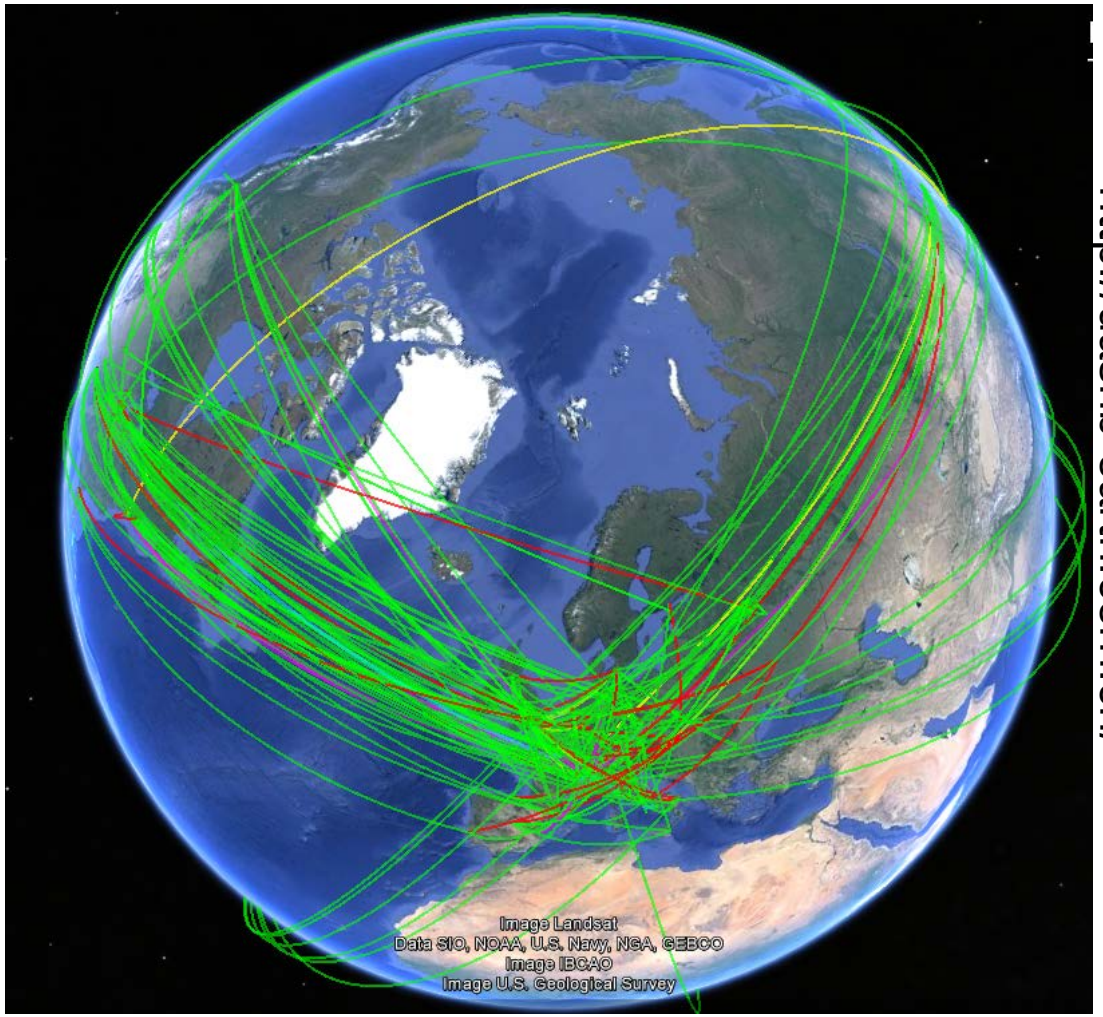
Computing Challenges for Run2

L. Poggioli, LAL, Orsay



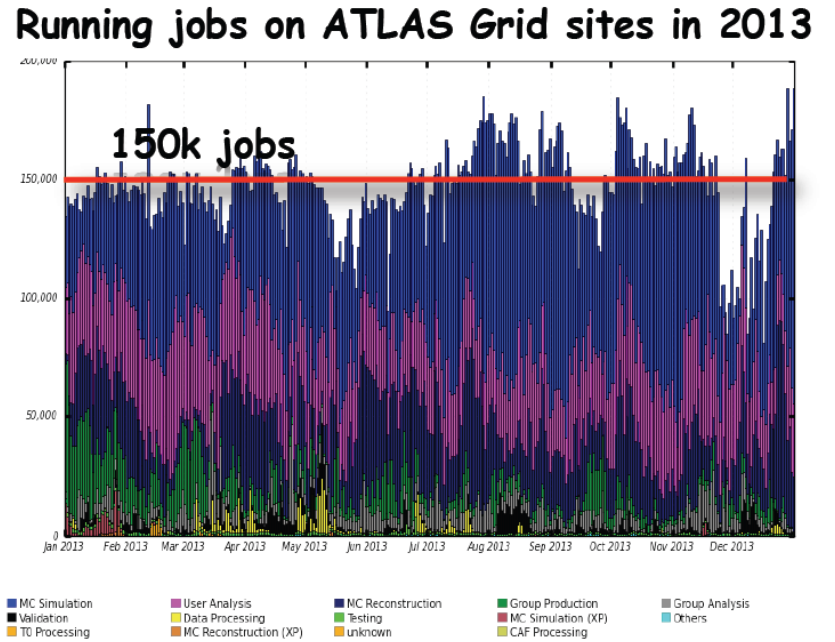
http://dashb-earth.cern.ch/

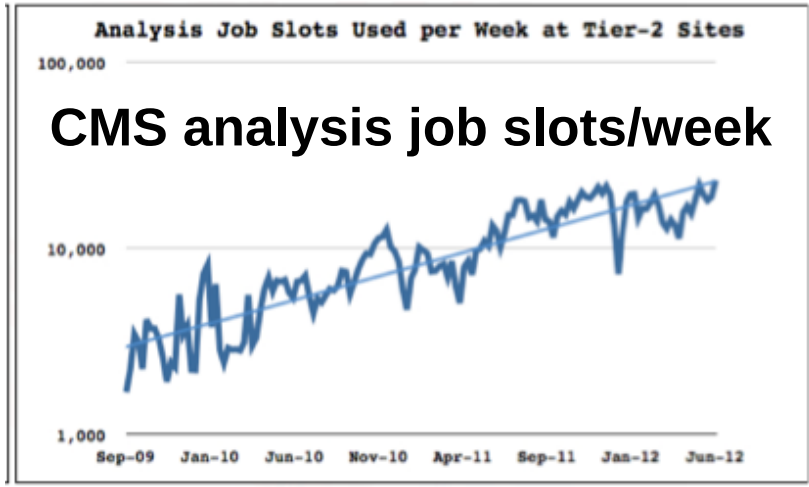
Outline

- Run1 outcome
- Run2 challenges
- Solutions
  - Model
  - CPU/Storage
  - Opportunistic resources
  - Network benefits
- Summary

Run1 Outcome

- ATLAS
  - 350M jobs in 2013
    - Analysis jobs > 50%
  - 1.2EB data read-in
    - > 82% by analysis
- CMS
  - 6pB data, 13B MC evts
  - 4pB/month transfers
  - MC/Data = 1.1





It worked beyond expectations!!

Analysis: Main driver of storage & I/O capacity

Run2 Challenges

- Flat budget constraints
  - h/w increase from Moore's law gain
    - Estimated factors of 1.2/year for CPU & 1.15/yr for disk & tape
- Data from Run-1
  - Proper data preservation
- LHC operation
  - HLT rate x 2.5
  - Pile-up > 30
    - -> Reco time x 2-2.5
  - 25ns bunch spacing
  - c.m. energy x 2
- 'New' detector
  - To be integrated in simul & reco

Computing Model 2010-2013

Network performance **breakthrough** (eg LHCONE 2011)

- Going away from hierarchical Model (T0-T1s-T2s)
- Dynamic data placement & deletion based on popularity
- T2→N-T1s & T2↔T2 exchanges – New T2D with data (LHCb)

