

Software and Computing Challenges at the High Luminosity LHC

Borut Paul Kersevan

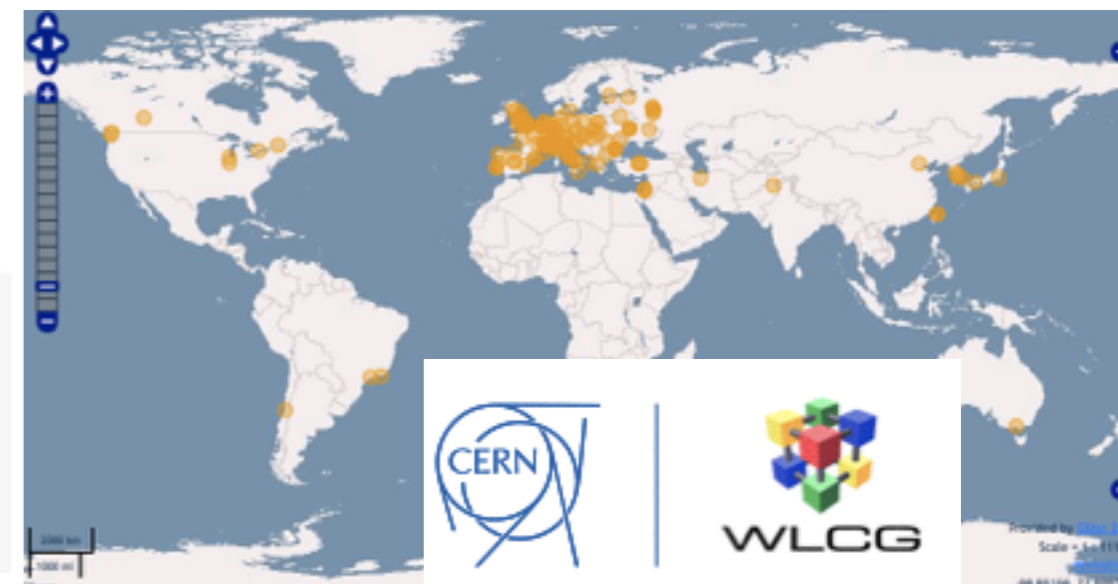
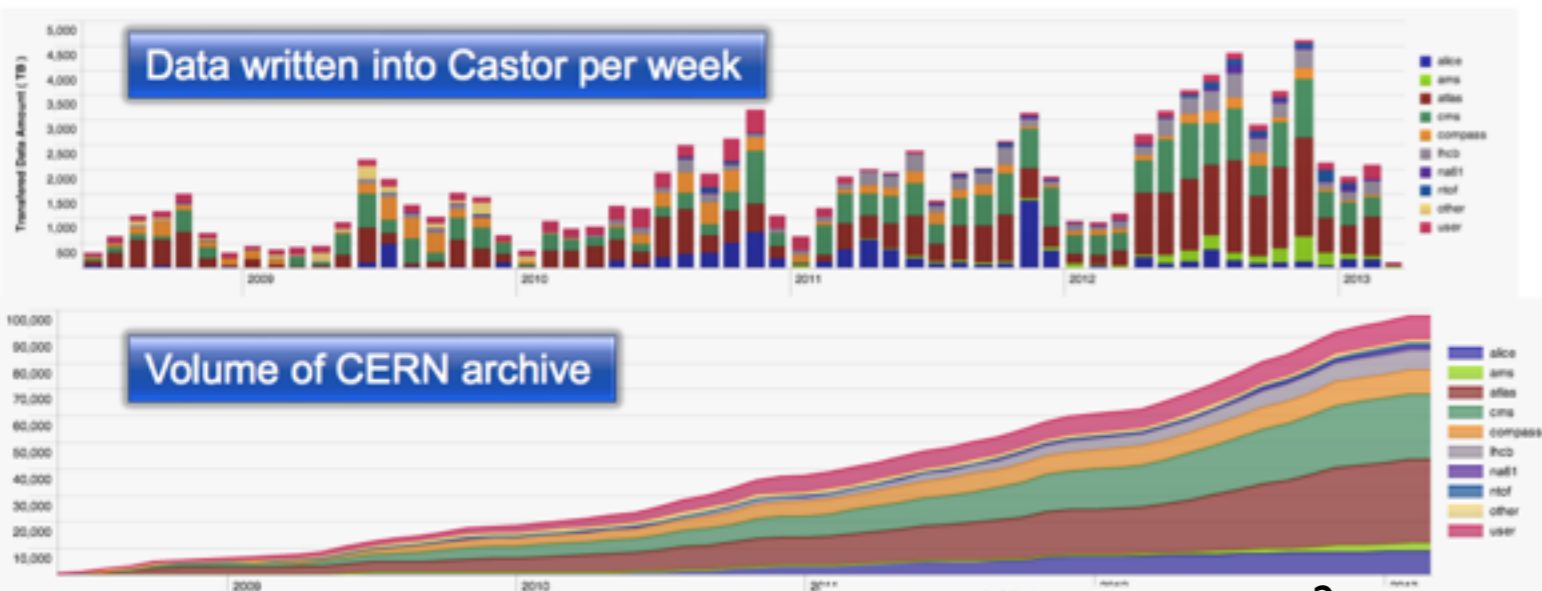
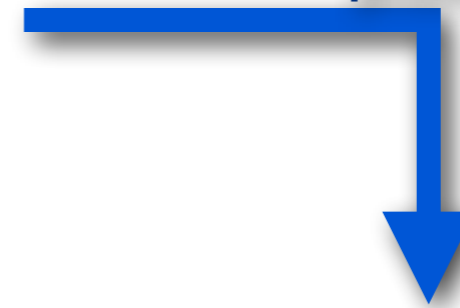
Jozef Stefan Institute,
Faculty of Mathematics and Physics,
University of Ljubljana

Introduction

- LHC delivered **billions** of recorded collision events to the LHC experiments from proton-proton and proton-lead collisions in the Run 1 period (2009-2013) and the ongoing Run 2 (2015-2018).
 - This translates to multiples of 100 PB of data recorded at CERN.
 - several 100 PB more storage needed across the **Worldwide LHC Computing Grid** to provide space for archival, replication, simulation and analysis.
- The challenge how to process and analyze the data and produce timely physics results was always substantial but in the end resulted in a great success.



Data transferred from CERN
across the world for access and
processing

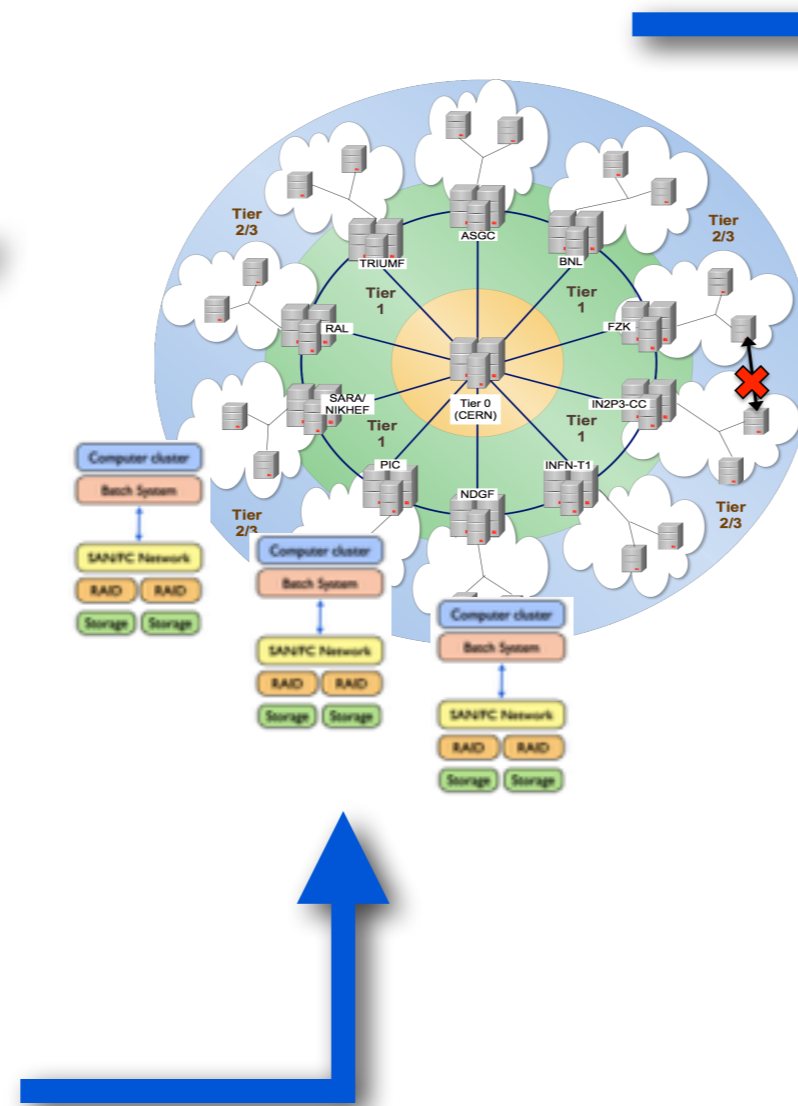
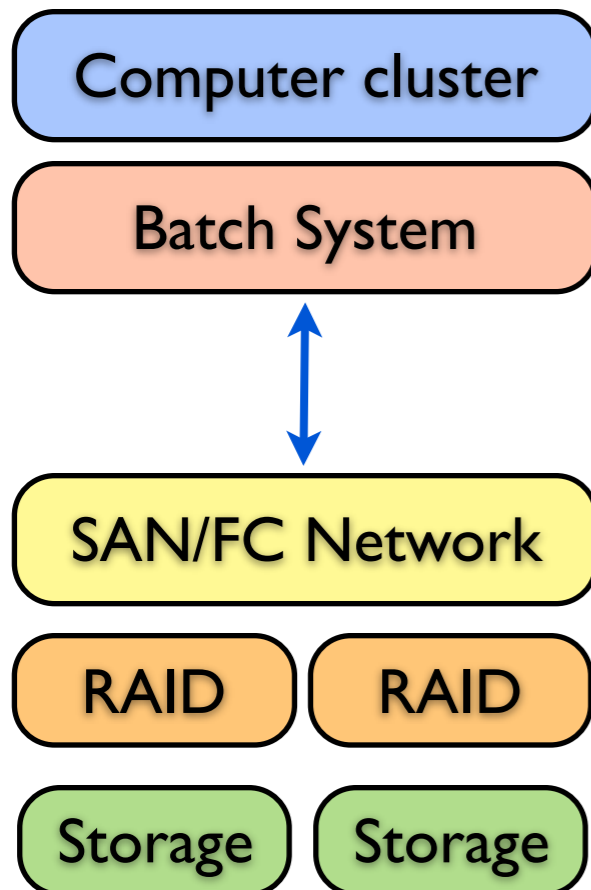


Distributed Computing Environment Technologies

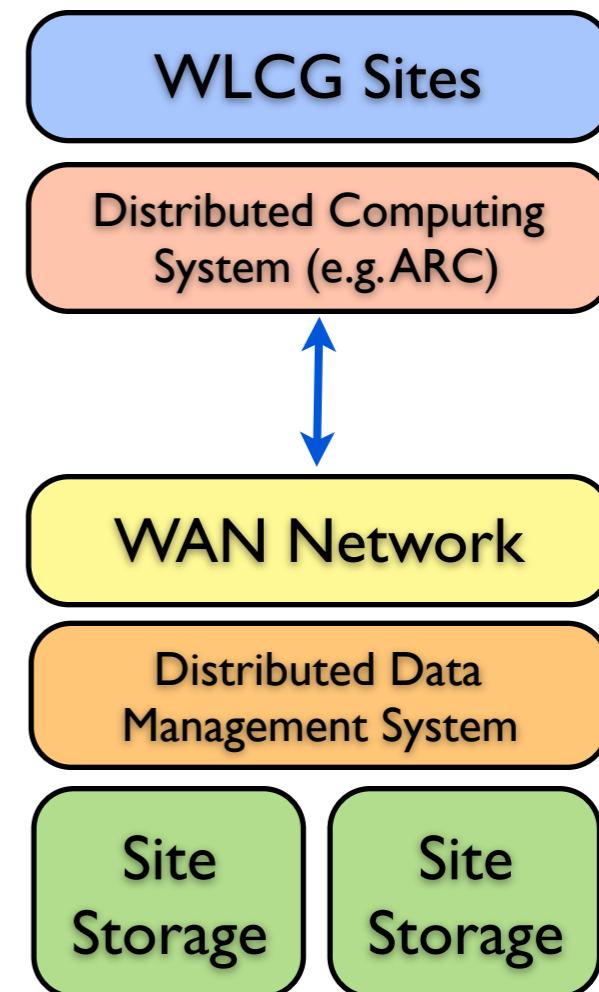


- Distributed computing introduced a new scale w.r.t. a local computing facility in terms of data and job management.
 - No 'industrial' standard or simple rules on what are optimal solutions for data placement and job brokerage to ensure the optimal usage/minimal job latency/..
 - **The LHC experiments composed their Computing Models based on best knowledge of the new and evolving system, including their experiment specifics.**

