ATLAS Data Management: Evolution Challenges

Vincent Garonne

Kicktipp / ADC Euro2016

		Journées									Total		
Pos.	Nom	1	2	3	4	5	16	Qu	Se	Fi	Bon.	Victoires	Tot
1.	IPERLUCA	17	16	4							0	1,33	37
2.	Tomas.Kouba	23	12	2							0	1,00	37
3.	TheCrow	13	13	2							0		28
4.	Ivan	14	10	3							0		27
5.	vincent	11	8	3							0		22
6.	tartan_army	10	6	4							0	0,33	20
7.	tjavurek	7	7	2							0		16
8.	AleDiGGi	3	6	4							0	0,33	13

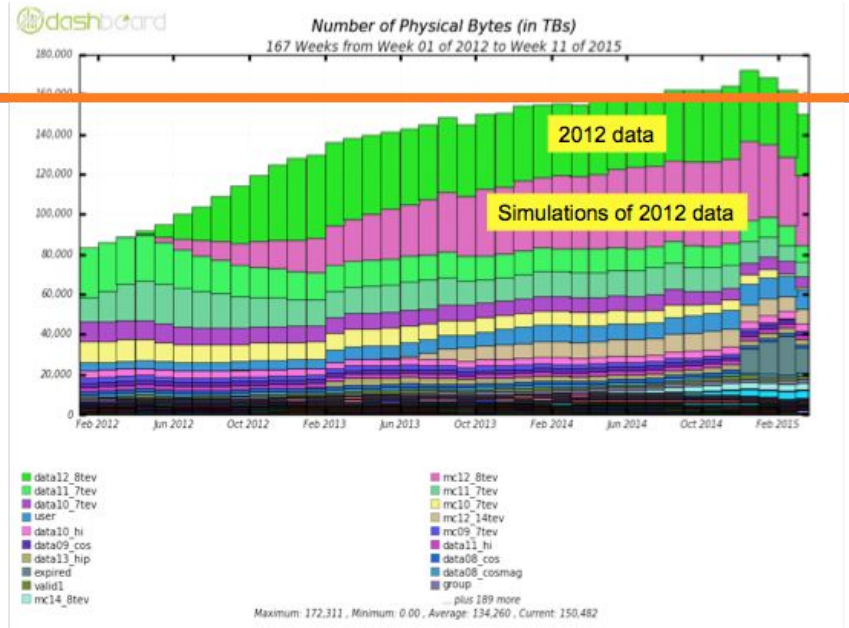
ATLAS DDM: Current numbers

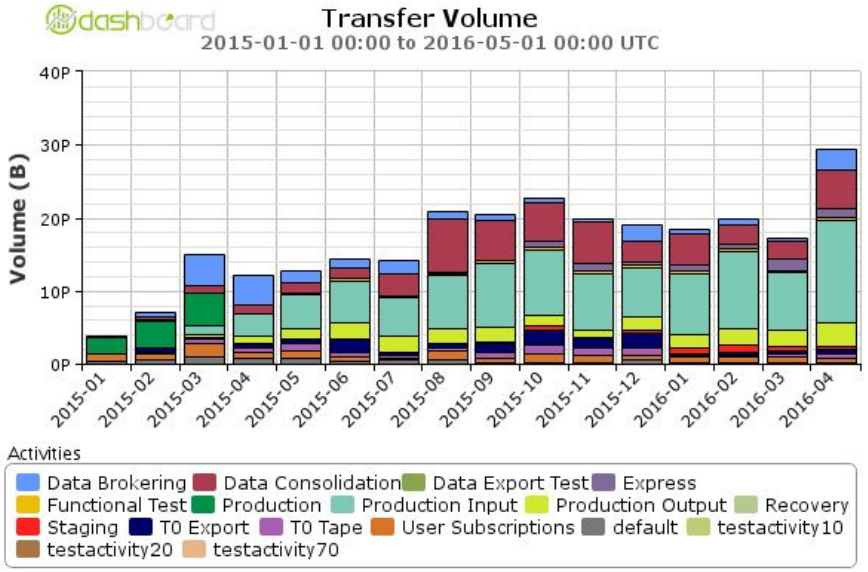
The ATLAS DDM System/Rucio has demonstrated very large scale data management:

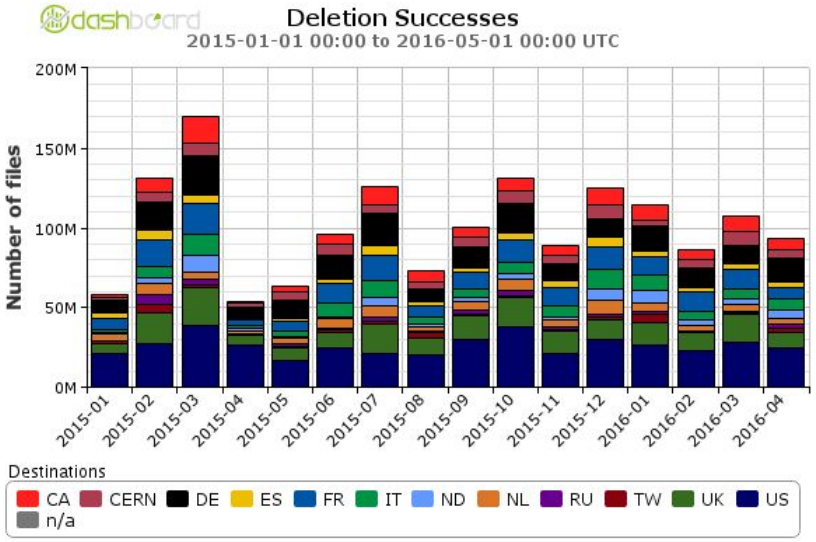
- 2000 ATLAS users
- 200 PB on 130 sites
- 1B file replicas
- 2.5 M file transfers/Day, 1.5 PB/Day
- 6 M deleted files/Day, 2 PB/Day
- 1M jobs/day