Planned data distribution

Jobs go to data

Multi-hop data flows

Poor T2 netwking across regions

Planned & dynamic distribution data

Jobs go to data & data to free sites

Direct data flows for most of T2s

Many T2s connected to 10Gb/s link

Limitations of current model & tools

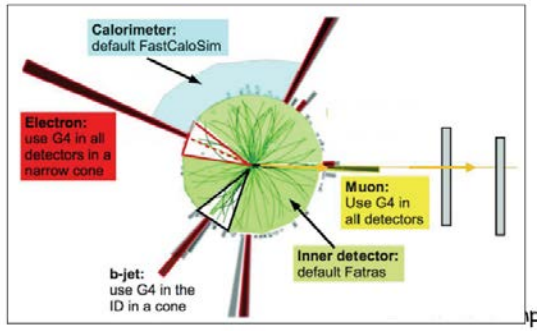
- Partitioning of resources
  - Analysis vs Central Production - T1s versus T2s
- Data distribution management & Production systems limits to scale to new conditions
- Memory increase of MC pile-up digi & reco
- Multitude of data format for analysis

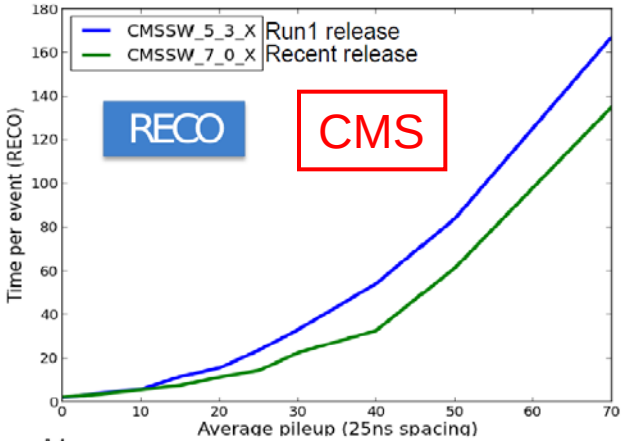
-> Gain needed in Simulation (CPU), Reconstruction (CPU, memory), Analysis (Data format, disk space, CPU)

Run2: Extrapolation & extension of end of Run1 framework

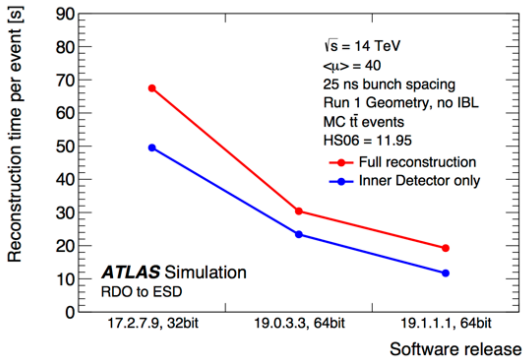
CPU Optimization

- Better usage of resources
  - Less MC/data than in Run1
  - Prompt Reco/No Repro (LHCb)
  - More Fast wrt Full sim
  - Optimization Fast/Full





ATLAS



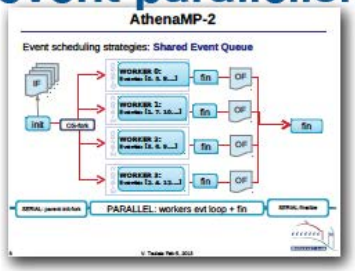
- Software improvements
  - Optimize track seeding
  - Use vectorized trigo. functions
  - Use faster algebra libraries, simplify data model

See M. Elsings's talk

Software Changes

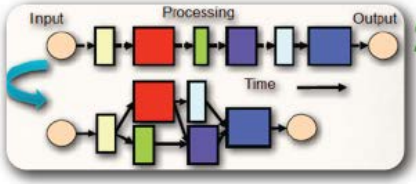
- All expts embarked in deep changes
  - Cf. HEP Software Foundation
- Memory footprint reduction
  - Using **multicore** jobs
  - Baseline for reconstruction
- Memory sharing
  - Using **multithreading**
- Revision of data models
- Vectorization
  - To exploit new architectures (GPU)

event parallelism



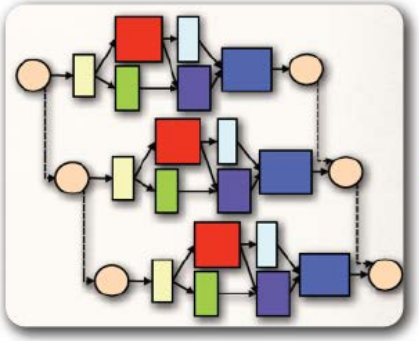
Today

algorithm parallelism



2015-2016 ?

event & algorithm parallelism



LS2 or before?