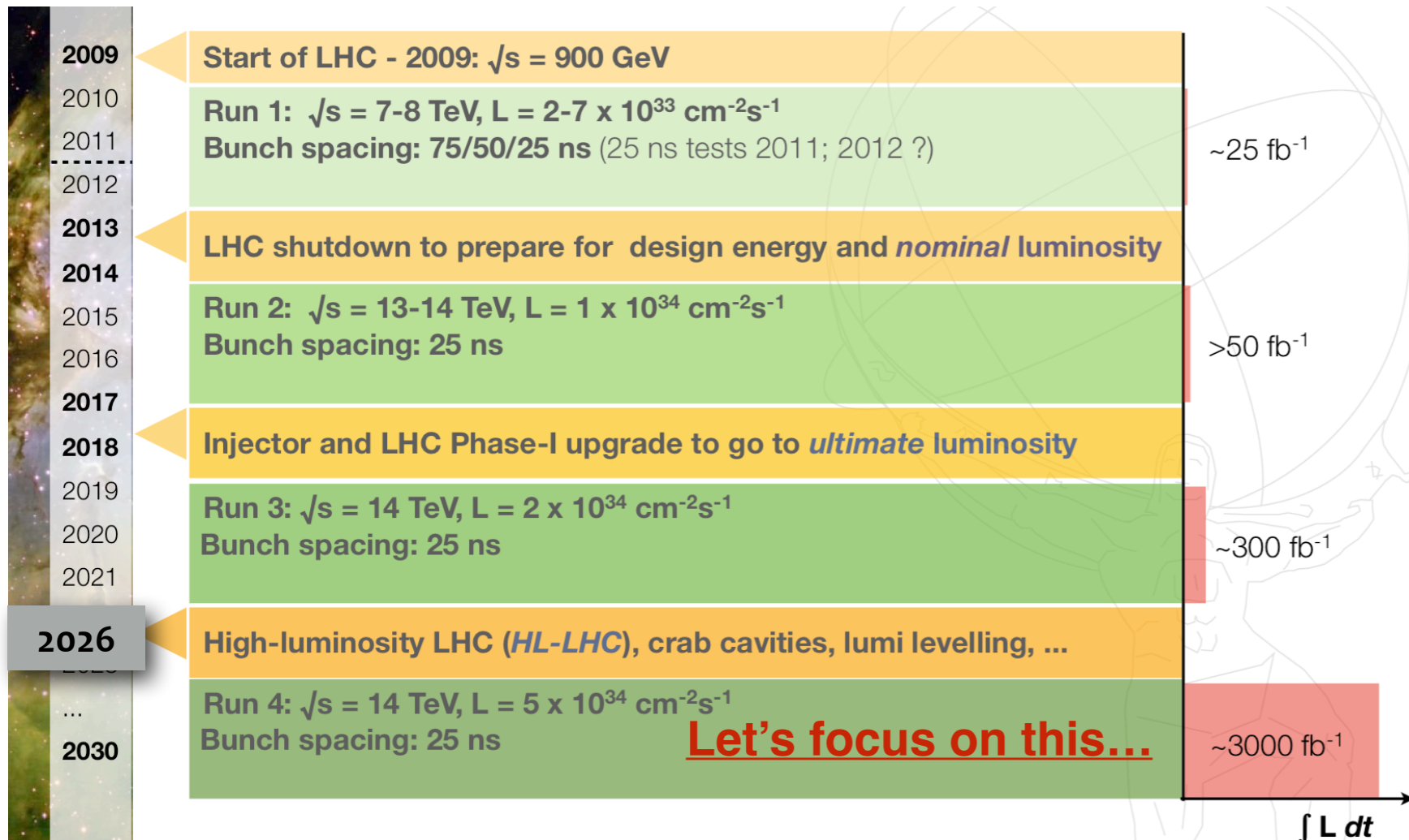
Local computing facility



Distributed Computing 'Facility'



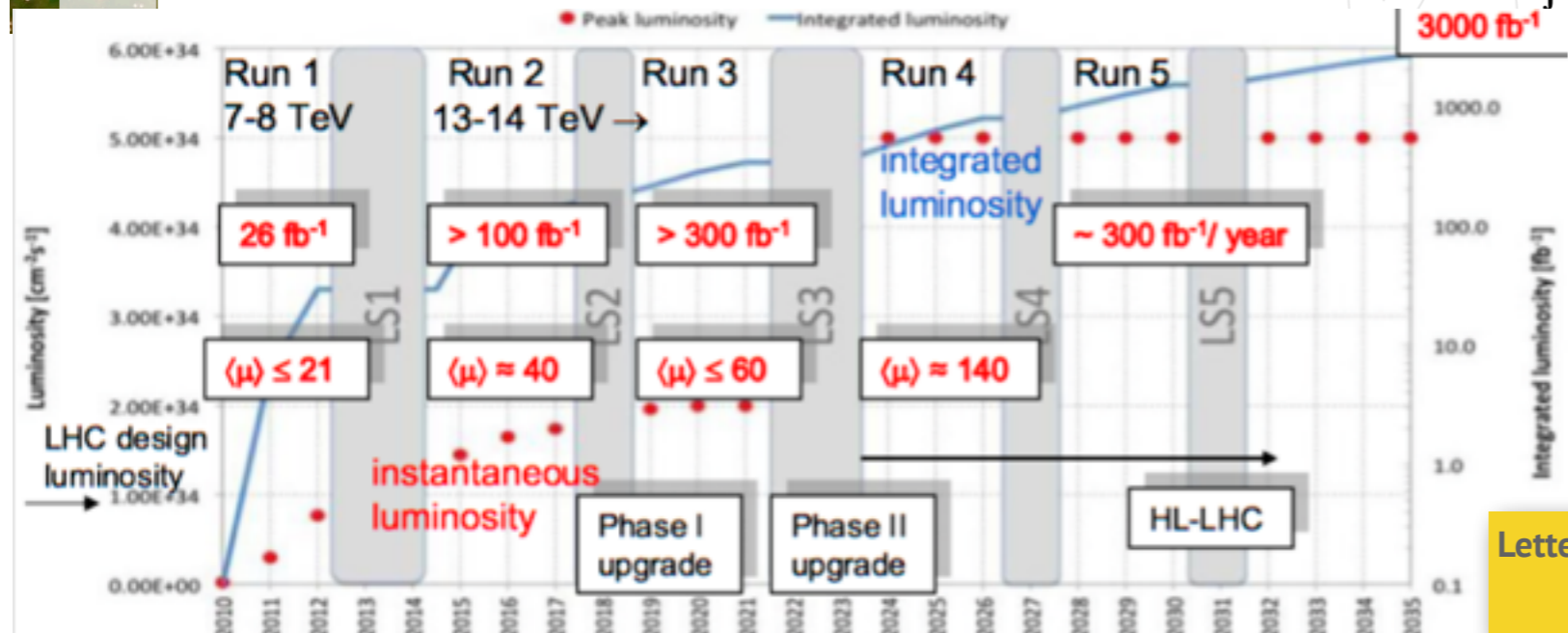
LHC Upgrade Timeline - the Challenge to Computing Repeats periodically!



HLT: Readout rate 1 kHz



HLT: Readout rate 5-10 kHz



Letter of Intent for the Phase-II Upgrade of the ATLAS Experiment
<https://cds.cern.ch/record/1502664>

ATLAS S&C 10 years from now..

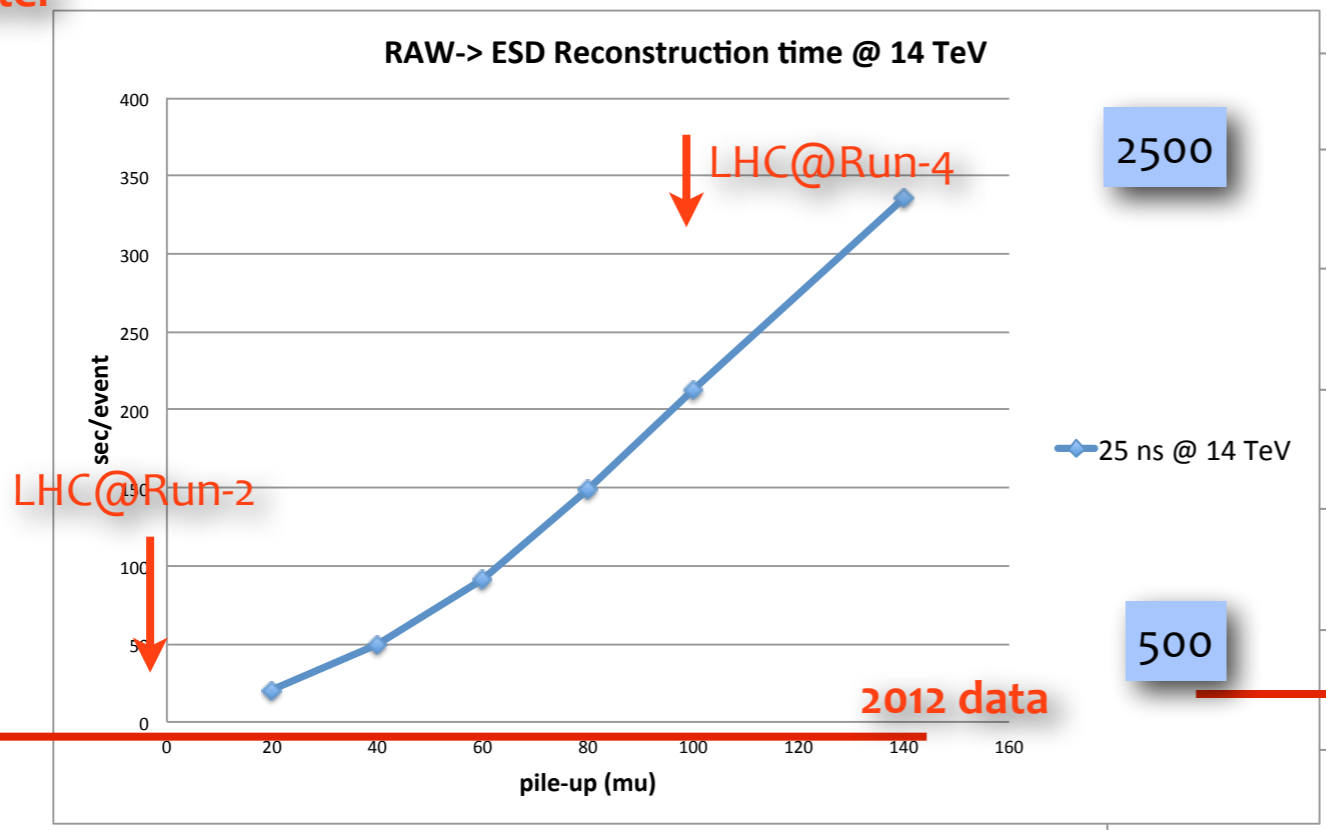


- **Resource use and availability: What can we say/assume about ATLAS software & computing 10 years from now (Run-4/HL-LHC)?**
 - Boundary conditions:
 - The HLT data collection rate w.r.t. Run-2 increases by **~ an order of magnitude**
 - the simulation statistics also **scales accordingly**.
 - **The pile-up also increases** by roughly an order of magnitude (factor 5-10) - **implications on event size and processing time** (e.g. reconstruction CPU/event, analysis..).