It's a lot of data !

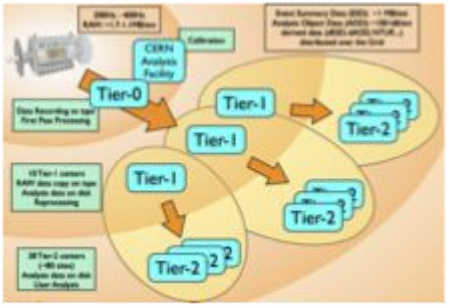
Average ATLAS traffic: 10GB/s





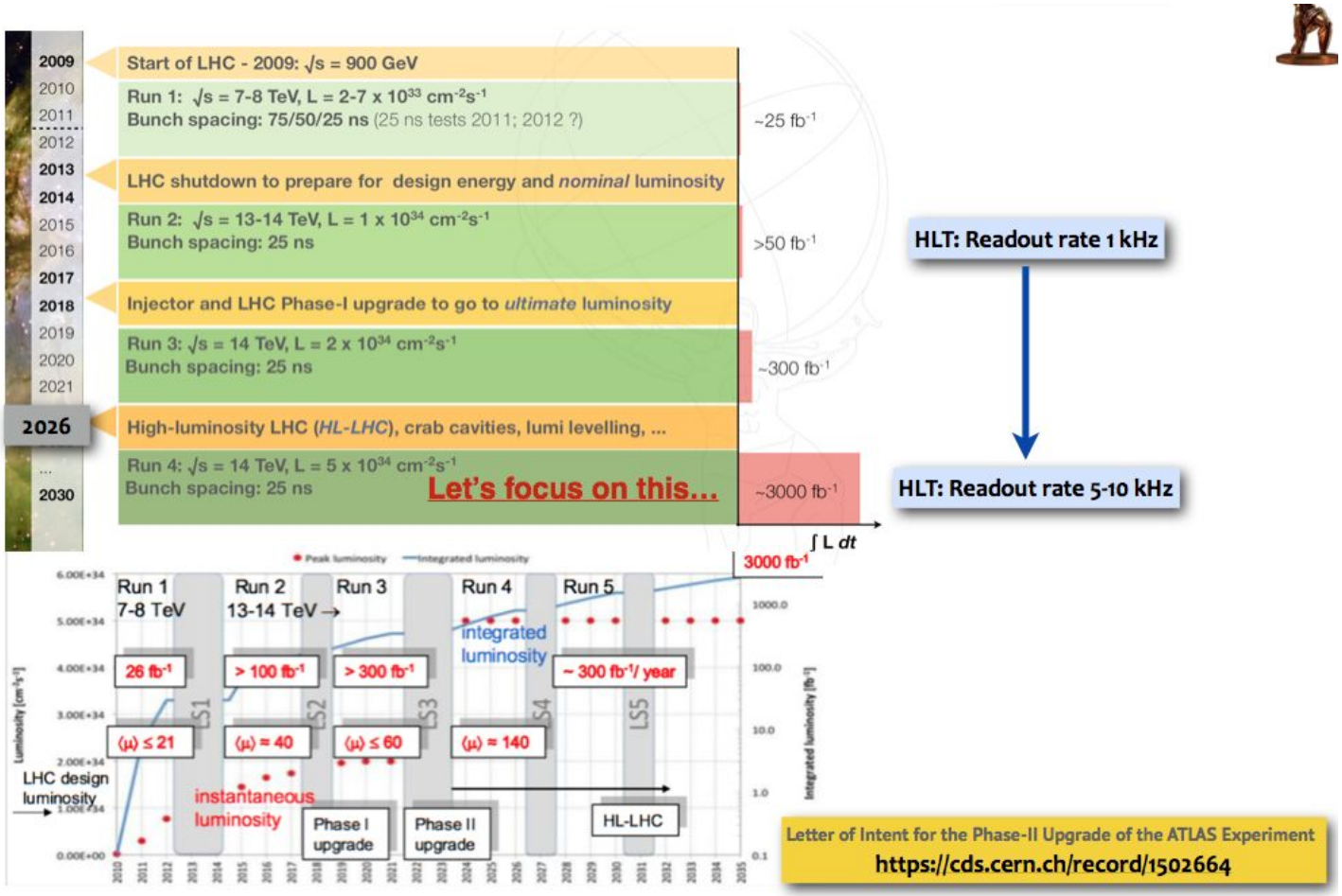


Current Challenges

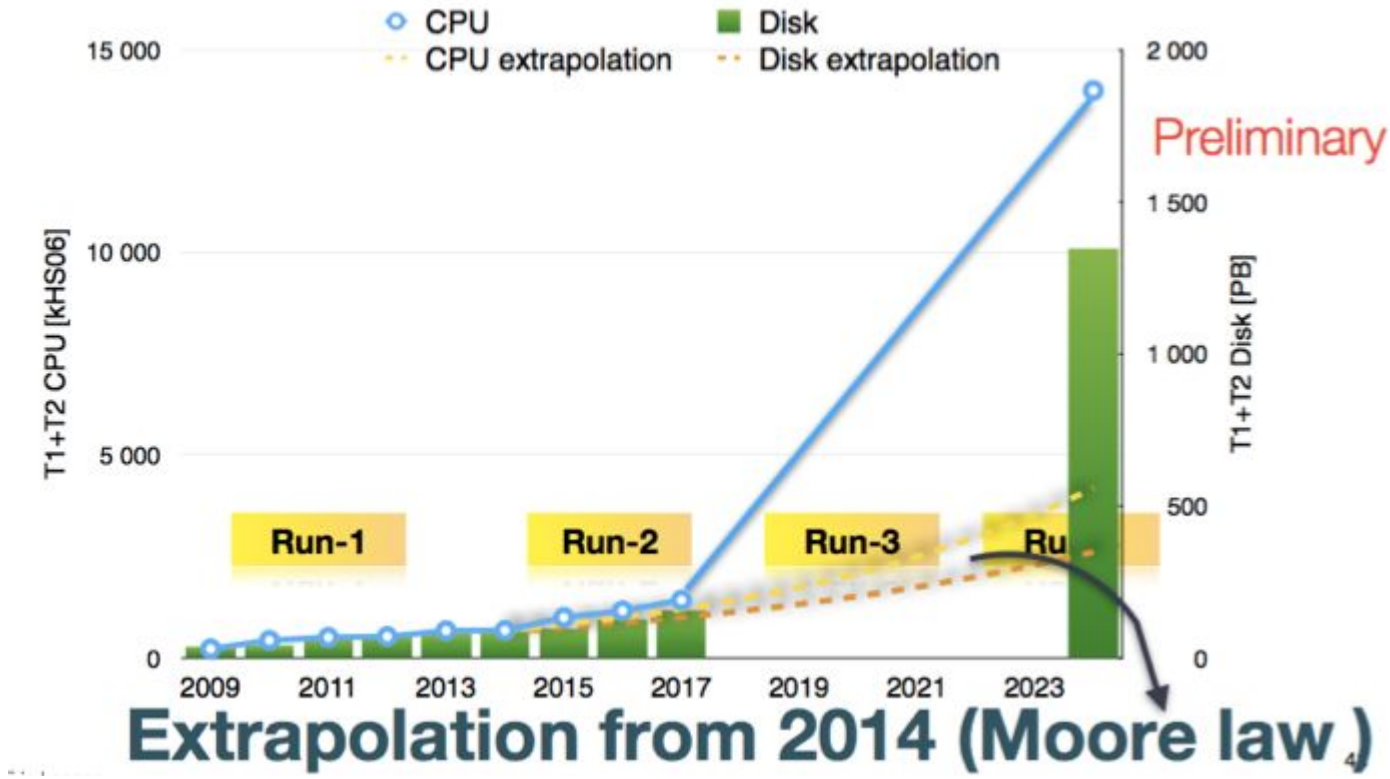


- New trends in data management
  - Original model was based on network being the weak point
  - But network has proven to be cheaper and better than expected
  - Break the rigid hierarchical model of data flow and sending jobs to data
    - Dynamic data placement
    - Remote data access over wide area network
- Event-level workflow instead of file-level
- Need more CPU and disk but with flat budget → opportunistic resources
  - High Performance Computing (supercomputers)
  - Volunteer Computing (general public)
  - Cache