Storage

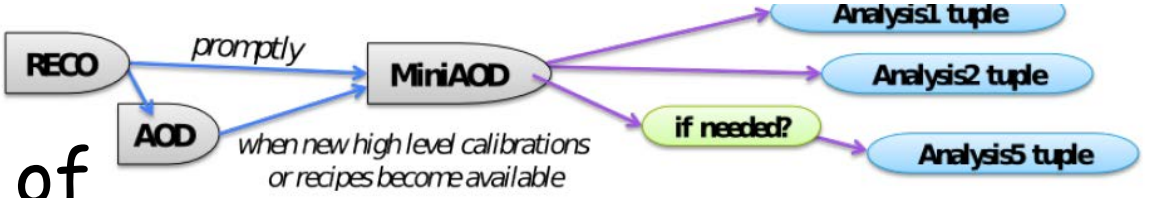
- Analysis formats
  - Reduce # types (xAOD/ATLAS), smaller (miniAOD/CMS, MDST/LHCb) -> Gain in space
  - Limit # replicas at T1s & T2s
- Disk
  - More efficient use of **dynamic** placement
  - More agressive **deletion** of non-popular data
- Tape
  - More usage of tape (~5x cheaper/TB than disk)
    - **Centrally** organized activities will read more from tape
  - Decoupling of Disk & Tape at T1s (CMS) for user

Workload / Data Management

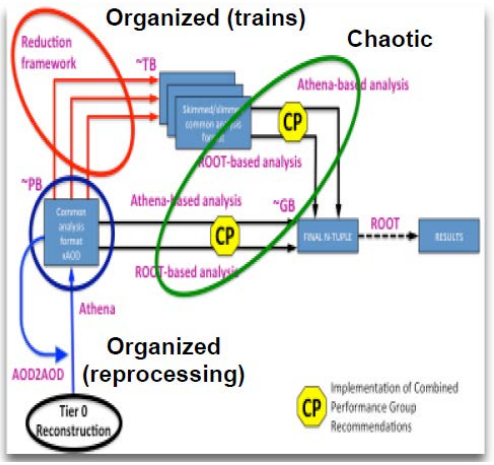
- Workload
  - Less separation T0/T1s/T2s
    - T1s can take T0 load, T2s do reprocessing (T1s task)
  - Unify analysis & production
    - Single queue (CMS), same engine (ATLAS)
    - Better reactivity to analysis loads
- Distributed Data management
  - New scalable architecture (eg ATLAS)
    - Built-in replication policy (space & netwk optimization)
  - Streamlining
    - Limit # catalogs for handling data (LHCb, ATLAS)
    - Use more powerful protocols for transfers (FTS3)

Analysis Model

- Goal: Minimize (i.e. Common) analysis formats & optimize analysis tools (submission,…)
- CMS: MiniAOD
  - Replaces dozens of Group Ntuple/trees - Small size (50kB)
  - Improved elements for job resubmission & task completion



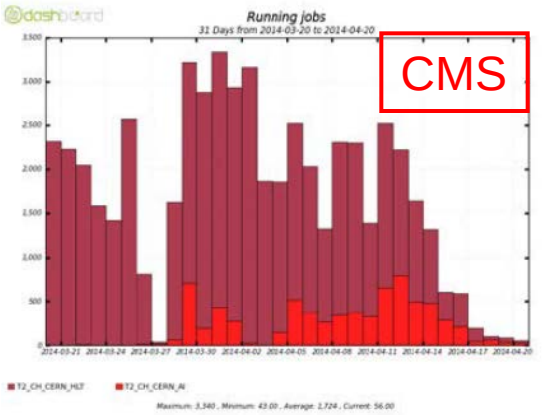
- ATLAS: xAOD
  - Data reduction frawework (PB->TB)
    - Group data sample centrally produced

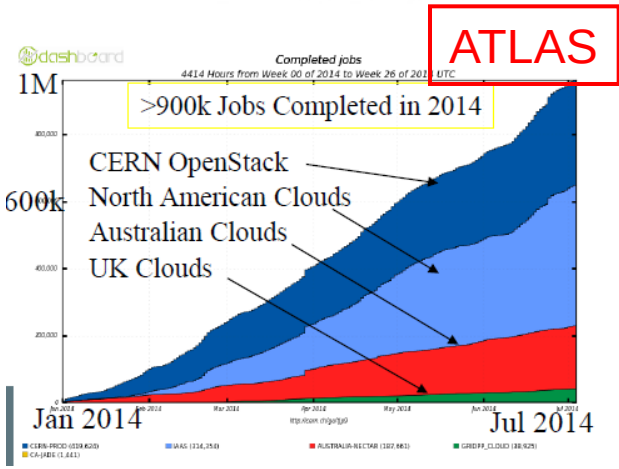


- LHCb: Generalized use of MDST

Opportunistic resources (1)