R. Seuster

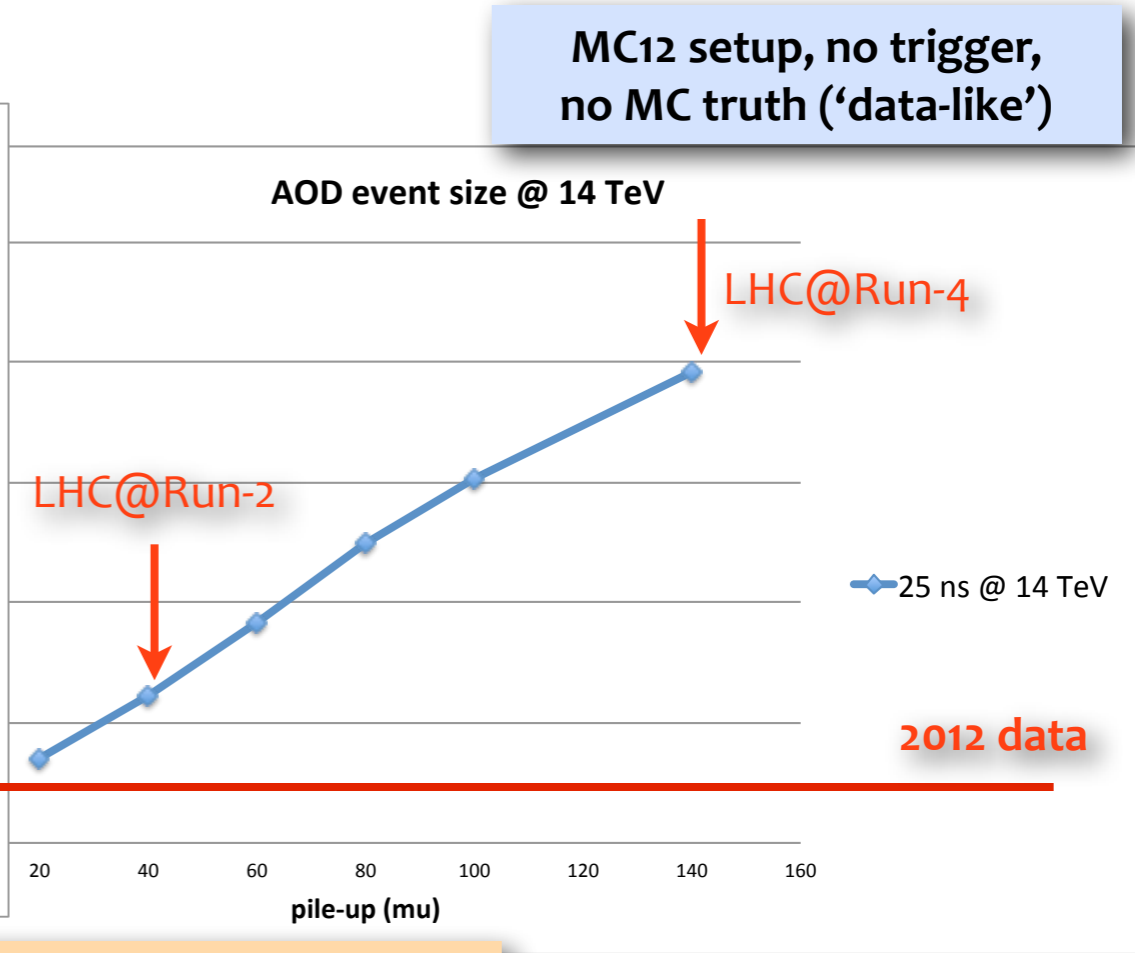
400

100



2500

500



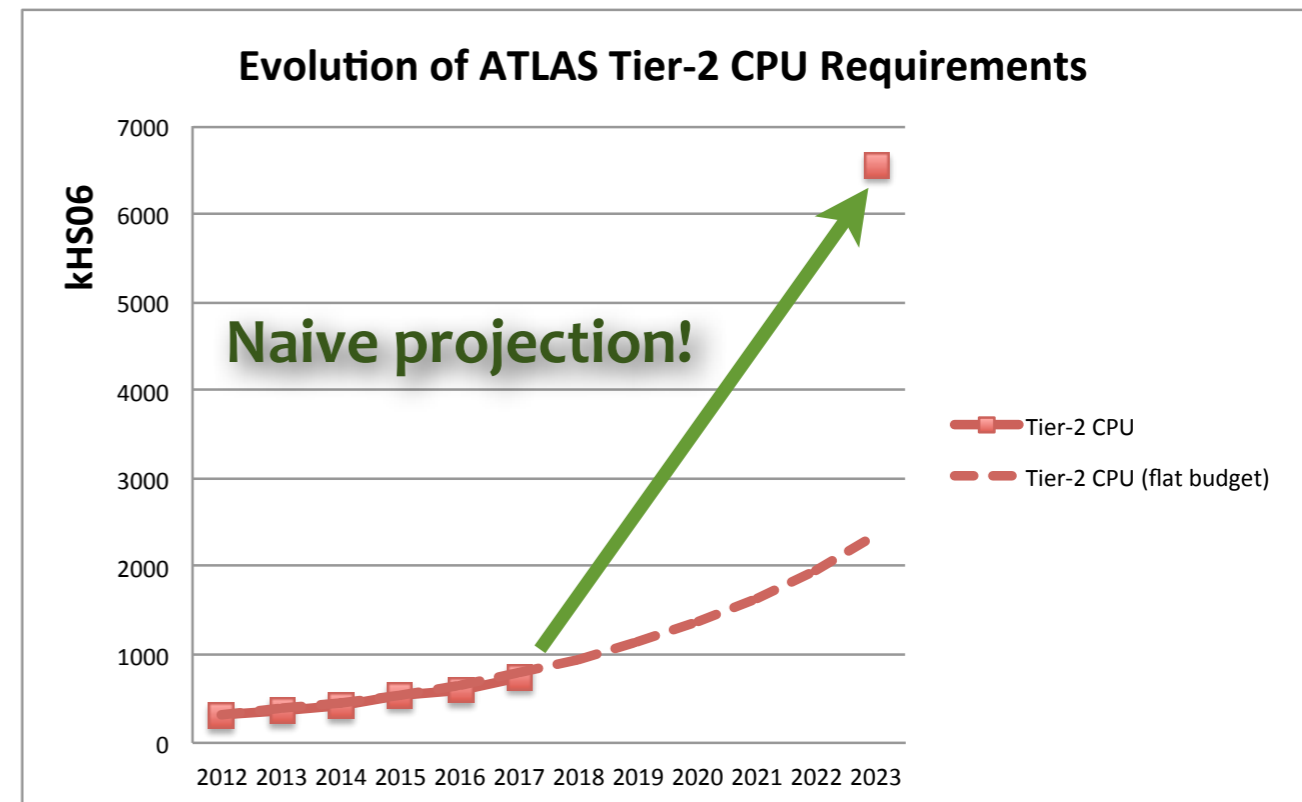
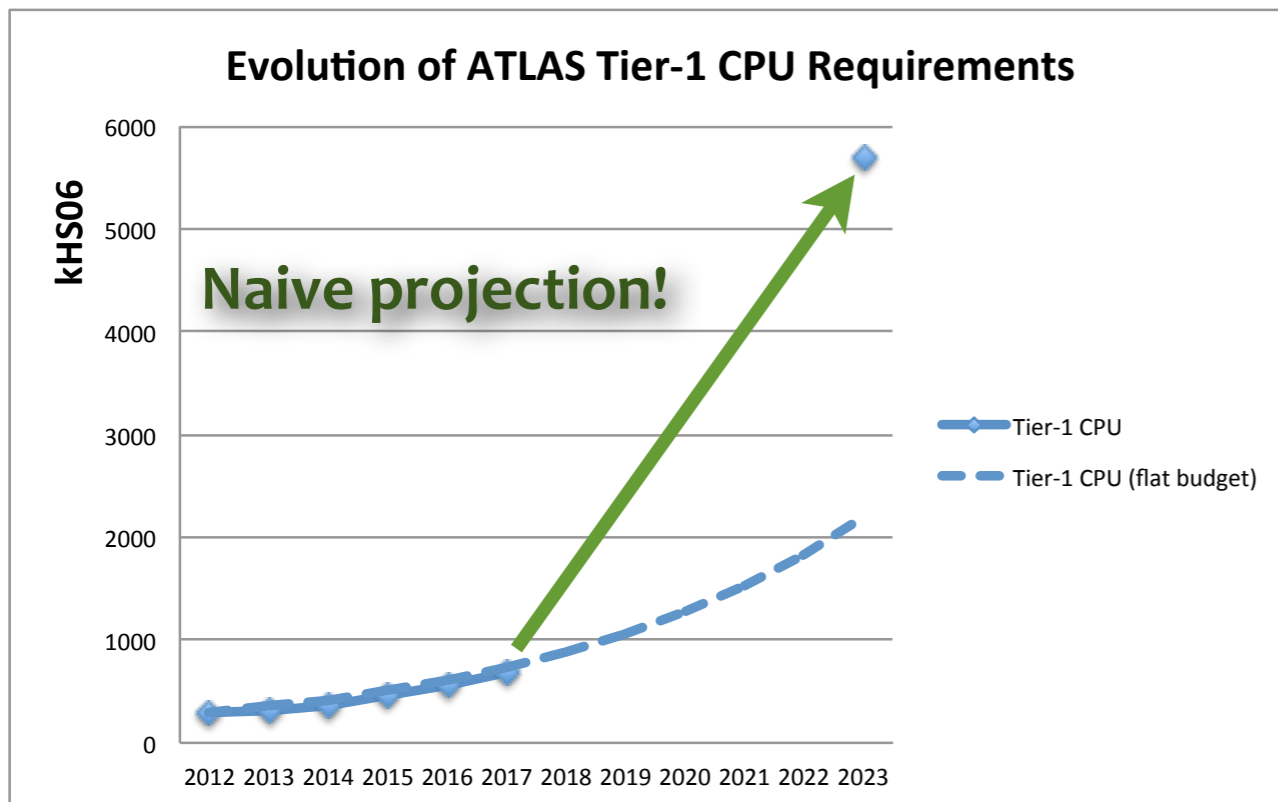
We need to keep the event record sizes and CPU/event for reconstruction on the same level as Run-2 (roughly 2012 values)!

ATLAS S&C 10 years from now..



- Resource increase w.r.t. ‘flat budget’:
 - assuming an order of magnitude increase in data and MC statistics w.r.t. 2017 (Ratio data:MC = 1:1)
 - Assuming same parameters (CPU, event size) as in 2017 (**means work!**)

CPU: A drastic deviation from ‘flat budget’! - We will not get it (probably)....



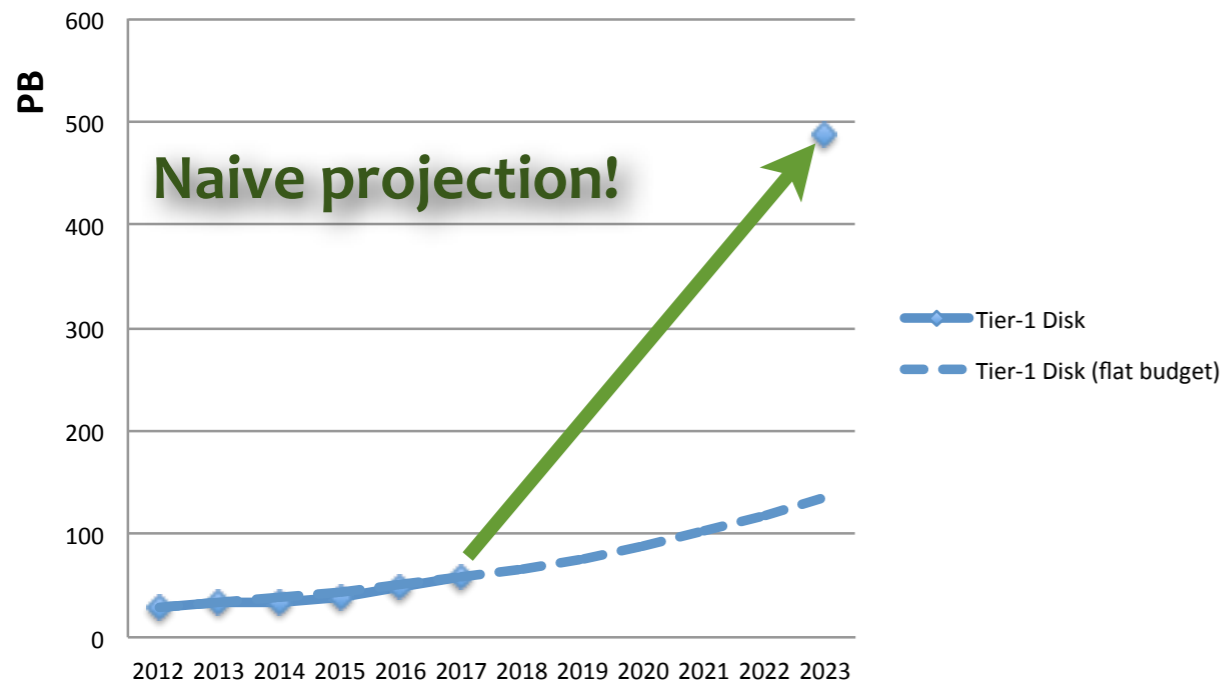
ATLAS S&C 10 years from now..



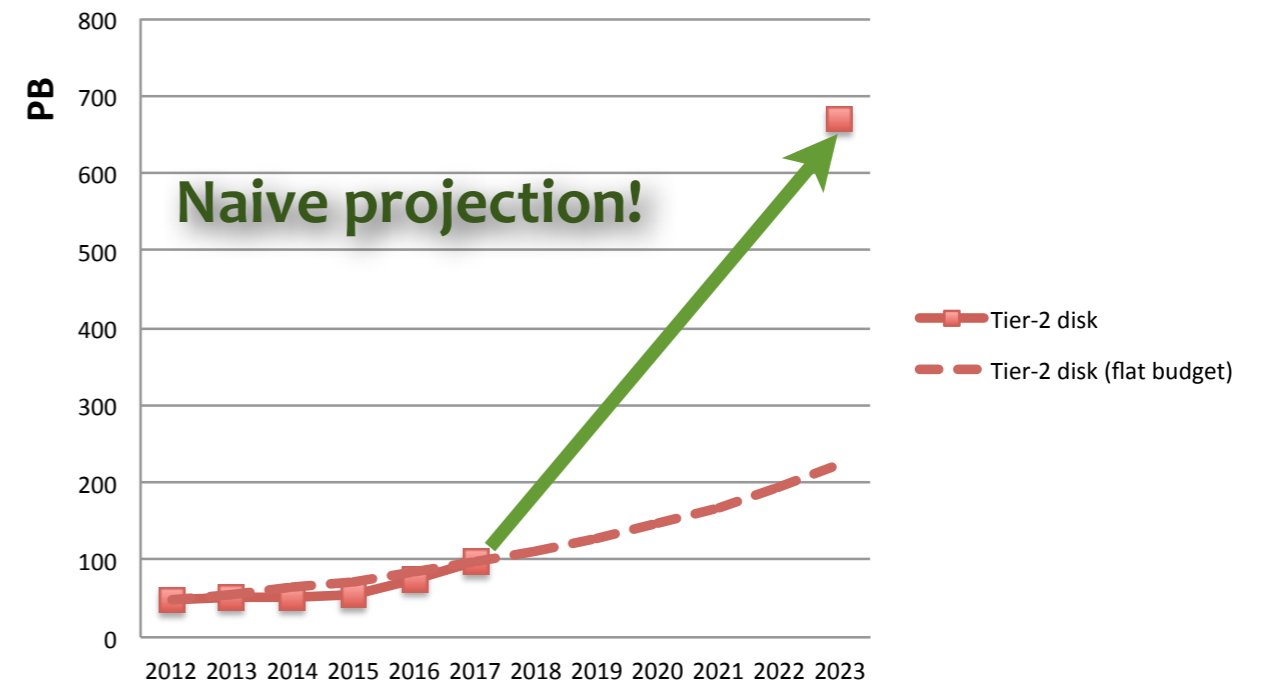
- Resource increase w.r.t. ‘flat budget’:
 - assuming an order of magnitude increase in data and MC statistics w.r.t. 2017 (Ratio data:MC = 1:1)
 - Assuming same parameters (CPU, event size) as in 2017 (**means work!**)

Disk: A drastic deviation from ‘flat budget’! - We will not get it (probably)....

