolling out xrootd gateways to Ceph in
- containers on worker nodes
- •
- Other efforts at RAL involving containers
- –
- providing more flexible computing infrastructure
- –
- making it easier to use public clou

Singularity (V Brillault)

- No slides shared

Collaboration between security teams (R Wartel)

- No slides shared

Ongoing work (L Valsan)

- No slides shared

WLCG Data Steering Group

- Set up group to define long term WLCG strategy for DM and track relevant technical themes
- Group met 4 times: https://twiki.cern.ch/twiki/bin/view/LCG/WLCGDataSteeringGroup
- Areas
  - Infrastructure diversity (inc. colocation other communities)
  - Common global strategy (common vision?)
  - Quality of Service
  - Infrastructure architecture (multi-site storage, federations, caches…)
  - Operations (storage accounting, monitoring. ….)
  - Enable future analysis
  - AAI & federated identity
  - Priorities (archive storage implementations, diversity).
- Next: Meetings on priorities. Cross-cutting discussions. Exchange forum?

Storage resource reporting proposal

- 5 proposed requirements
  - All services (total used space); storage − total used and total free space; provide a structured file holding space info; provide subdir resource reporting; storage dumps
- Principal remaining discussion points
  - “json file”
  - Storage dumps
- Full text
  - https://docs.google.com/document/d/1yzCvKpxsbcQC5K9MyvXc-vBF1HGPBk4vhjw3MEXoXf8/edit#

Object stores

- Description: something that stores objects. An object is data and metadata.
- Clever data placement algorithm

- Commercial clouds − complicated cost model. Protocols.
- Amazon S3 | OpenStack swift API
- Ceph: Object and block storage.
- CephFS − adds Meta Data Server
- Most replicate data. Erasure Coding (EC) can be seen as an extension to RAID.
- Current WLCG computing models work very well with Object Stores

SKA data products (R Bolton)

- SRC and SRC-Alliance Requirements and Goals. Produce and archive advanced data products from SDP data products
- Relationships
- Image cube examples
- Correlated visibility samples

Multi-site storage with dcache (T Mkrtchyan)

- Components: Door; Pool; Namespace; Poolmanager
- Multi-site deployment
- Network: CELL messages (inter componenet communication); ZooKeeper (service discovery)
- Fault tolerance; upgrades…

- dCache has a long tradition in providing
- federated storage for WLCG
- ●
- The configuration flexibility allows to control
- data placement and replication
- ●
- Fault-tolerant setup is recommended for a
- distributed deployment
- ●
- We solve technical issues, sites have to

Data caches on the OSG

- How big should the cache be?
- What is the minimum perfornance ?
- How should caches be authenticated?
- Building on StashCache. Have extended CVMFS to read metadata.
- USCMS large-scale proxy cache setups

- We have two interesting demos centering around the Xrootd caching proxy:
- •
- CMS is probing scale
- in total volume and integration with complex job infrastructure.
- •
- OSG is probing CVMFS-based access and multi-VO authentication setups (maybe non-
- GSI-based in the near future?). Smaller overall working set size. Partnering with LIGO.
- •
- Is this accessible for a generic “small site” with a few tens of TB?
- •
- Unclear for CMS: can efficiently access via remote IO.
- •
- Majority of our production workflows are streaming-based (not cache friendly).
- •
- Non-streaming-based workflows are very IO-intensive: not great for “small sites.”
- •
- Analysis is a good use case (repeat use, moderate IO requirements), but requires
- moderately-large working set size (few hundred TB).
- •
- Tests by the UCSD / Caltech team will provide illumination this year

Regional (HTTP/WebDAV) federations (R Sobie)