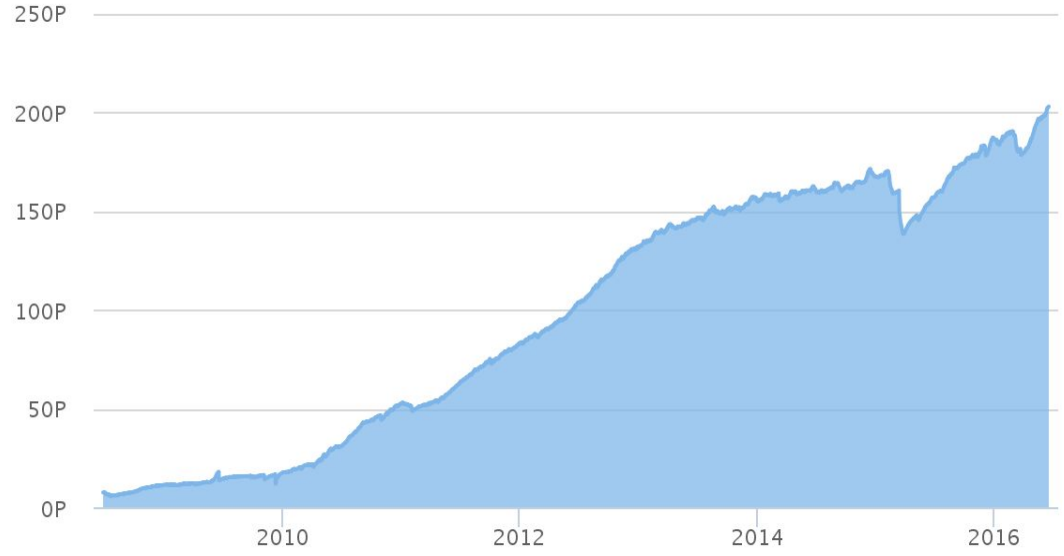
LHC Upgrade Timeline



ATLAS Resource Needs at T1s & T2s

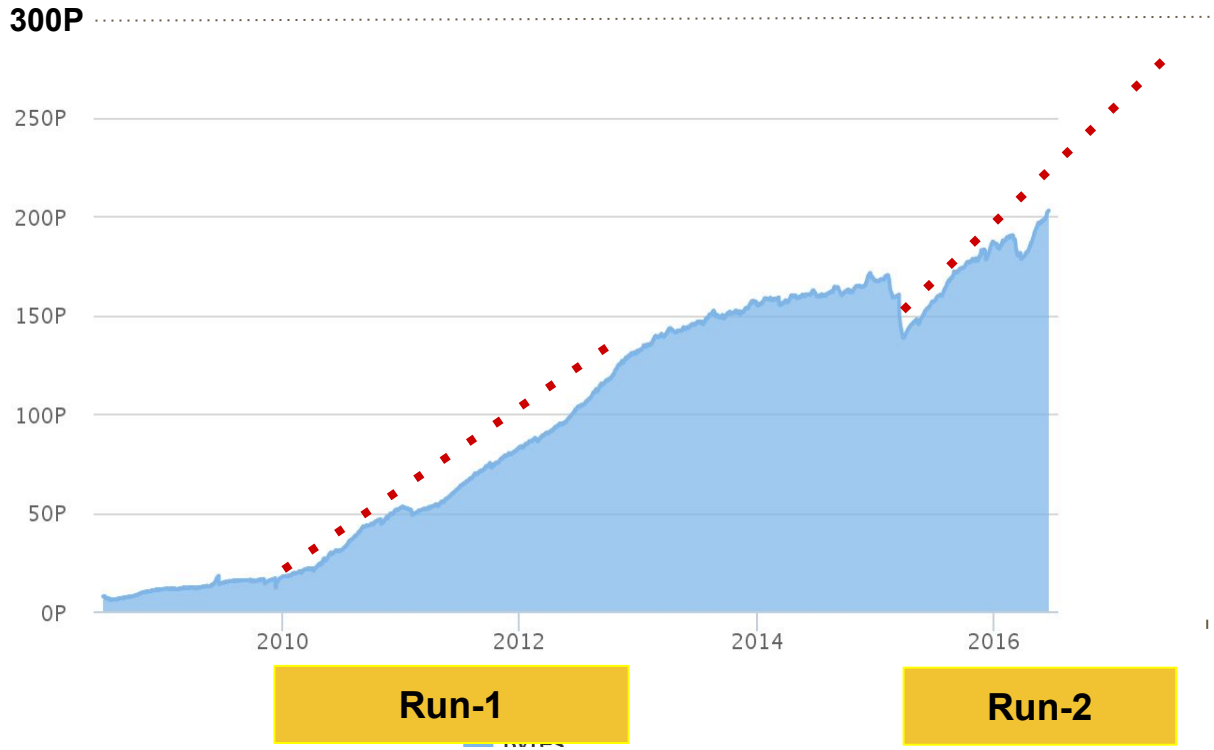


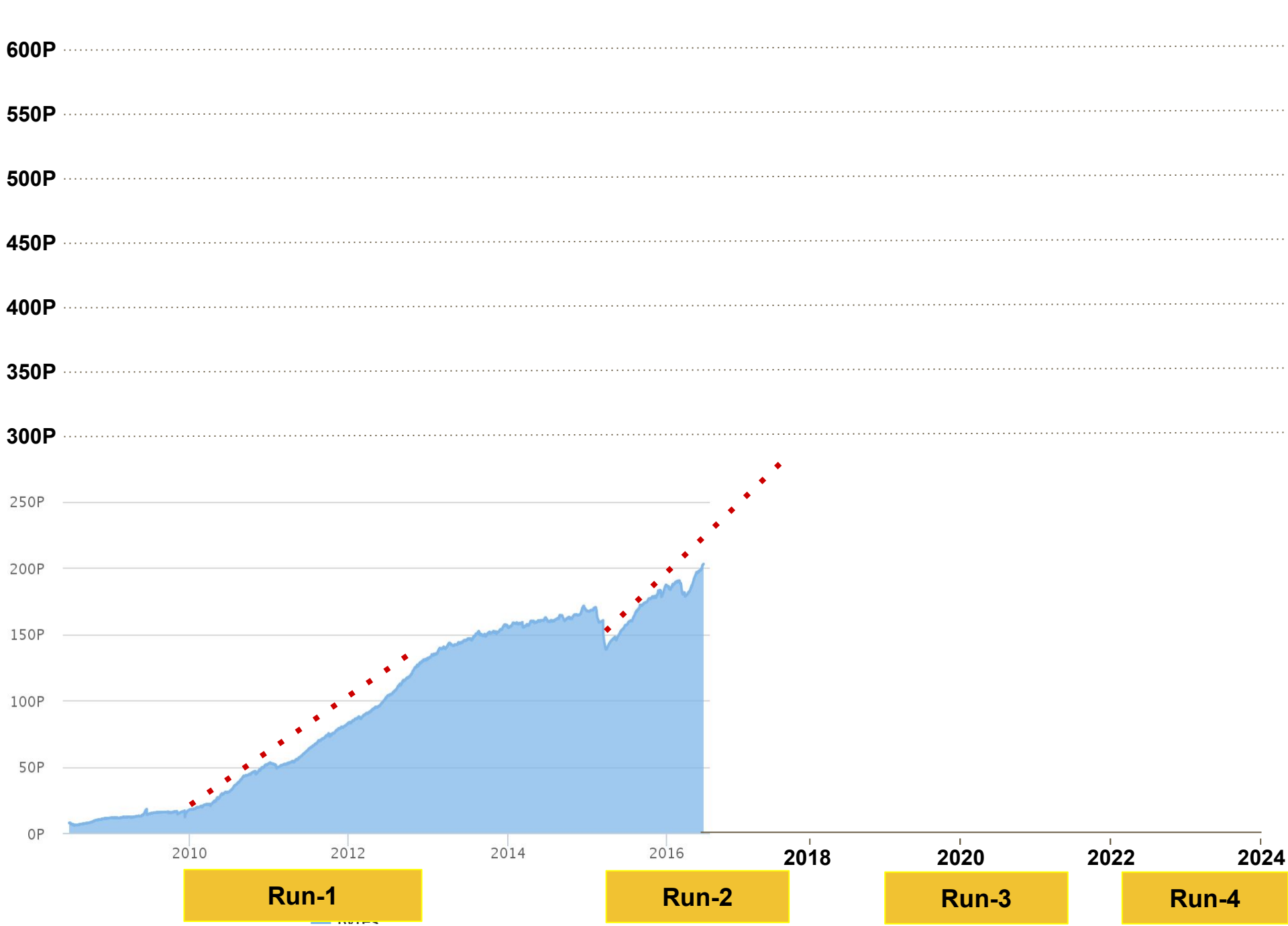
Slide by E. Lancon



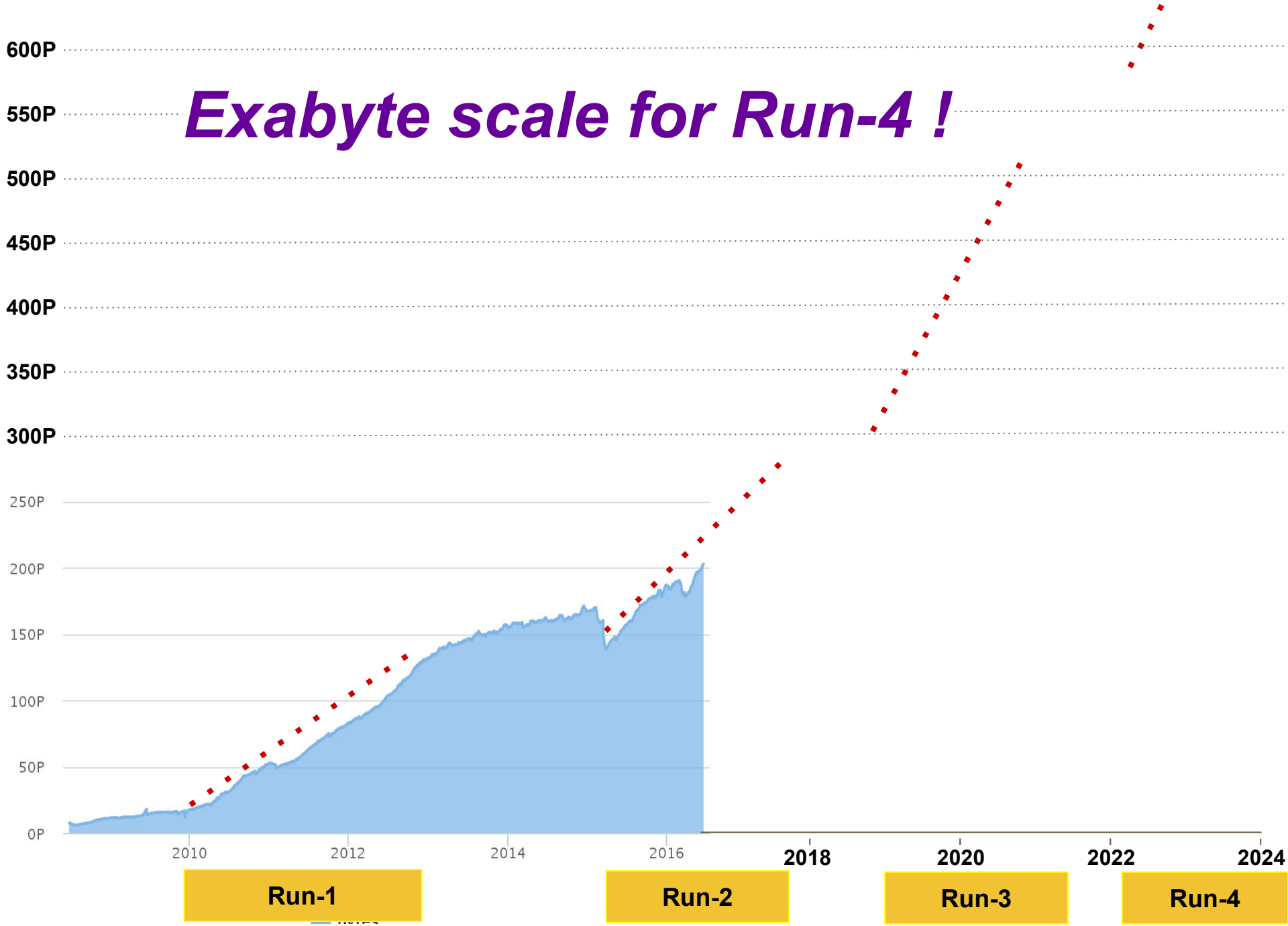
Run-1

Run-2


300P
250P