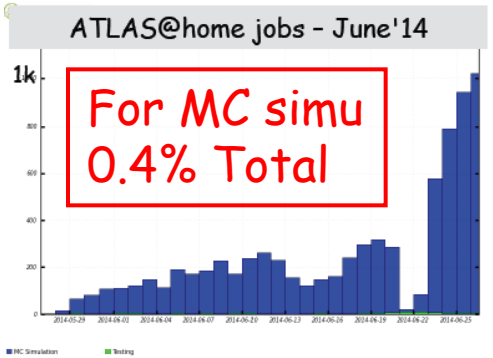
- Virtualization
  - Ask for resources thru interface and get access & control of (virtual) machine, i.e. job slot on Grid
  - High Level Trigger (HLT)
    - Use resource between fills
      - Not for LHCb (Farm used between fills for deferred HLT processing)
    - Expect CPU power ~T1 or big T2
  - Clouds usage
    - Academic (eg OpenStack @ CERN)
      - eg 6K cores at CERN-T0 for Heavy Ion reprocessing (CMS)
    - Commercial (Amazon EC2, Google)





Opportunistic resources (2)

- Super Computers (HPC)
  - From Peta to ExaFLOPS
  - Large # CPU cycles can be used parasitically
    - eg MC simulation (10% Grid production, 10-20k cores)
  - Issues: I/O & outbound connectivity
- Volunteer Computing using BOINC
  - Used by LHCb & ATLAS: **Free!!**
    - Solution for Institute desktop clusters
    - Can work at event level
      (Cf. ATLAS event service)



- Extra-unpledged resources at sites
  - eg T3s resources, opportunistic, as in Run1

Remote Data Access (1)

- Networking keeps on progressing fast
  - x10 every 4.25 yrs / Already 100Gb/s among US
- -> Jobs can access data remotely via network
  - Allows better usage of storage resources
  - Breaks the 'jobs go to data' Grid paradigm!!
  - Better suited to Analysis jobs
- Protocols
  - http: Allows direct download files from Grid to local
    - Not in quality production today
  - Xrootd: Allows direct data acces in ROOT & analysis s/w (ATLAS, CMS, LHCb in deployment)

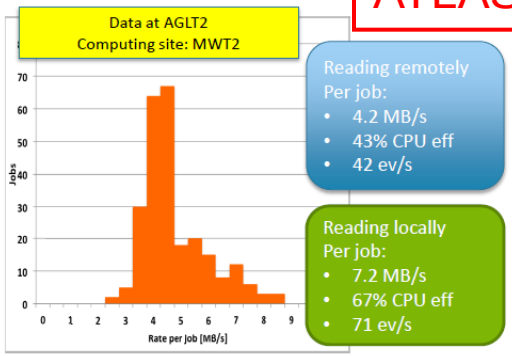
Remote Data Access (2)

- Remote access modes using Xrootd
  - Grid job recovery if data access issue at site
  - Run jobs at site w/o data & access files remotely
- CMS: Anydata Anytime Anywhere
  - Access 20% data over network
  - Small loss in efficiency local vs remote
  - Mostly for Analysis, intend to use for reprocessing