Evolution of ATLAS Tier-1 Disk Requirements



Evolution of ATLAS Tier-2 Disk Requirements

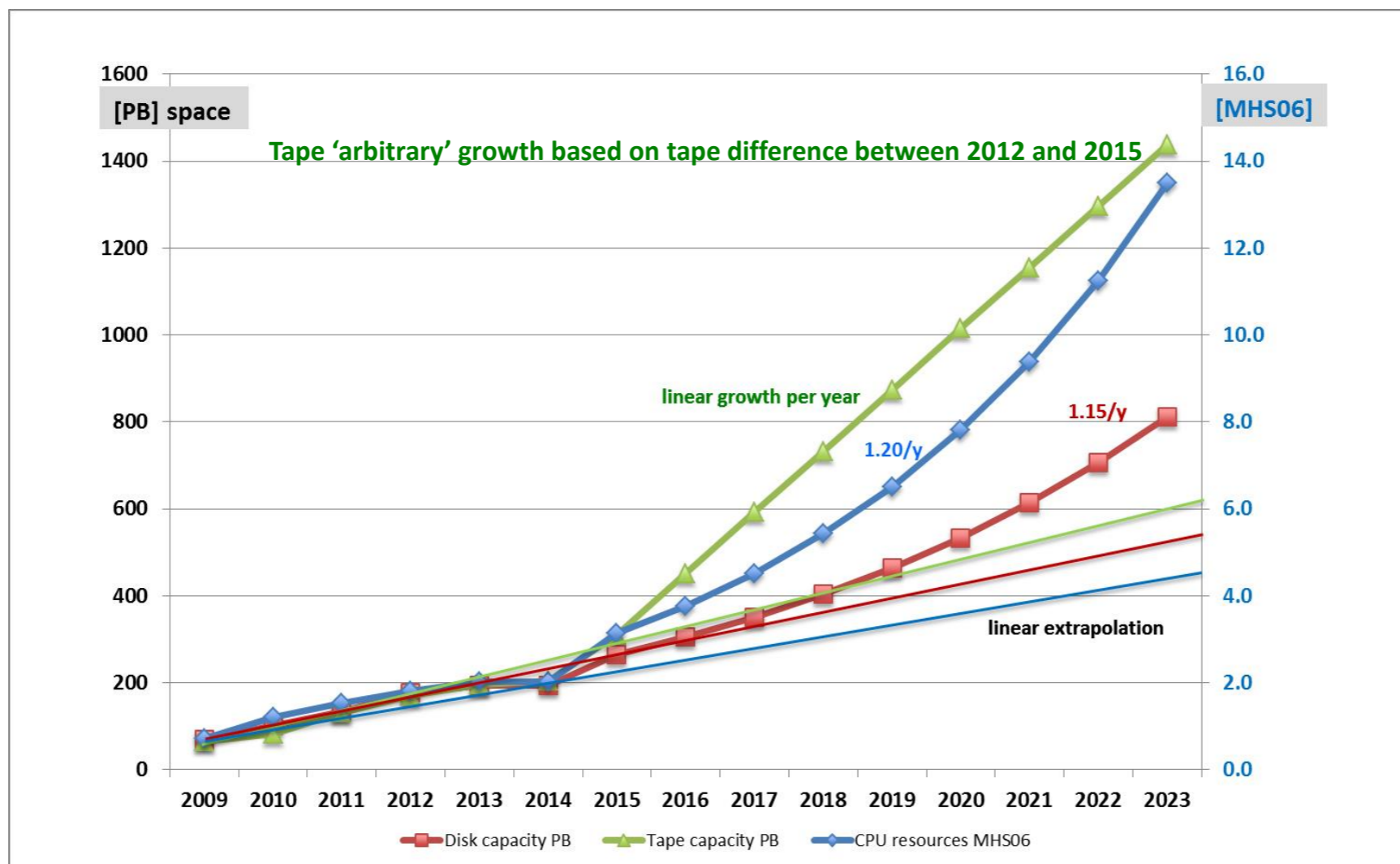


We need drastic improvements to our computing model and new tools to handle it!

'Flat Budget Interpretation'



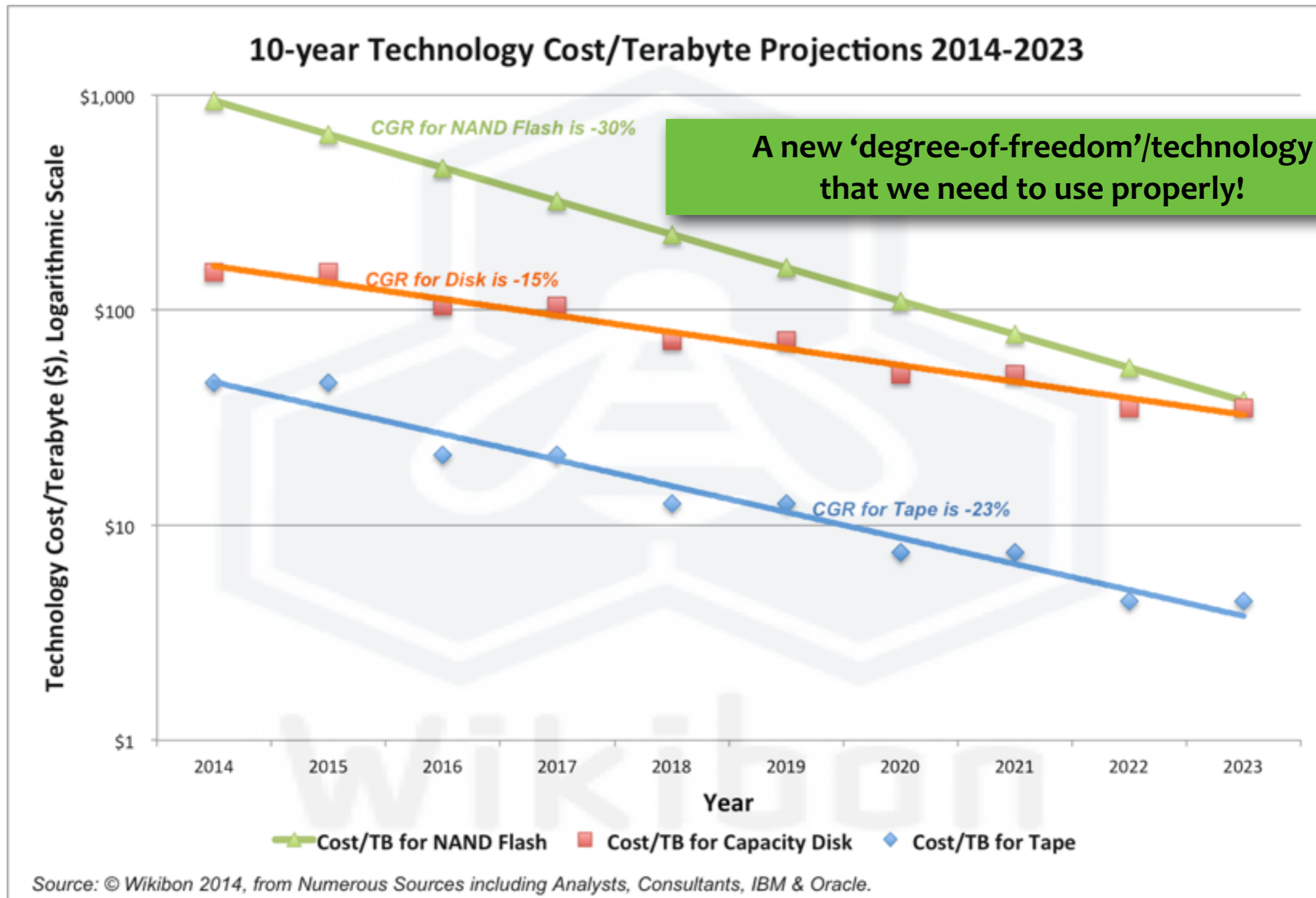
- The 'flat budget' resource increase was made by B. Panzer from CERN/IT for the purpose of Experiment Computing models for Run 2, presented to LHCC.
 - evaluated at factors of **1.2/year for CPU** and **1.15/year for disk**.
 - this will also appear in the technology chapter of the LHCC document, the plot is taken from there: <http://cds.cern.ch/record/1695401>



'Flat Budget Interpretation'

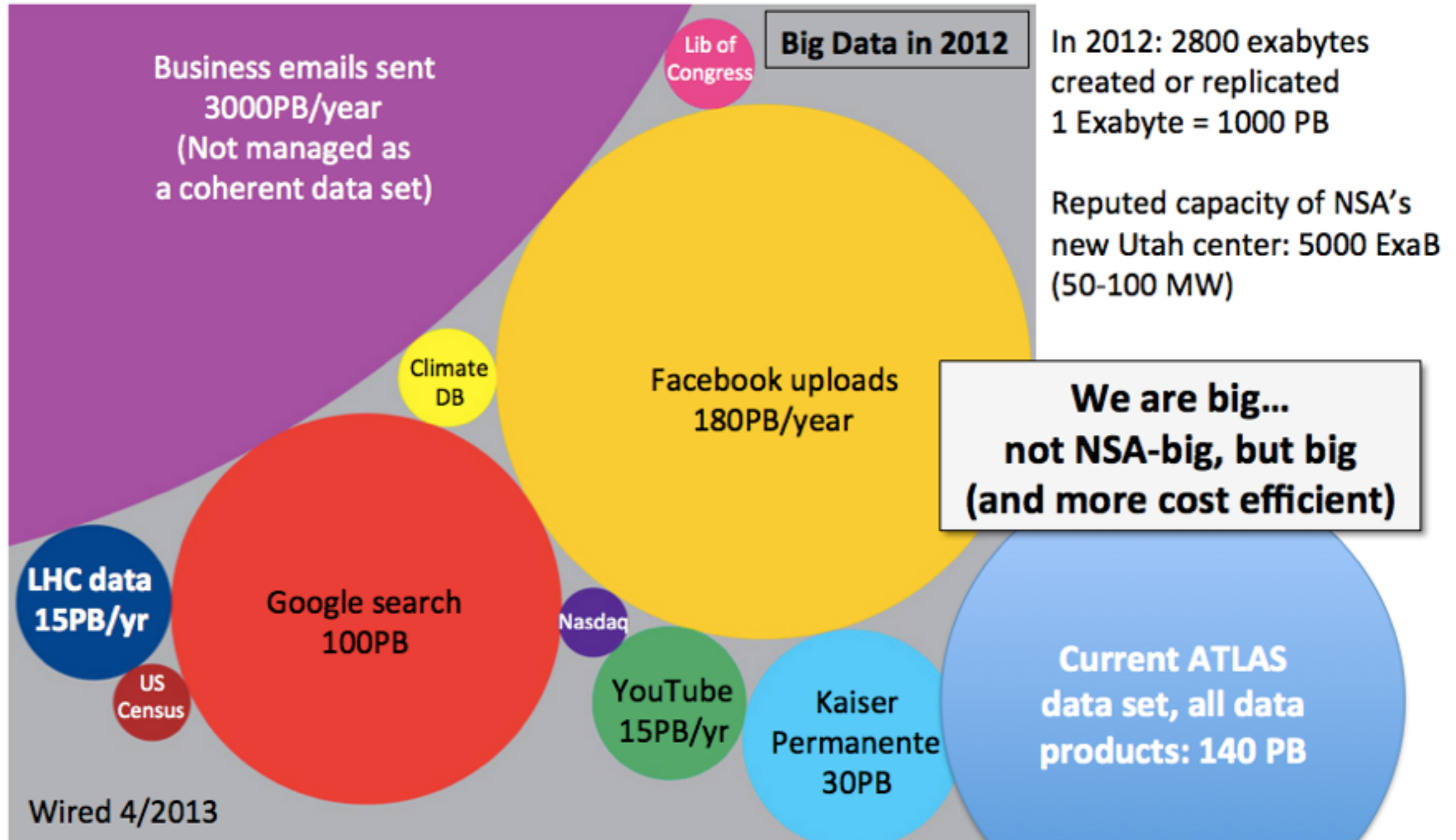


- This in fact corresponds quite well to commercial studies:



Comparing to the World Outside

Data Management Where is HEP in Big Data Terms?



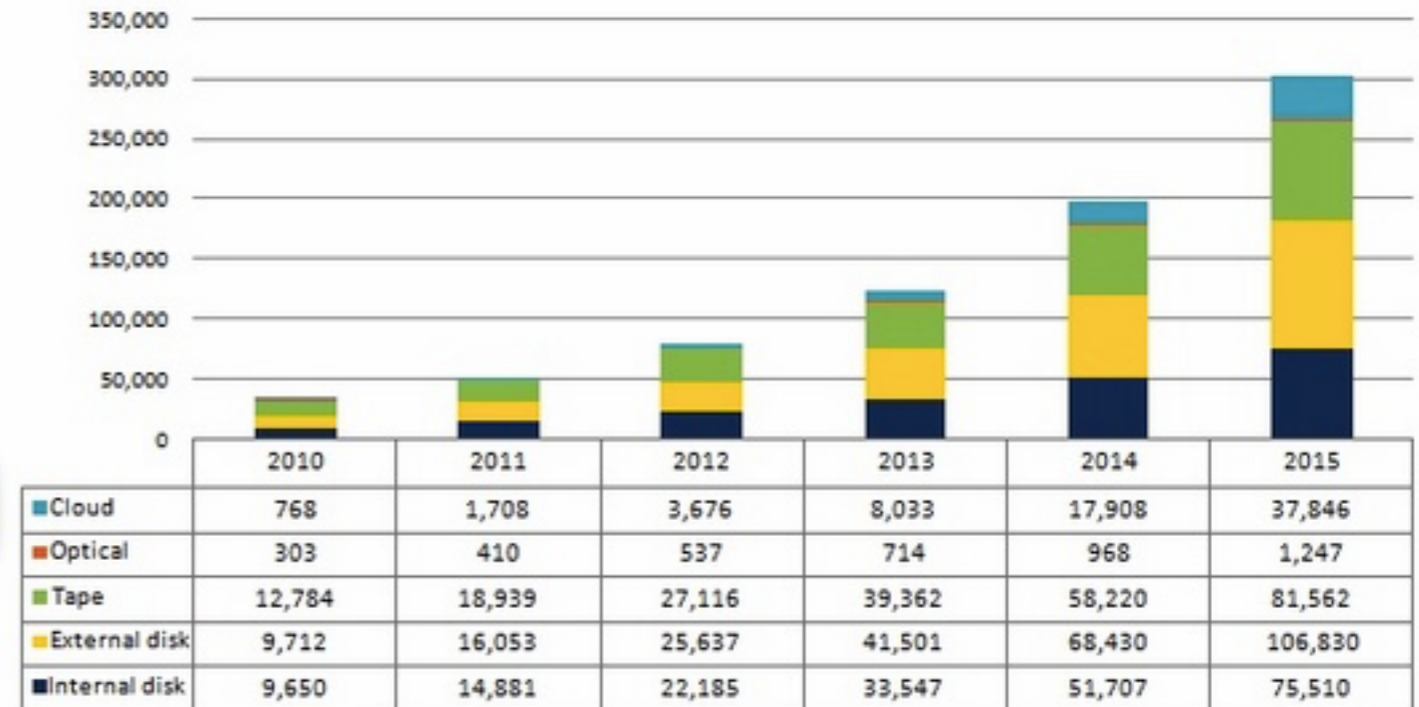
<http://www.wired.com/magazine/2013/04/bigdata/>

Comparing to the World Outside



- Looking at the trends, the world gained an **order of magnitude of storage** over the last five years..