200P
150P
100P
50P
0P
2010
2012
2014
2016
Run-1
Run-2


600P
550P
500P
450P
400P
350P
300P
250P
200P
150P
100P
50P
0P
2010
2012
2014
2016
2018
2020
2022
2024
Run-1
Run-2
Run-3
Run-4

Exabyte scale for Run-4 !



ATLAS DDM: Current Numbers (2016)

The ATLAS DDM System/Rucio has demonstrated very large scale data management:

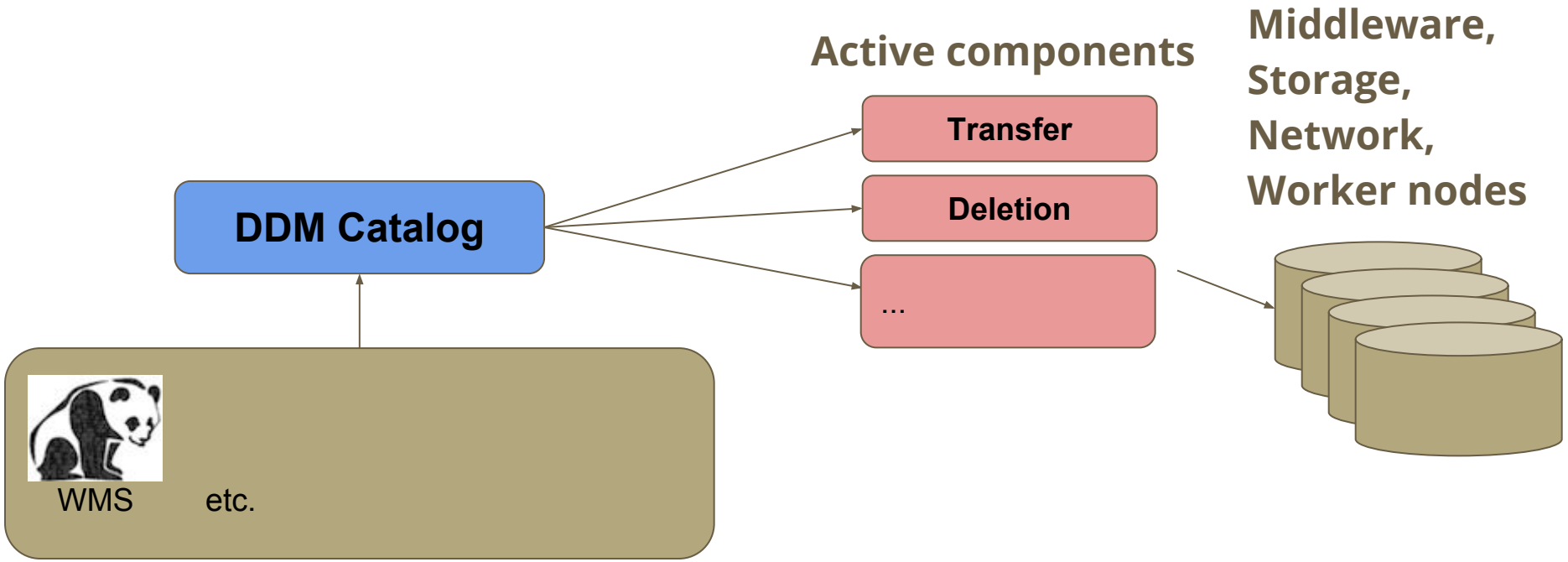
- 2000 ATLAS users
- 200 PB on 130 sites
- 1B file replicas
- 2.5 M file transfers/Day, 1.5 PB/Day
- 6 M deleted files/Day, 2 PB/Day
- 1M jobs/day

ATLAS DDM: Future Numbers (2026)?