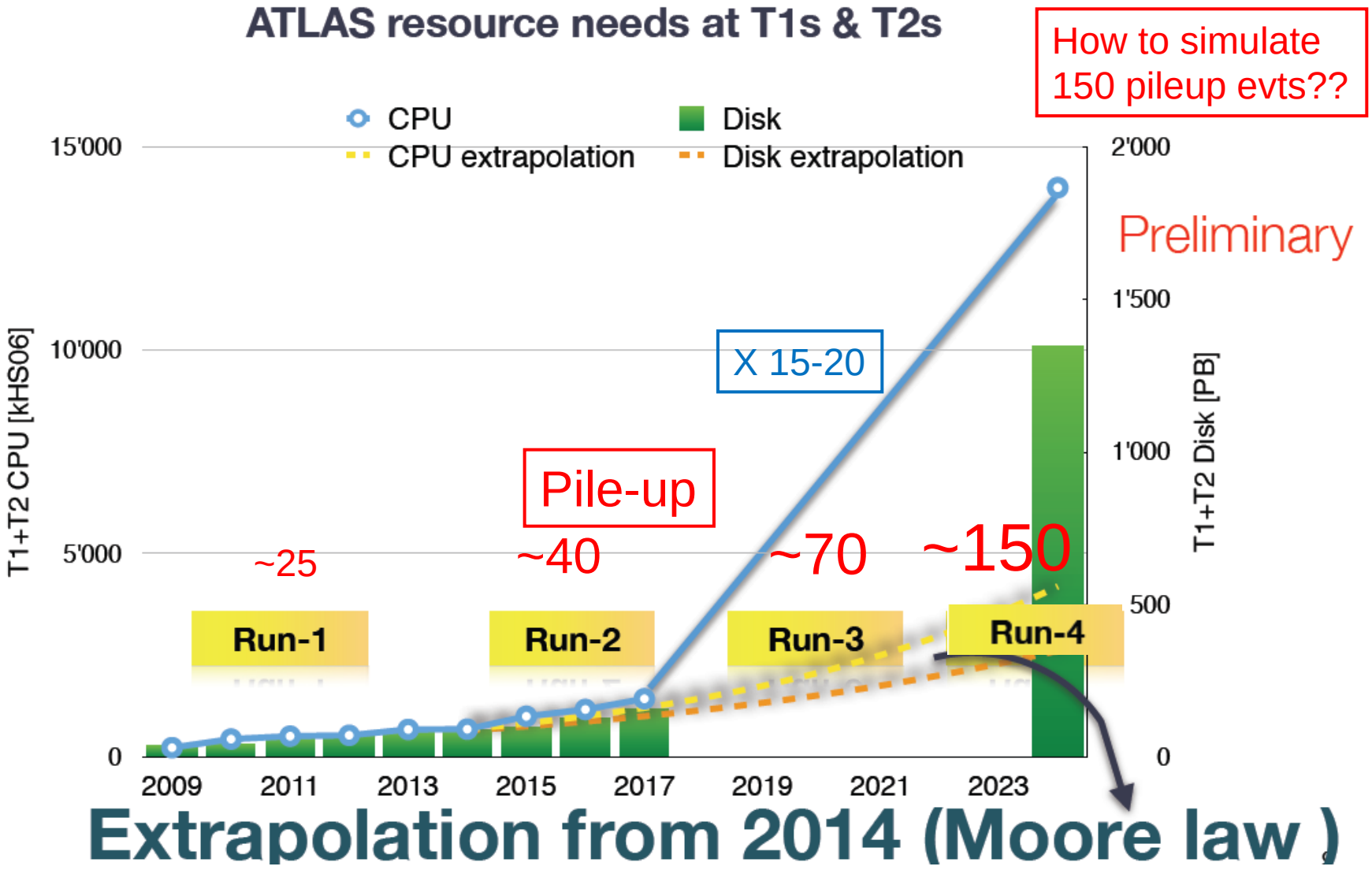
CMS

- ATLAS: Federated ATLAS Xrootd
  - Recovery mode OK
  - Running remotely mode under test
    - Potential impact of network saturation

ATLAS



Beyond Run2


ATLAS resource needs at T1s & T2s
How to simulate 150 pileup evts??
CPU
CPU extrapolation
Disk
Disk extrapolation
Preliminary
15'000
10'000
5'000
0
2'000
1'500
1'000
500
0
T1+T2 CPU [kHS06]
T1+T2 Disk [PB]
X 15-20
Pile-up
~25
~40
~70
~150
Run-1
Run-2
Run-3
Run-4
2009
2011
2013
2015
2017
2019
2021
2023
Extrapolation from 2014 (Moore law)

Summary

- Run1 completed successfully
  - A lot of experience gathered
  - Computing acknowledged as key component
- Run2 is an evolution of Run1
- Many ideas investigated for Run2
  - Cf. LHC Computing Model Update document
  - Role of Tiers, use of network, data federation, clouds, opportunistic resources
  - Big efforts by experiments to optimize & gain in resource (CPU, memory, storage)
  - All these ideas being tested at full scale now
- Manpower is an issue