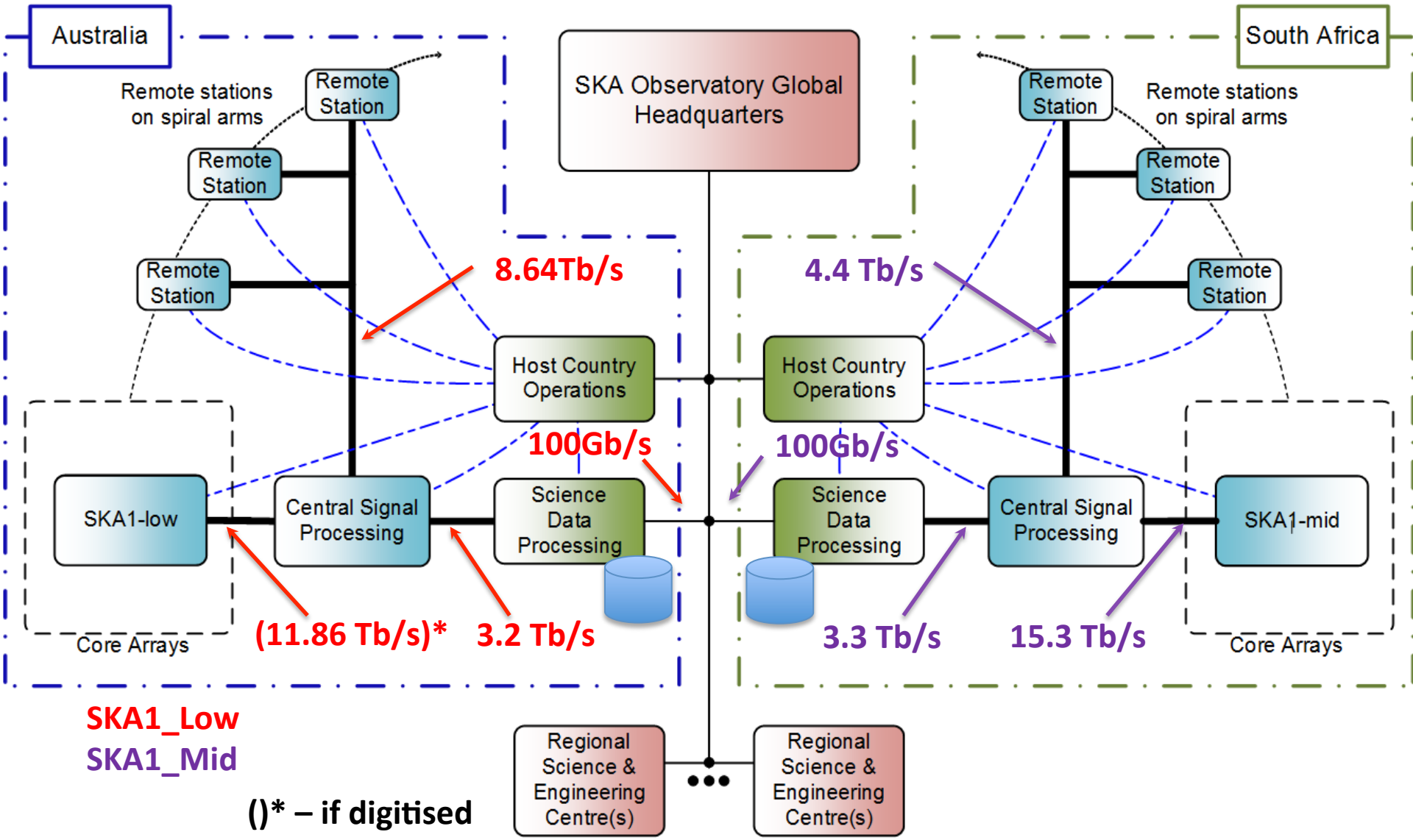
- Multiple development
- instances of
- DynaFed
- –
- DynaFed
- in front of
- empty object storage
- : CERN
- , RAL, INFN, UVIC
- –
- DynaFed federating existing/full storage:
- INFN, UVIC
- •
- Object or conventional storage
- –
- Accessed using WebDAV
- •
- Plan to start running ATLAS test jobs (summer 2017)

Transition to diskless configuration for RO-16 & R)-14 (V Emmanouil)