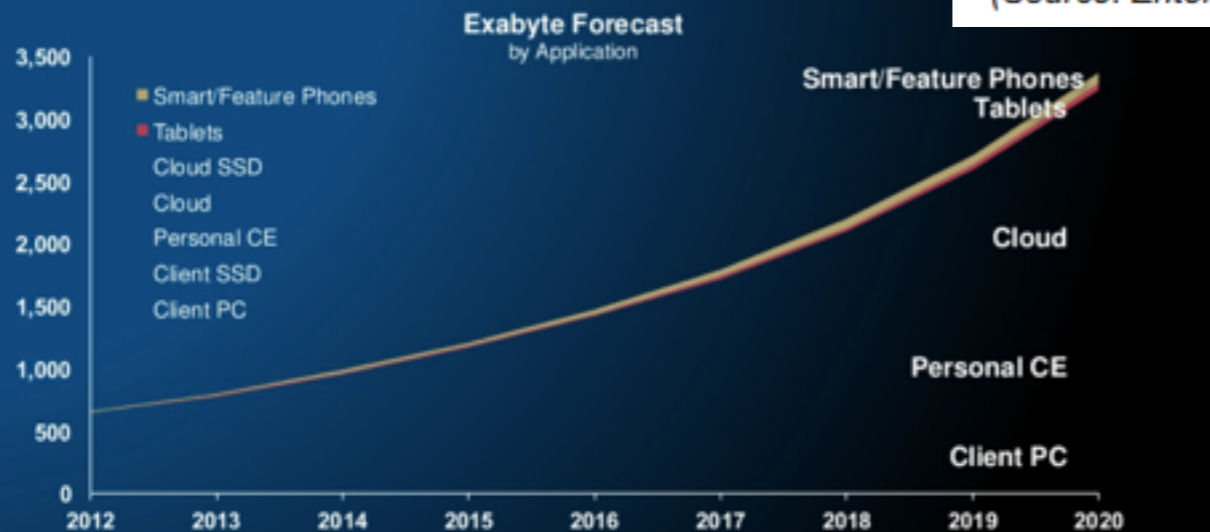
Total WW Digital Archive Capacity
by Media, 2010-2015
(Petabytes)



And the trend is expected to continue

(Source: Enterprise Strategy Group)

2012–2020 Storage Trends



Of the exabytes stored on HDDs and SSDs, over 85% in 2020 will be stored on HDDs

Source: HGST analysis

... an interesting point is that tapes are a viable technology in all projections.

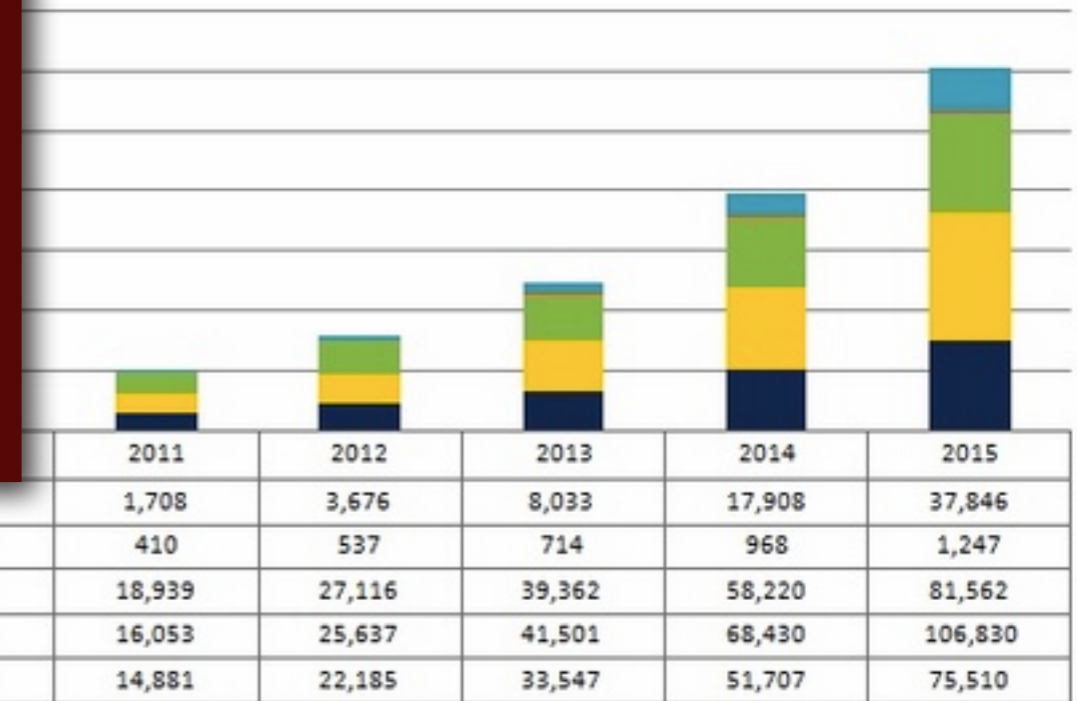
Comparing to the World Outside



- Looking at the trends, the world gained an **order of magnitude of storage** over the last five years..

The punch line here is that the data storage of HL-LHC experiments will NOT become a 'trivial' problem ten years from now.

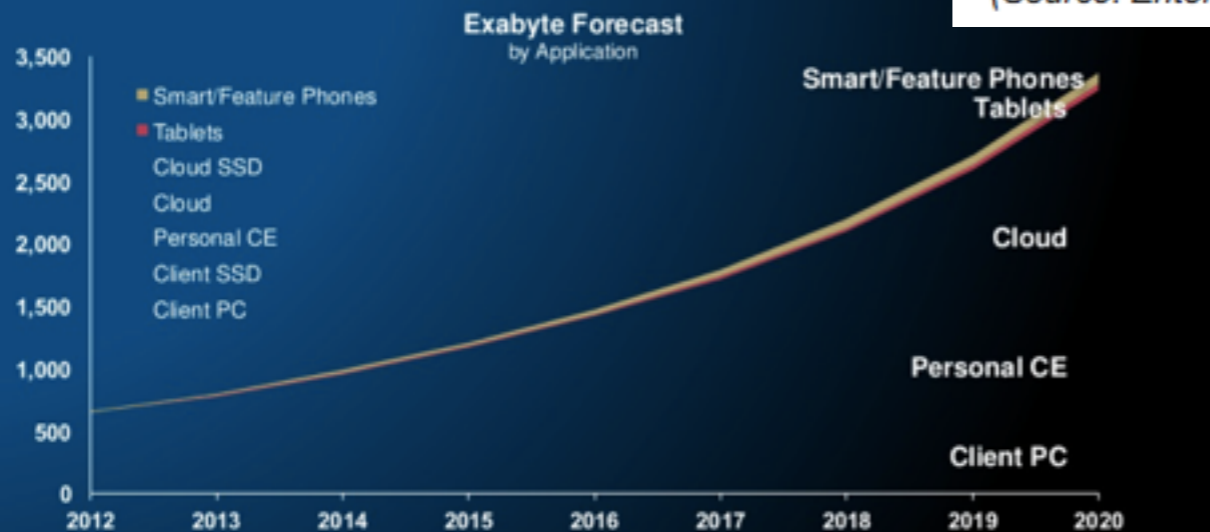
Total WW Digital Archive Capacity by Media, 2010-2015 (Petabytes)



And the trend is expected to continue

(Source: Enterprise Strategy Group)

2012–2020 Storage Trends

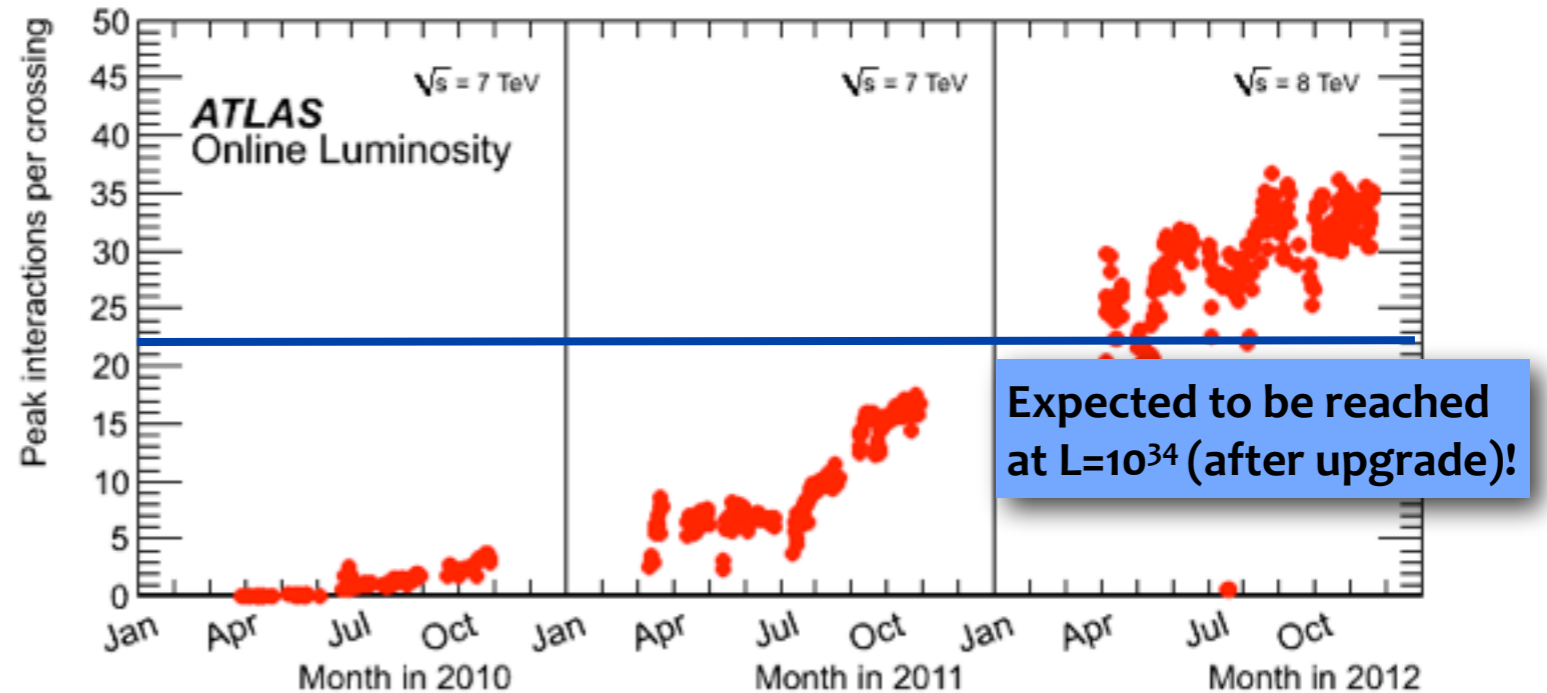
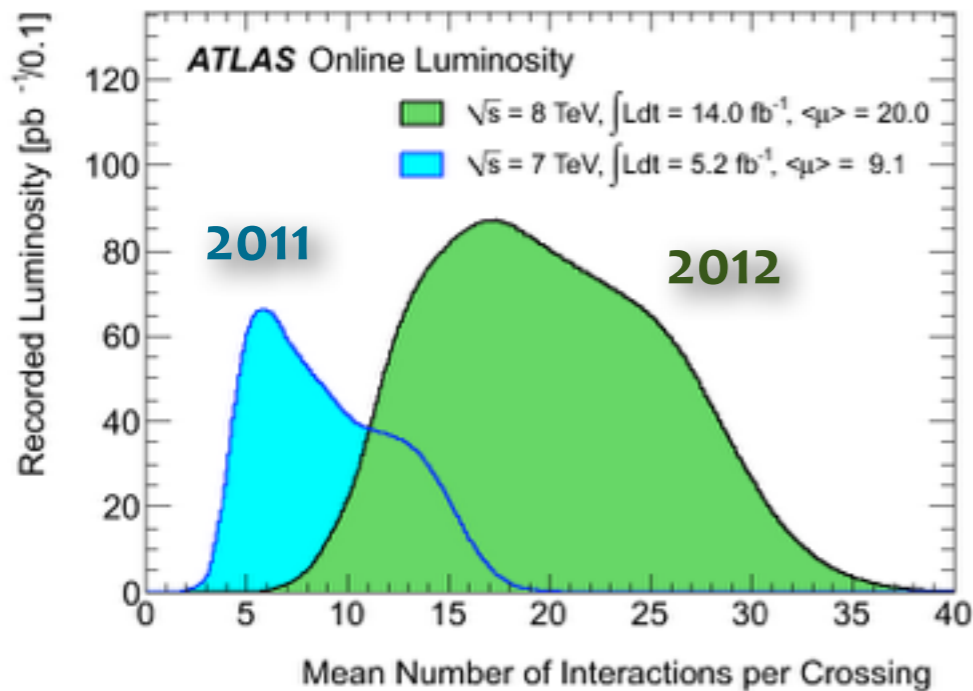