The ATLAS DDM System/Rucio ~~has~~ **will** demonstrate~~d~~ very large scale data management:

- 2000 ATLAS users
- 200**0** PB on 130 sites
- 1**0** B file replicas
- 2**5** M file transfers/Day, 1**5** PB/Day
- 6**0** M deleted files/Day, 2**0** PB/Day
- 1**0** M jobs/day

DDM Evolution: Current Logical Overview



Future of Catalog ?

- Most of DDM implementations for the LHC are based on catalog
  - It's convenient to have one global and fast index for job scheduling
  - Easier to manage, few misses and availability > 99.99 %

- DDM has to scale with the (cumulative) number of data objects and operations
  - Data object(s) can be containers(s), dataset(s), file(s), event(s)
  - Most of the operations are generated by the workload management system, i.e., PanDA

Catalog Scalability & Database Technologies

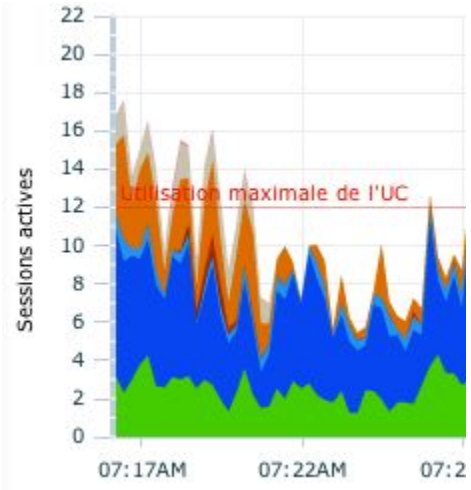
- For the scalability, it'll follow the advances in databases, open and standard technologies
- Rucio has a flexible design with no dependence on particular RDBMS implementation
  - Relational database management system
    - Use cases: Real-time data, transactional

  - Non relational structured storage
    - Use cases: Popularity, accounting, log analysis, ML