- RO-04 and RO-16 diskless configuration is functional and stable
- The amount of processed data from RO-16 and RO-16 with respect to the total processed data (RO-
- 07 + RO-16) is very small ~(1TB)
- We will not expect network issues between RO-07 and RO-16 because of diskless configuration,
- current network traffic is too low
- The
- remote transactions (due to RO-14 and RO-16) on RO-07's DPM are order of 21.5%
- We should follow this utilization in the case that we would like to increase the current CPU
- capacity (local and/or remote) on those sites.
- The contribution of RO-14 and RO-16 correspond up to 30% MC simulation activity amogst the
- four 4 sites in Romania.
- The solution can be improved with the use of ARC-CE
- Cache capabilities and control the concurrent number of data transfer connections

Data Transport Requirements SKA

Areas of overlap: Data Moving tools; Replica managers and Data placement; Storage systems; Traffic flows between Data Centres; Federated AAI and user access



ATLAS Network Requirements

- Moving to a Nucleus and Satellite Model.
- Designation depends on capacity and throughput.

Nucleus	Now	5 year (2022)	10 year (2027)
Storage Capacity (PB)	2	5	12.5
Total CPU (kHS06)	40	100	250
LAN (Gb/s)	40	200	1000
WAN (Gb/s)	20	60	200

Disk-less	Now	5 year (2022)	10 year (2027)
Total CPU (kHS06)	20	50	125
WAN (Gb/s)	4	20	100

ALICE: WAN Numbers @T2s

- Total WAN traffic are 1/11 of LAN traffic
  - WAN total T2s = 650MB/sec => **12.5KB/sec/core** (52K cores)
- Our 'advisory rule' for T2s is **"100Mb WAN network per 1K cores"**
  - Usually the available bandwidth is superior to the above
  - ~Independent of T2 role – a diskless T2 has to export ~2x the produced data; T2 with an SE exports one copy and accepts copies from other centres.
- Importance of tuning. LHCOPN recommended.

CMS networking requirements

- LAN and WAN connectivity at the same level at many sites
- ♣
- US sites have typically 100Gbit/s
- ♣
- European sites (several) 10Gbit/s with a large spread
- >
- CMS schedules typically 10-15% of jobs with remote data access
- ♣
- The “pentaly” in CPU efficiency is a drop of ~10% for remote access
- ♣
- Large gain in flexibility
- >
- Computing model likely to evolve towards scenarios that require fast interconnets
- >
- Need to improve CMS transfer system and scheduling system to better exploid
- network metrics

WLCG Network Throughput WG (S McKee)

- WG has established a working infrastructure to monitor and
- measure our networks
- –
- Proven record on fixing existing network problems and improving
- transfer efficiency
- –
- Stable production infrastructure (issues still require follow
- -
- up)
- •
- Mid
- -
- term evolution topics to discuss
- –
- Network capacities planning
- –
- needed ?
- –
- Network utilization monitoring
- •
- Both site
- -
- level and WAN (REN)
- –
- Evolving and integration of monitoring data
- •
- New sources, dashboards and network stream
- –
- Network Analytics
- •
- Alerting/Notifications and anomaly detection
- –
- SDN Networking demonstrators and testbeds
- •
- Testing new technologies such as
- OpenVSwitch
- , OVN,
- OpenDaylight,etc
- .

Cost model

- We need a
- model
- to play with - know the impact on costs:
- →
- Efect if disk from T2s is moved to T1s? Or disk federatons?
- →
- How the costs can be improved by evolving WLCG (HL-LHC era)
- - This could be useful for
- VOs
- to know where to put eforts:
- →
- quantfy the advantages or disadvantages of e.g remote I/O vs local processing,
- memory vs cores, produce/store or reproduce data, fast sim vs full sim, ...
- →
- Cost efects on whatever
- sw
- improvement (investment cost vs real savings)
- →
- Cost of opportunistc resources inclusion into the system
- →
- Running workfows where they are most efcient (in terms of cost)
- →
- Evaluatng the cost of the inefciencies in the system. Where to focus?
- →
- Cost of the network → remote reads and data transfers costs
- - This could be useful for
- FAs/sites
- to be cost-efectve

Efficiency and costs (J Elmsheuser)

- Known bottlenecks
- Batch on EOS (BEER?). Use under-utilised EOS nodes for CPU payload processing
- ATLAS computing workflow performance understading meetings and analysis. Look at memory consumption, i/o etc. for every job.
- Geant4 − optimisation 1
- ATLAS reprocessing − job memory profile I. Merge over cores at end not efficient for node.
- CMS: Multi-threaded applications.