Of the exabytes stored on HDDs and SSDs, over 85% in 2020 will be stored on HDDs

Source: HGST analysis

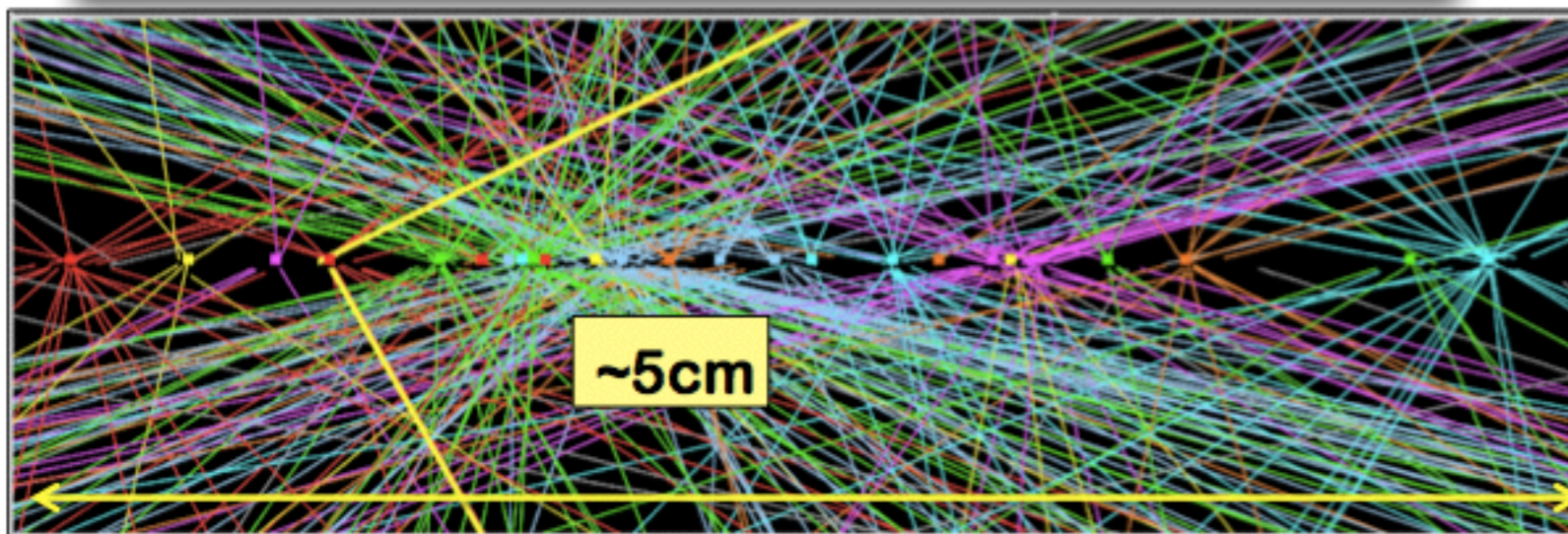
... an interesting point is that tapes are a viable technology in all projections.

A Challenge in 2012: Mastering Pile-up



$$\mu = L \times \sigma_{inel} / (n_{bunch} \times f_r)$$

Z → μμ event from 2012 data with 25 reconstructed vertices!

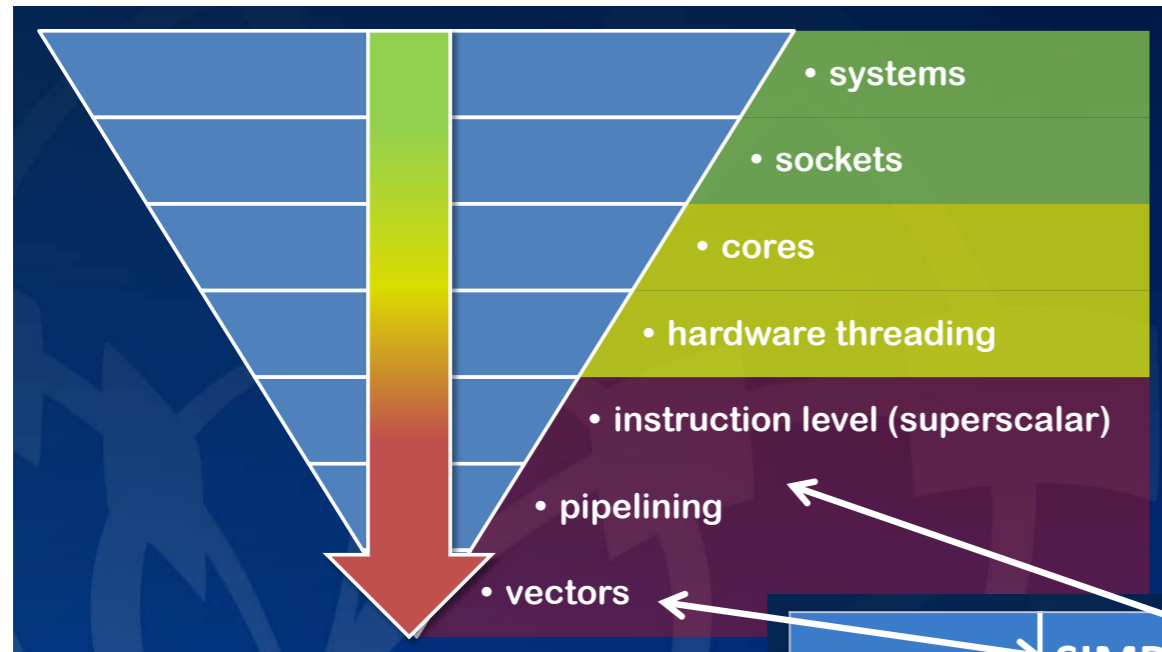


It will get MUCH WORSE in HL-LHC!

Huge efforts invested until now (years!) to minimize the impact of pile-up on physics:

- Develop robust fast triggers.
- Optimize reconstruction and identification of physics objects.
- Precise modeling of pile-up in simulation.
- Improving computing model to handle 2x trigger rate and 2x event size.

Adapting to Modern CPUs



Gains from the different levels of parallelism are multiplicative

Lower ones are harder to use by software

Andrzej Nowak / CERN OpenLab

Efficiency of CPU usage on new hardware: HEP reaches few percent of the speedup gained by fully optimized code or by typical code

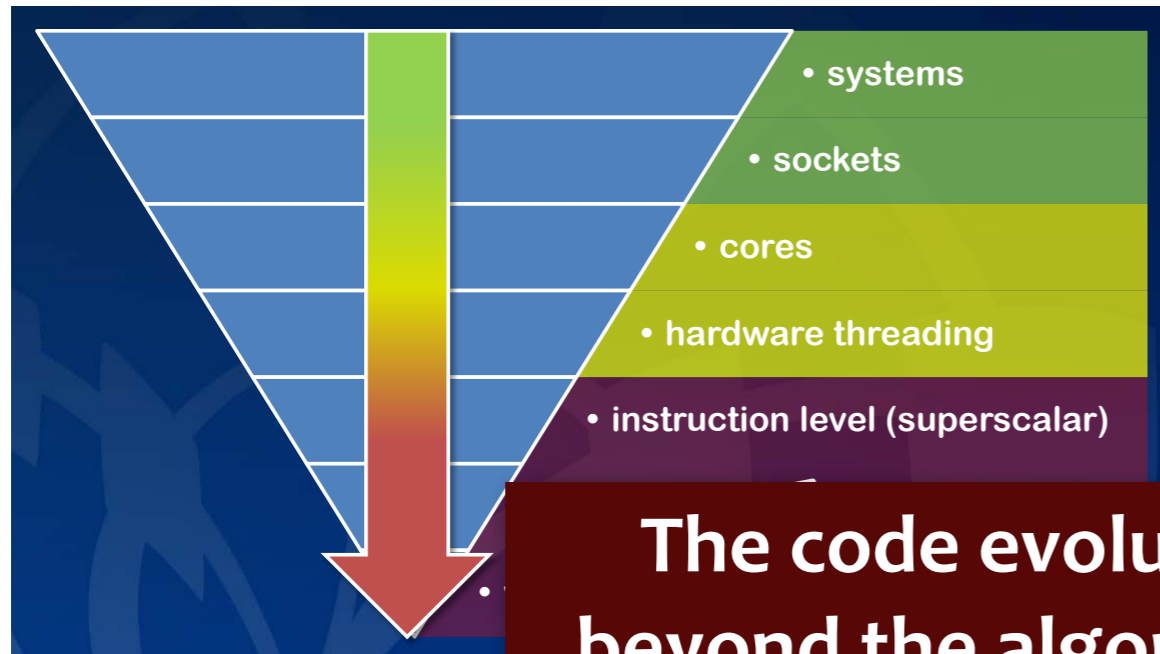
Omnipresent memory limitations hurt HEP - to be overcome first

	SIMD	ILP	HW THREADS	CORES	SOCKETS
MAX	4	4	1.35	8	4
TYPICAL	2.5	1.43	1.25	8	2
HEP	1	0.80	1	6	2

	SIMD	ILP	HW THREADS	CORES	SOCKETS
MAX	4	16	21.6	172.8	691.2
TYPICAL	2.5	3.57	4.46	35.71	71.43
HEP	1	0.80	0.80	4.80	9.60

... well, the code is written (mostly) by physicists, for physics

Adapting to Modern CPUs



Gains from the different levels of parallelism are multiplicative

Lower ones are harder to use by software

The code evolution needs to go beyond the algorithm optimization and incorporate massive parallelization, vectorization & co. to be able to utilize the new CPU (& GPU!) technologies.

Efficiency of CPU on new hardware reaches few percent the speedup gain by fully optimized code by typical code

Omnipresent memory limitations hurt HEP - to be overcome first

CERN OpenLab

CORES	SOCKETS
8	4
8	2
6	2

	SIMD	ILP	HW THREADS	CORES	SOCKETS
MAX	4	16	21.6	172.8	691.2
TYPICAL	2.5	3.57	4.46	35.71	71.43
HEP	1	0.80	0.80	4.80	9.60

... well, the code is written (mostly) by physicists, for physics

Adapting to Modern CPUs



The story becomes even more complicated when we add the (GP)GPUs and many-core (XeonPhi) architectures into the game.

Computing Growth is Not Just an HPC Problem

Microprocessor Performance "Expectation Gap" over Time (1985-2020 projected)

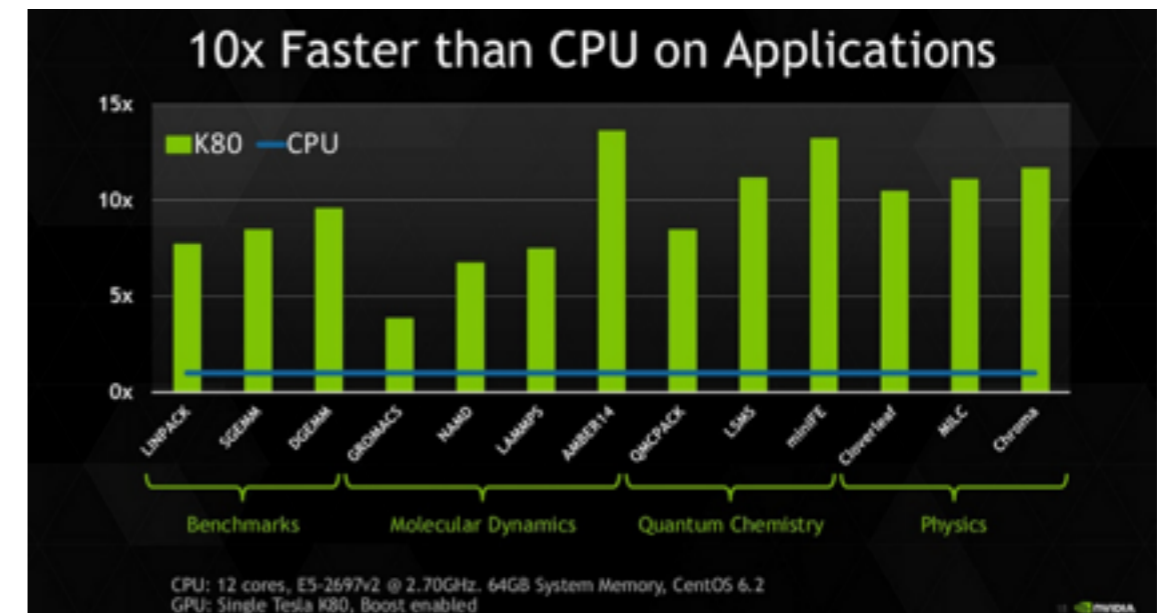


TESLA

WHAT IS GPU COMPUTING? GPU APPLICATIONS SERVERS AND WORKSTATIONS

NVIDIA Home - Products - High Performance Computing

NVIDIA TESLA P100
Unleashing Infinite Compute for the Modern Data Center



1997: THE FIRST INTEL® TERAFLUP COMPUTER consisted of: **9,298** INTEL PROCESSORS and occupied: **72** SERVER CABINETS

THE INTEL® XEON® PHI™ COPROCESSOR will provide: **1** TERAFLUP OF PERFORMANCE and occupy: **1** PCIe SLOT



[Click to learn more](#)

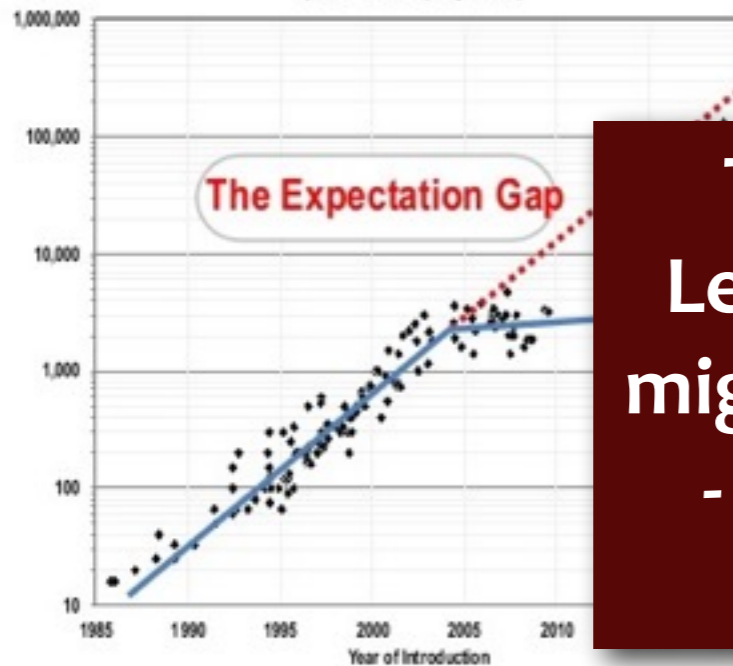
Adapting to Modern CPUs



The story becomes even more complicated when we add the (GP)GPUs and many-core (XeonPhi) architectures into the game.