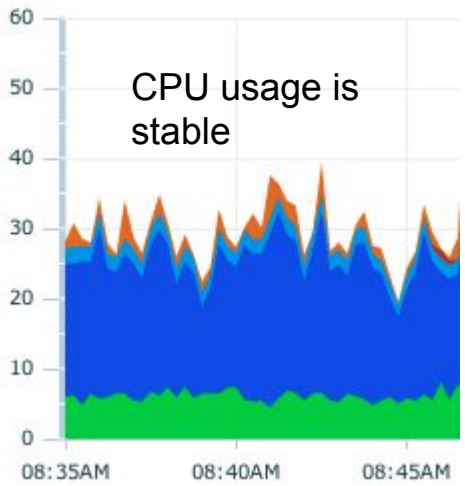
DB: Scaling tests - Load Increase

Load

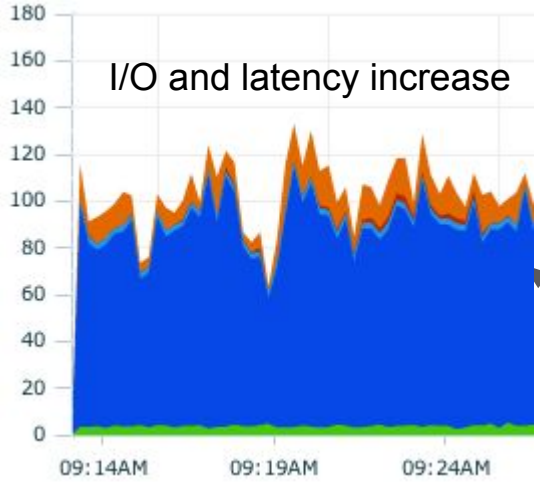
X1



X2



X4



Latency

Throughput

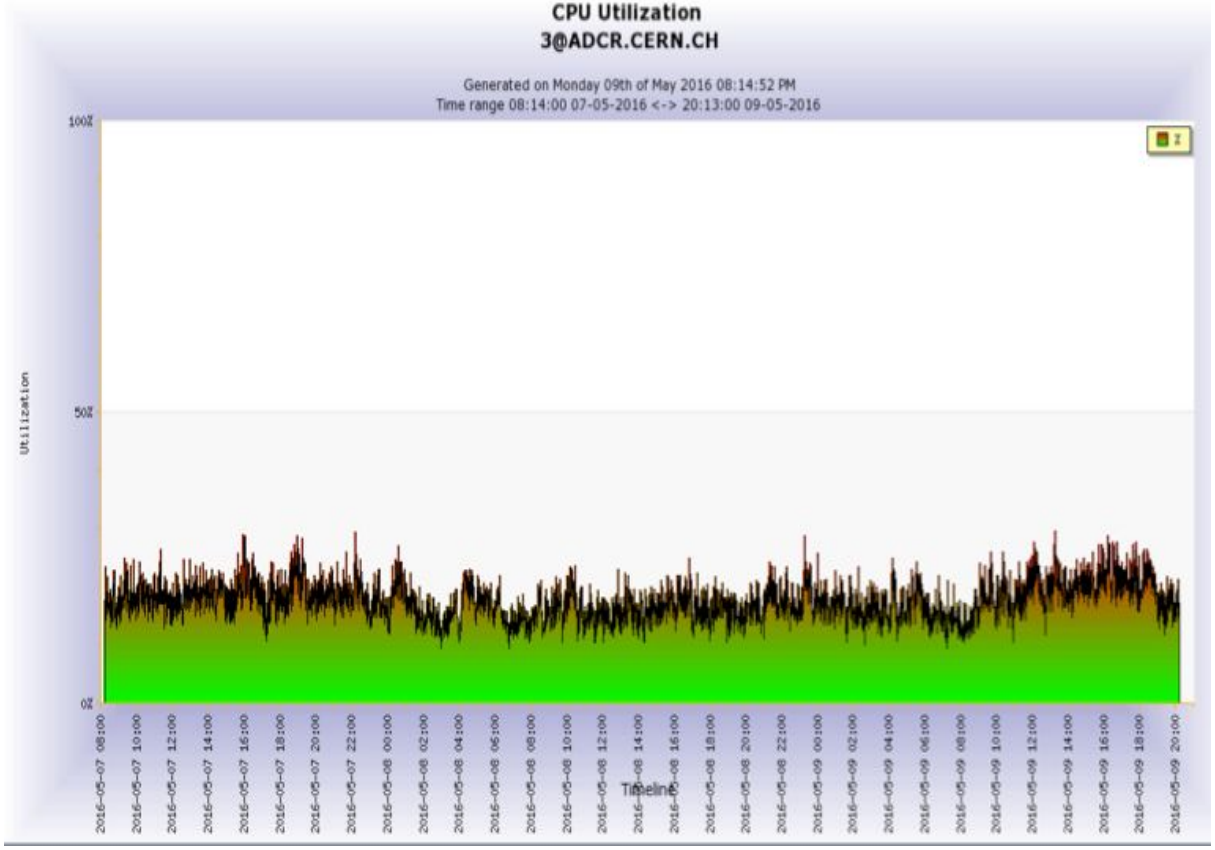
900 I/O ops.s-1
20 MB.s-1

2000 I/O ops.s-1
30 MB.s-1

3200 I/O ops.s-1
40 MB.s-1

Oracle: Current Status

Thanks to various improvements in the Rucio code and in the DB objects, Rucio load on the database is not high.



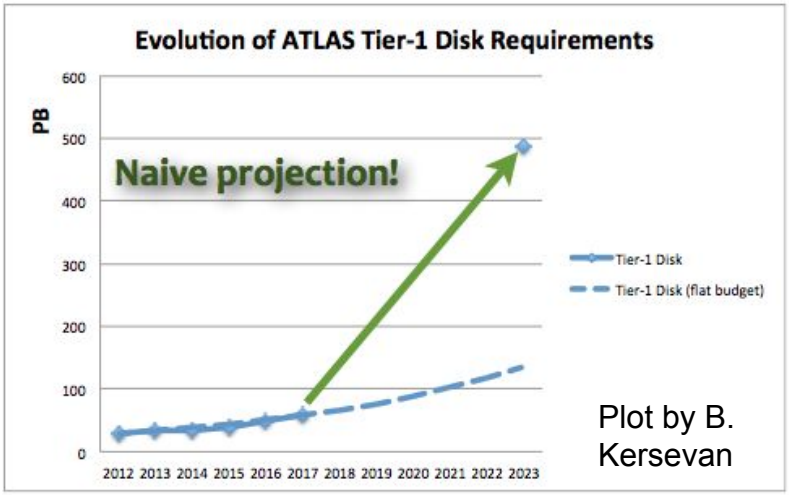
DBAs remarks:

- The Rucio load on the DB is acceptable. Keeping it low is important.
- A factor 4 more load is possible.
- New HW for the ADCR database is expected to be in place in Q1 of 2017 to serve the ADC applications from 2017 to 2020.
  - Factor 10 doable ?

Personal Comments

- I have observed a general trend to have more and more (physics) metadata in DDM to facilitate data selection, discovery and analytics
  - Should DDM and the metadata part be strongly coupled ?
- Production, analysis and workload management workflows have a direct influence on the system' scalability
  - One scheduling change can have butterfly effect on DDM
  - Is the current strong separation between DM and WM an affordable model for run-4 ?

Storage & Network



Plot by B. Kersevan

- Resource increase w.r.t. 'flat budget':
- In ten years from now, the terabyte price on disk might be divided by 3
- Tapes foreseen to be viable for both capacity and cost
- Looking at the trends, the world gained an order of magnitude of storage over the last five years…
- We have to keep up to date with data storage/transfer technology and evolution
- The biggest predictable gain will come from network and will strongly influence the experiments computing models

⇒ Data storage of HL-LHC experiments will NOT become a 'trivial' problem ten years from now

Summary

- Many questions, no immediate answers