- There are several examples in the experiments grid systems or
- experiments payloads that could improve the efficiencies by a few
- percents each
- •
- But it take sometimes quite a long time to put them into production
- either due to physics constraints/priorities or since experts have to
- carefully maintain current systems

LHCb Recent Efforts on Compute and Software Optimization (S Rosier)

- "Fast stop"
- Code vectorization
  - Special library. Optimal speed up x8 for this sub-detector.
  - LHCb software stack buid for SSE 3 and 4.2. Soon only 4.2.
  - If providing WNs thrhgouh virtualisation need to export the CPU capabilities correctly
  - Training of the collaboration

Sharing Information, Dissemination and Training (M Schulz)

- Vectorisation, memory management, efficient numerical code, advanced features of C++…. Computing on FPGAs.
- Knowledge is dispersed.
- Tribes − fragmentation
- Vision: e.g. QA for information like stackoverflow.com.
- How?

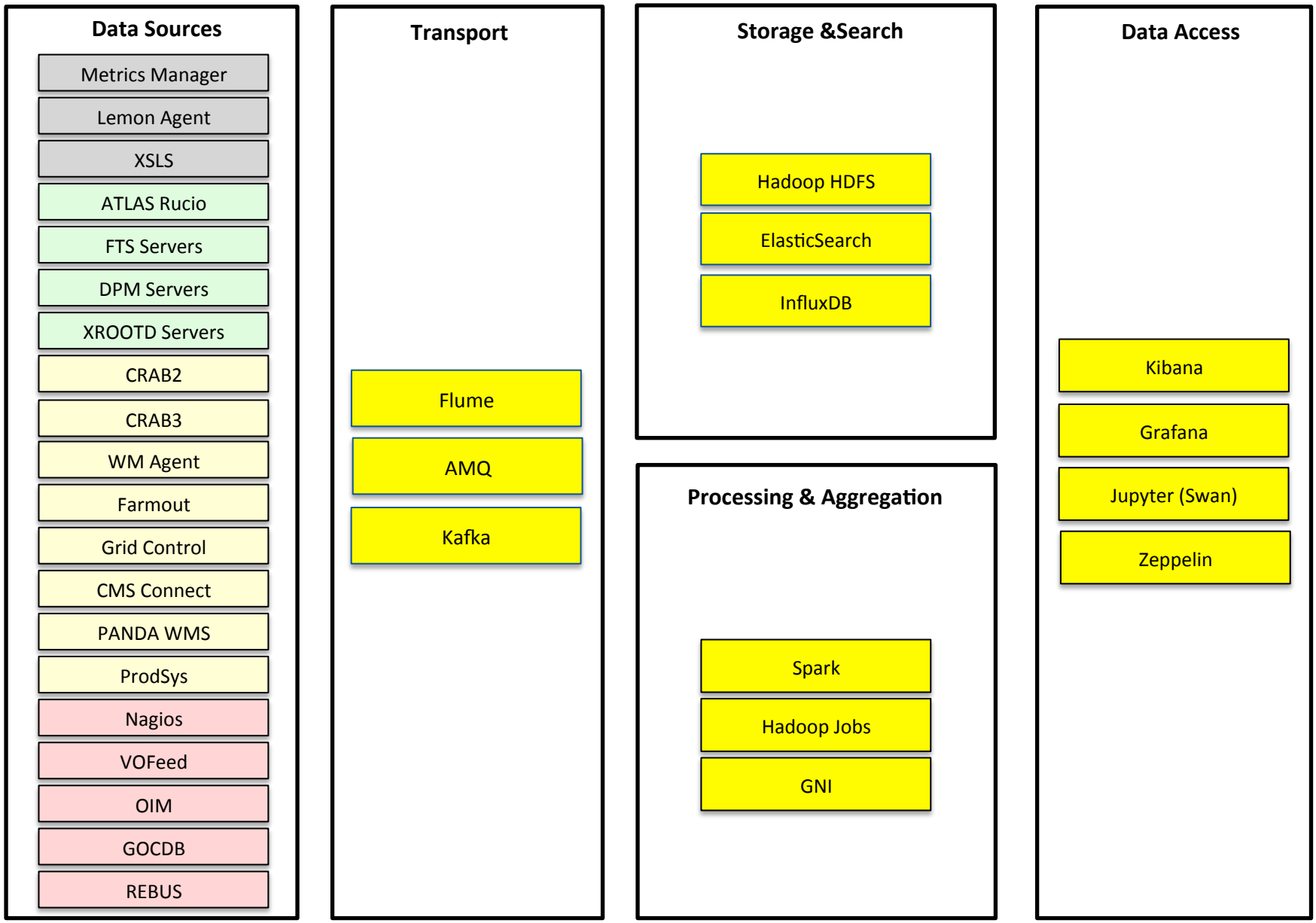
Discussion with astronomers

- A brief history of ….

Benchmarking WG report (D Giordano)

- HS06
- –
- Preliminary study still shows good agreement among HS06 and
- CMS MC
- ttbar
- ,
- when server fully
- loaded
- –
- Passive benchmarking
- •
- Discrepancies among HS06 and ATLAS
- reco
- jobs are within 10
- %
- –
- Need to better understand the reasons of the discrepancies for
- LHCb
- and ALICE
- SPEC2017 is now available:
- should start
- testing it
- Work
- in progress to setup a testbed for the HS06 successor
- –
- Support from the Experiments is mandatory here

Status and plans of WLCG Unified Monitoring (A Aimar)



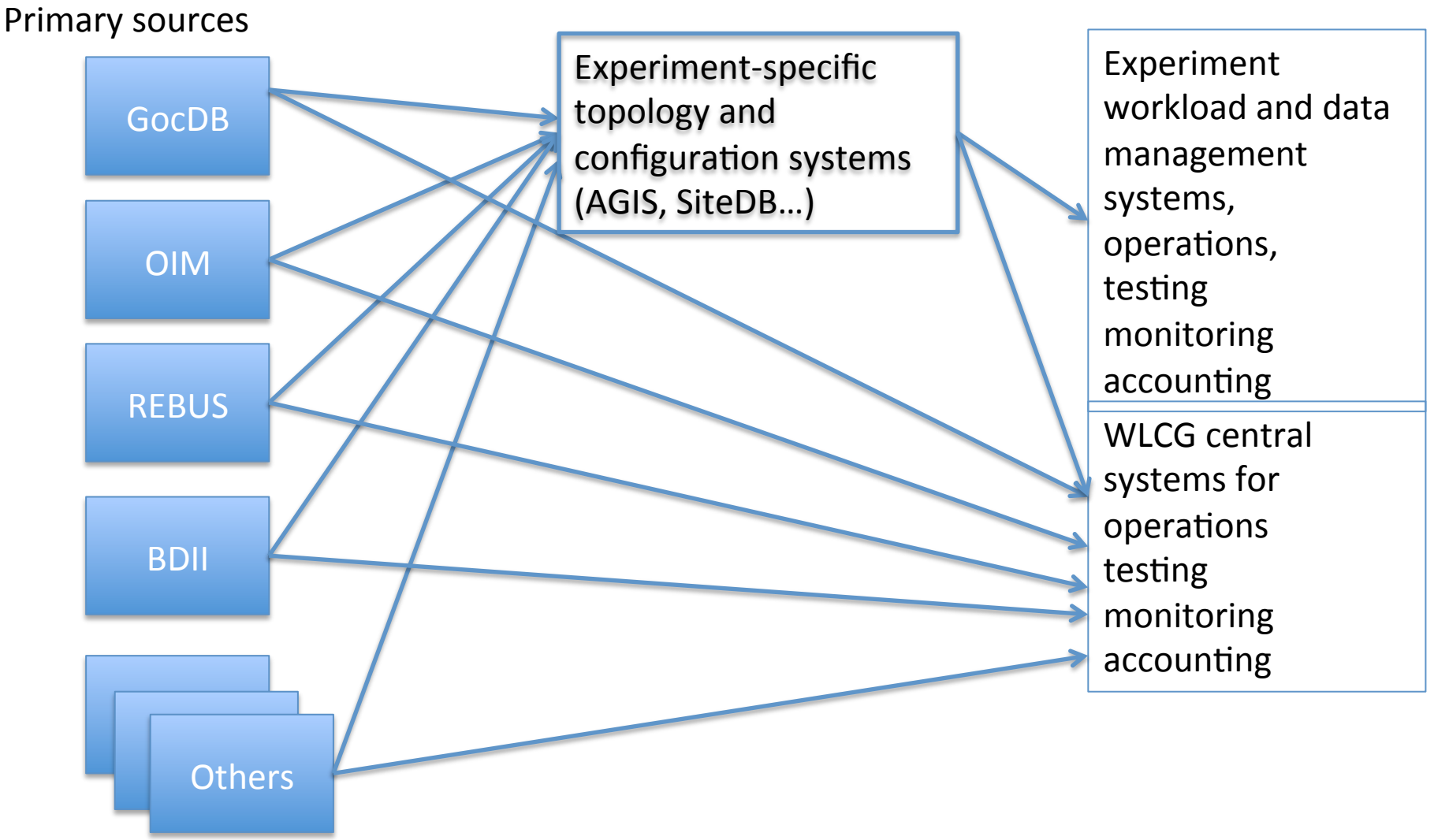
Elastic Search service at CERN (U Schwickerath)