Computing Growth is Not Just an HPC Problem

Microprocessor Performance "Expectation Gap" over Time (1985-2020 projected)



These boost the 'PFLOPs' of the Leadership HPC facilities today and might be part of every facility in 2026 - we need to be able to use them efficiently!

TESLA

WHAT IS GPU COMPUTING? GPU APPLICATIONS SERVERS AND WORKSTATIONS

NVIDIA Home - Products - High-Performance Computing

NVIDIA TESLA P100

on Applications

Application	Performance (approx.)
Benchmarks	100
Molecular Dynamics	200
Quantum Chemistry	300
Physics	250

CPU: 12 cores, E5-2697v2 @ 2.70GHz, 64GB System Memory, CentOS 6.2
GPU: Single Tesla K80, Boost enabled



[Click to learn more](#)

1997: THE FIRST INTEL® TERAFLUP COMPUTER consisted of: **9,298** INTEL PROCESSORS and occupied: **72** SERVER CABINETS

THE INTEL® XEON® PHI™ COPROCESSOR will provide: **1** TERAFLUP OF PERFORMANCE and occupy: **1** PCIe SLOT



Networking - probably good news...




- Networking is the one item that will most probably continue its progress & evolution further..
 - In terms of bandwidth increase.
 - In terms of new technologies (NaaS - Network as a (virtual) Service ?)



**Software-Defined Networks (SDN):
Bridging the application-network divide**

Inder Monga
Chief Technologist and Area Lead,
Engineering, Research and Software development

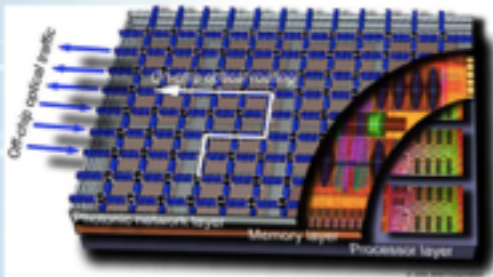
CHEP 2013



**Advanced Networking for HEP,
Research and Education
in the LHC Era**

Harvey B Newman
California Institute of Technology

A fun peek into the future...just imagine


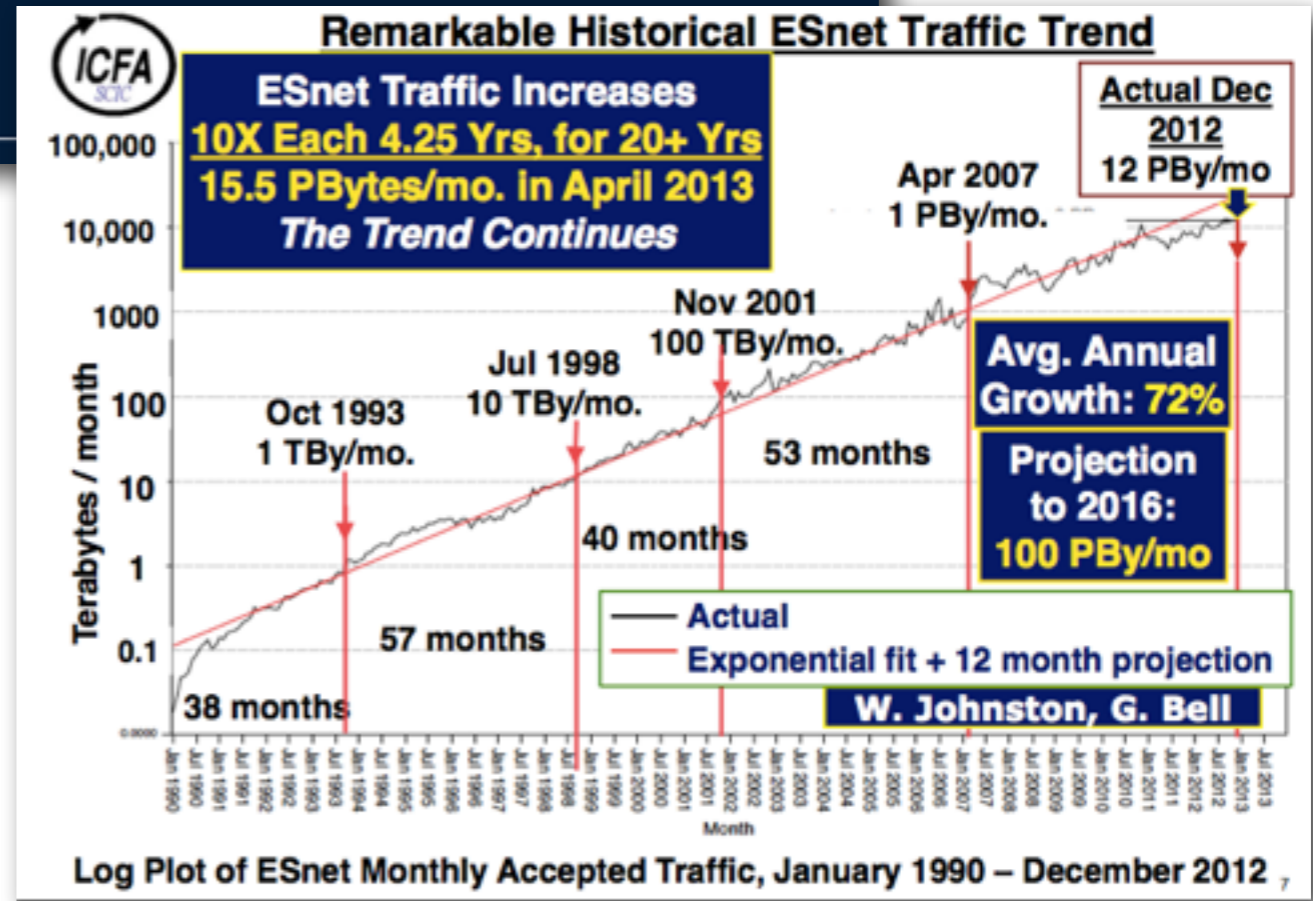


With silicon photonics integration, each chip will have a network interface
That implies each chip could be network addressable

If so, we could design servers without needing NIC cards – no difference between communication within the motherboard or outside.

With HEP applications like FAX, file systems or memory can be mounted remotely to my chip while 'streaming data for analysis.'

With SDN, can effectively route IP and non-IP protocols (like ROCE)
SDN could revolutionize how computing is done, are we ready for that?

Networking - probably good news...

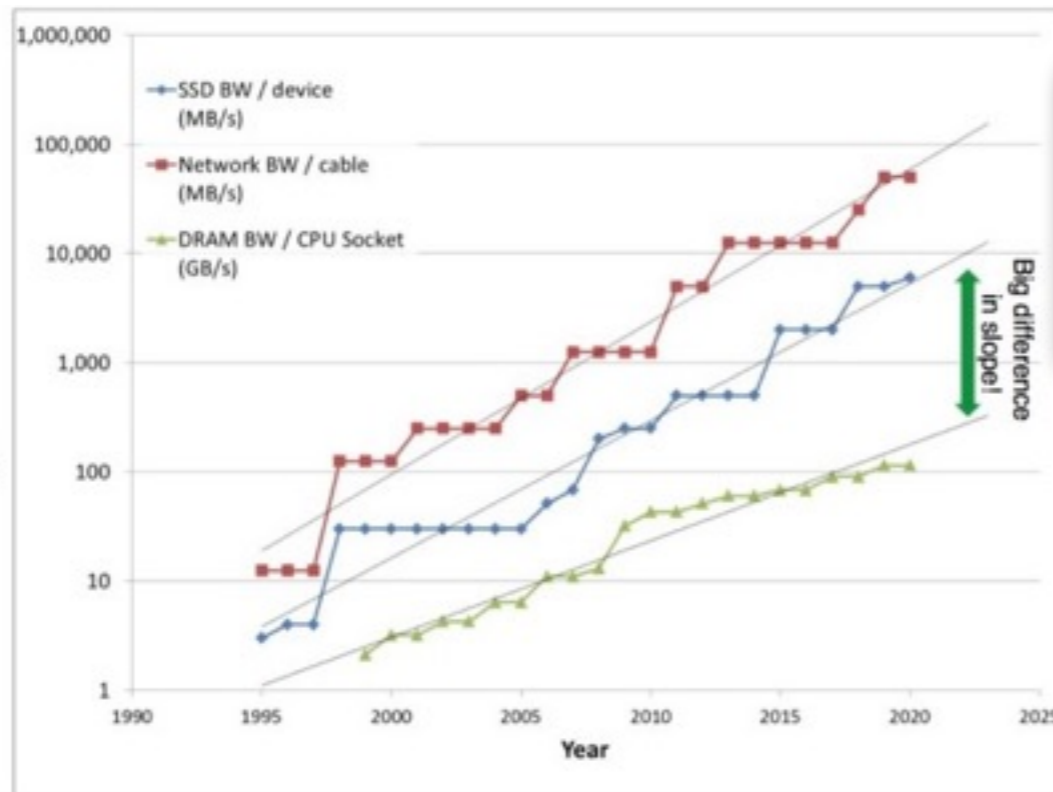


- When looking at network bandwidths w.r.t. SSD and DRAM, the future limiting factor for us might in fact be DRAM/CPU ... to be explored!

Network, Storage, & DRAM trends

Log scale

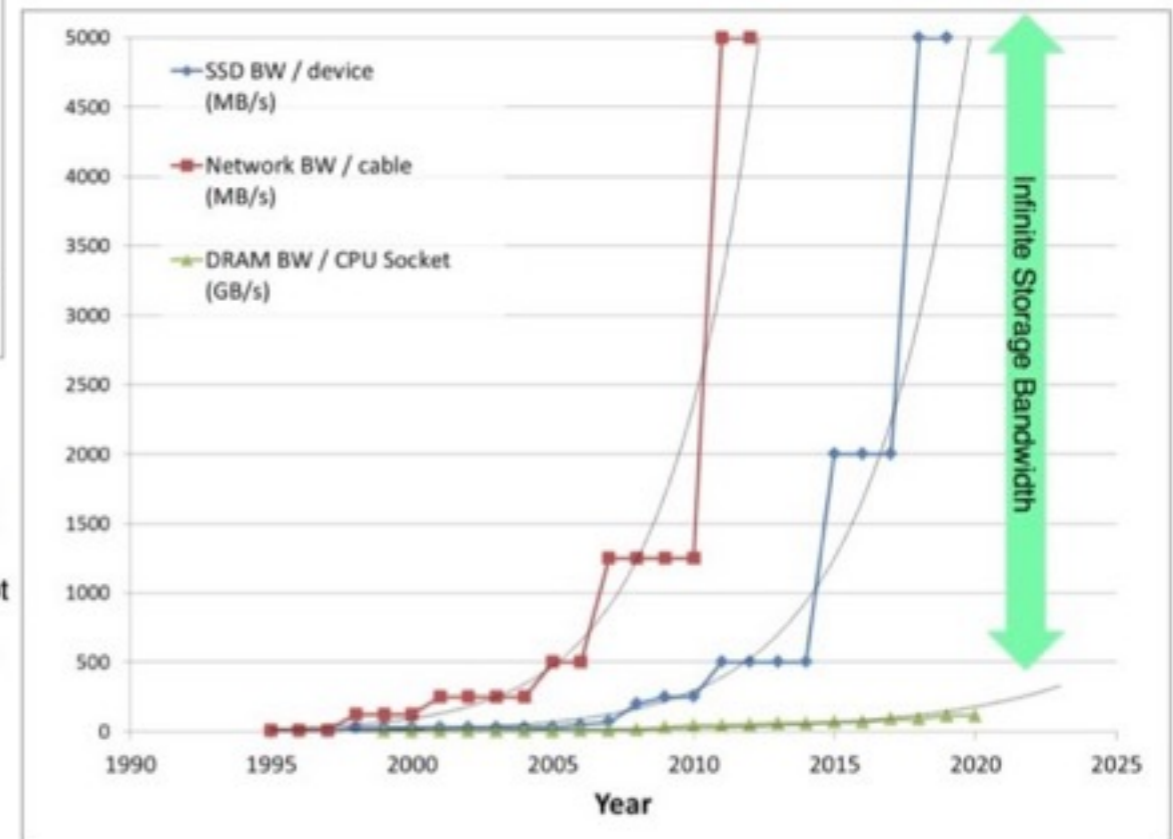
- Use DRAM Bandwidth as a proxy for CPU throughput
- Reasonable approximation for DMA and poor cache performance workloads (e.g. Storage)



LHC experiments are already reading files ~directly over WAN (xrootd protocol...) The factor is also the robustness/failure rate, not just the speed...