- Running a centralised Elasticsearch service at CERN
- ●
- Support 2.X and 5.X versions
- –
- Moving to 5.X
- –
- Index level security for 5.X Elasticsearch
- ●
- Lessons learned
- –
- Very different use cases and requirements
- –
- Careful tunings are needed on
- both
- client
- and service side
- ●
- Contact:
- elasticsearch-support@cern.ch

Monitoring in experiments (A Aimar)

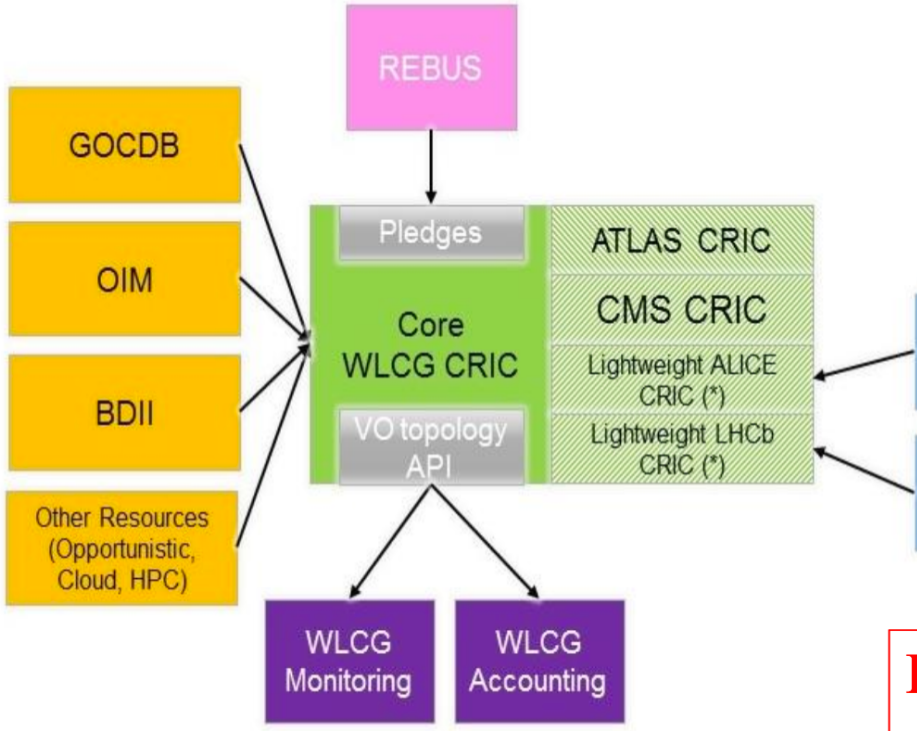
- All looking at external technologies instead of fully in-house solutions
  - ElasticSearch, Kibana, Kafka, Messaging, Spark, HDFS
  - less custom-made less control, but huge communities and free improvements
- ATLAS and CMS
  - have their infrastructure for deeper analytics studies
  - rely on MONIT for monitoring, and as migration from WLCG dashboards
  - use MONIT curated data with aggregation, enrichment, processing, buffering, dashboards, etc.
- ALICE and LHCb
  - run their own infrastructure
  - based on central IT services (ES, HDFS)
  - could share (some) data in MONIT, if needed

Information system evolution (J Andreeva)



CRIC (AD Girolamo)

Plugin based: Core and Experiments



(*) Maintained by WLCG to store very simple experiment topology information (i.e. experi

Core CRIC

- *Resources "Provided by"*
- Single entry point for WLCG topology and service configuration
- Consumes information from available information sources

Experiments CRICs

- *Resources "Used by"*
- Describes experiment topology
- Uses core CRIC and adds extra info needed by experiment

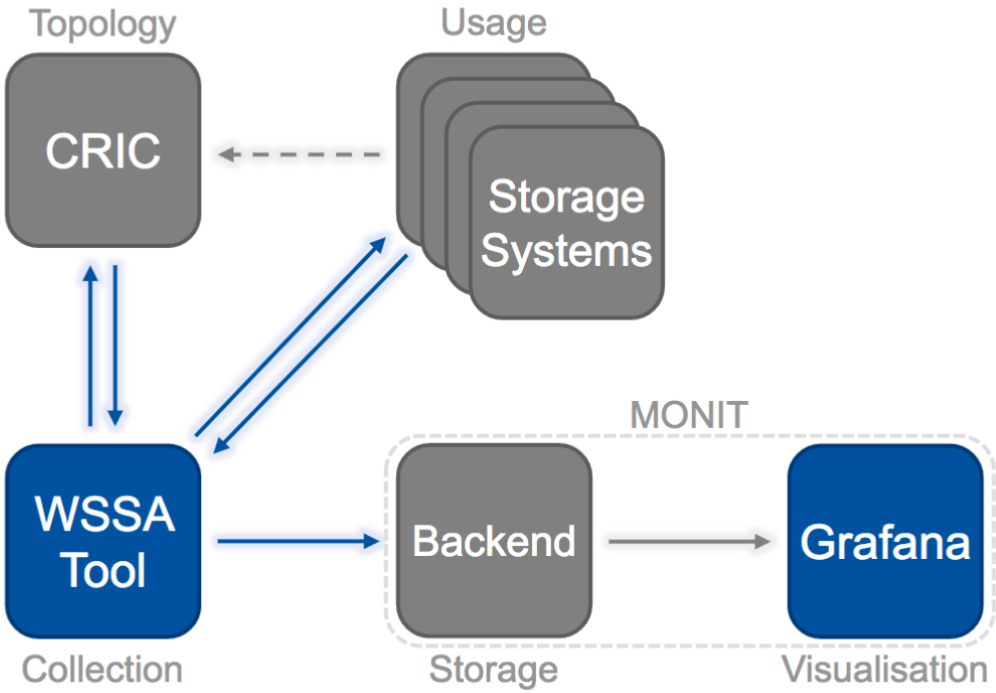
Storage Space Accounting (J Andreeva)

- Total used and total available for all distinct spaces available for the experiments

  (should be available by at least one non-SRM protocol)
- Number of files if possible but not strictly required
- Frequency not higher than once per 30 minutes
- Accuracy order of tens GB
- CLI or API for querying

Storage Space Accounting – prototype (D Christidis)

Architecture



IPv6 sessions

- WLCG and IPv6 (A Dewhurst)
- An Introduction to IPv6 (T Froy)
- Deploying IPv6 (F Prelz)
- DHCPv6 at CERN (E Martelli)

Thursday - exercises

- Data transfer
- perfSONAR
- Squif for Frontier & CVMFS