Linear scale

- Same data as last slide, but for the Log-impaired
- Storage Bandwidth is not literally infinite
- But the *ratio* of Network and Storage to CPU throughput is widening very quickly



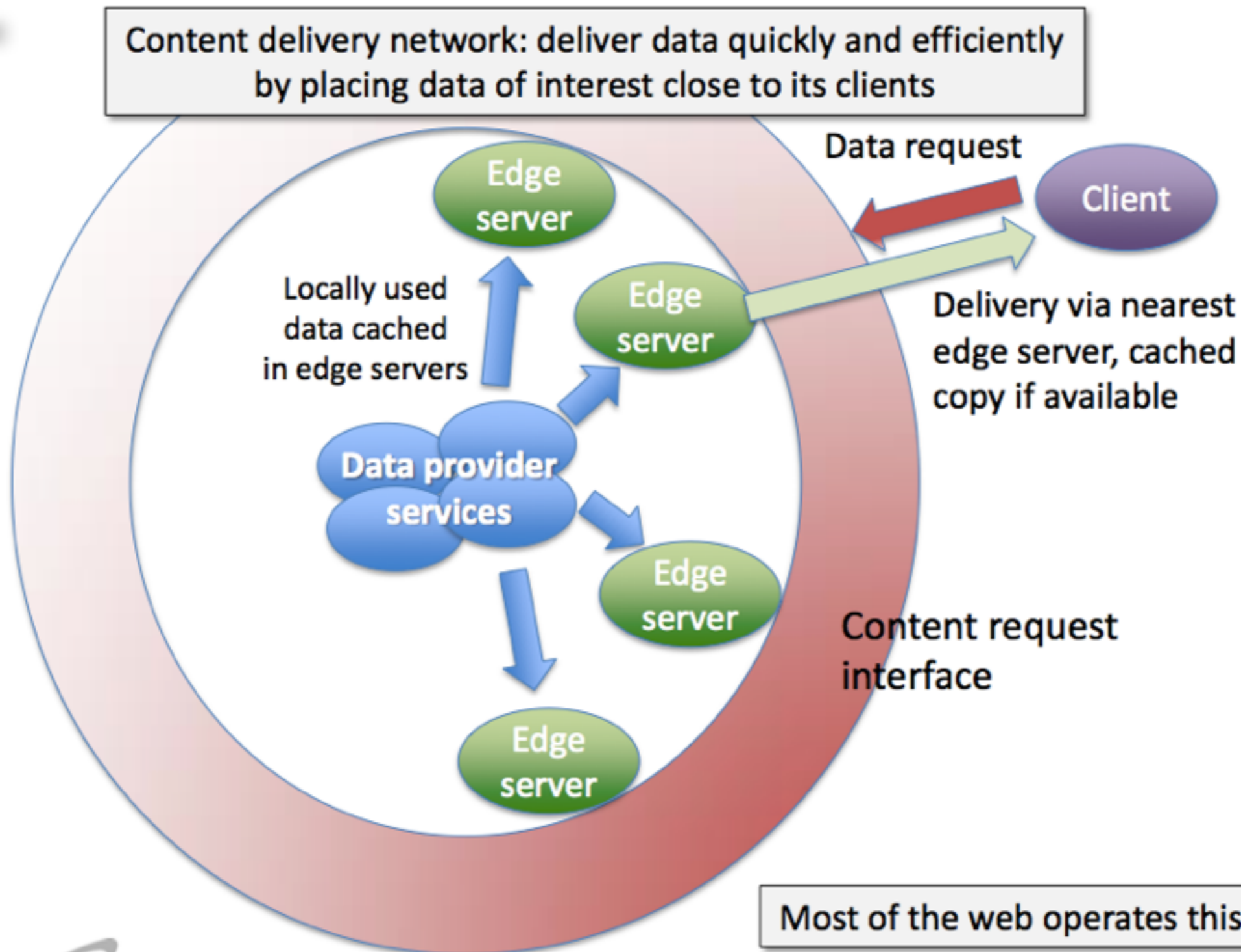
<https://itblog.sandisk.com>

More on Future Data Access...



The Content Delivery Network Model

T. Wenaus



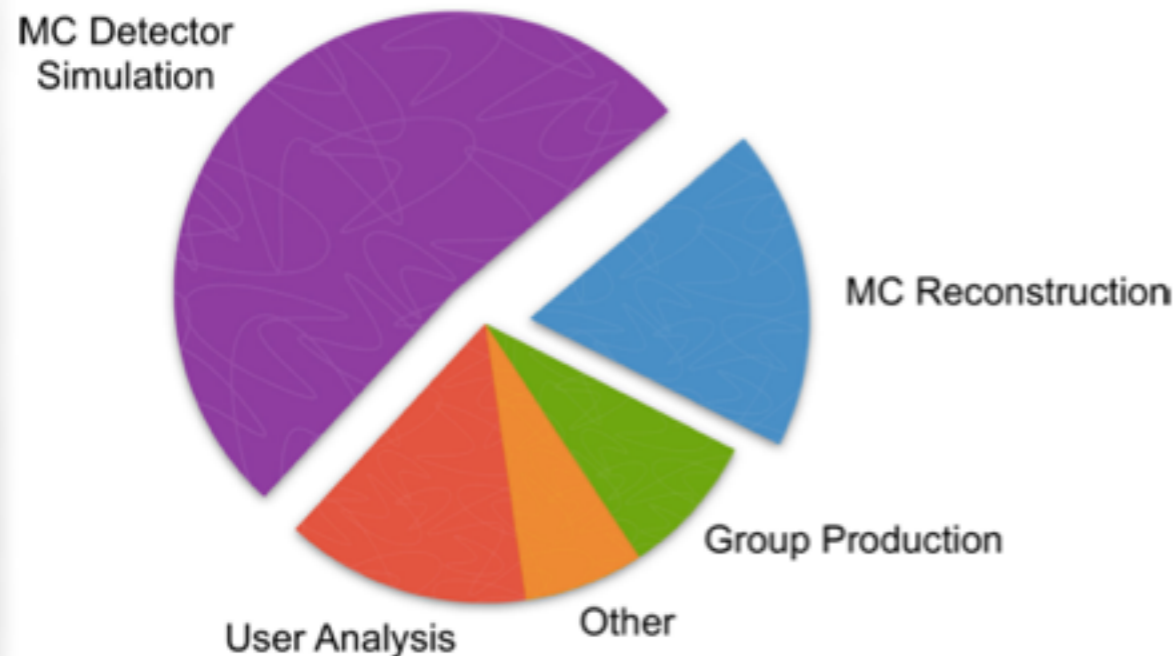
A Note on Simulation



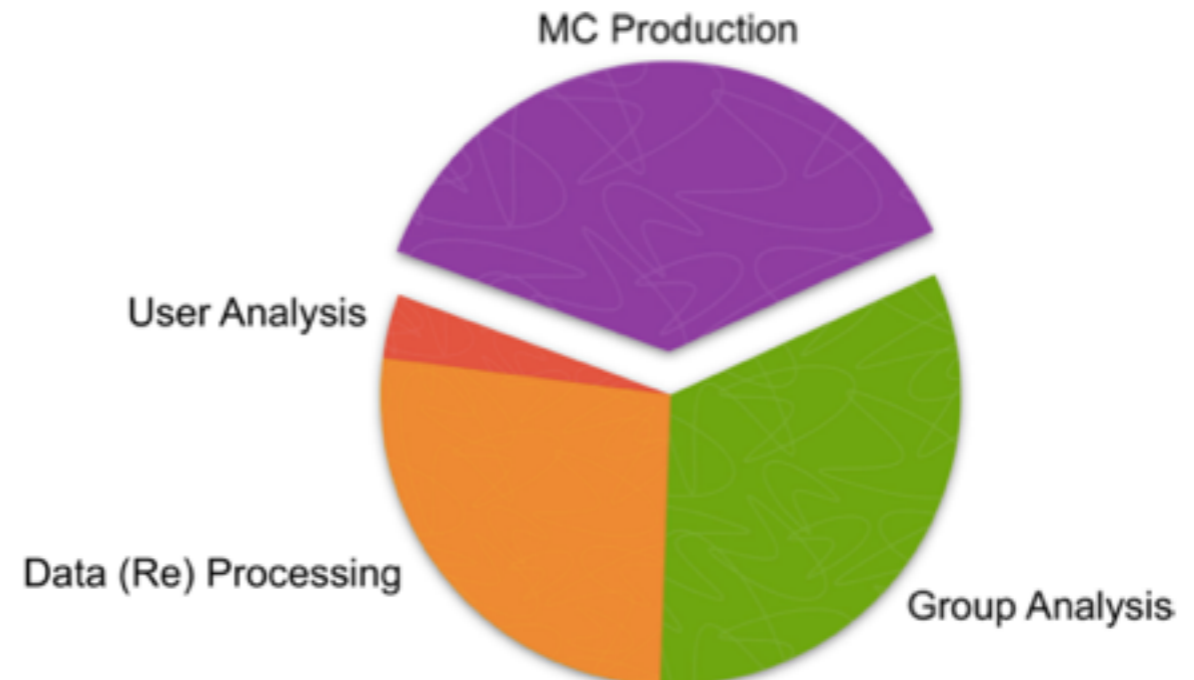
- ▶ Grid CPU usage dominated by MC production
- ▶ MC production takes up large fraction of Grid disk usage => **limitation**
- ▶ Precise detector simulation => highly CPU intensive
- ▶ Obstacle for physics analyses in need of large MC statistics => sensitivity limitation
- ▶ Higher luminosity and pileup => larger MC production needed

ATLAS Grid usage in 2012

ATLAS grid CPU utilization:



ATLAS grid disk utilization:



A Note on Simulation

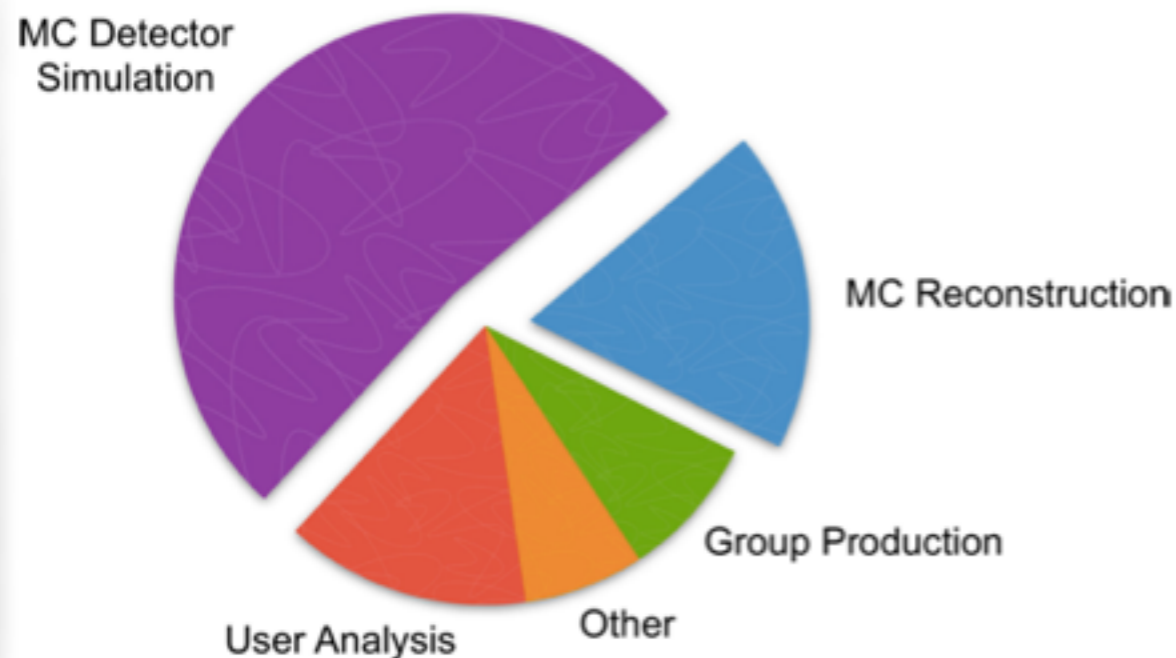


This we did not anticipate properly before LHC started...
Our Grid sites are High Throughput Computing focused...

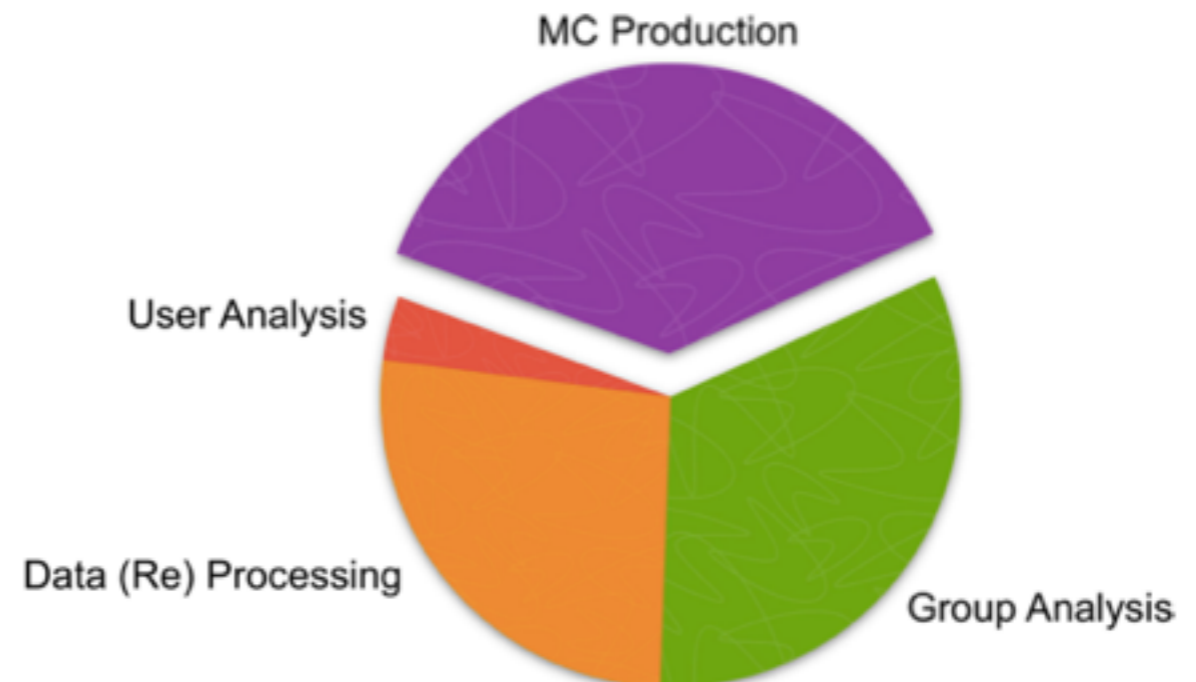
- ▶ Grid CPU usage dominated by
- ▶ MC production takes up large
- ▶ Precise detector simulation =>
- ▶ Obstacle for physics analyses in need of large MC statistics => sensitivity limitation
- ▶ Higher luminosity and pileup => larger MC production needed

ATLAS Grid usage in 2012

ATLAS grid CPU utilization:



ATLAS grid disk utilization:

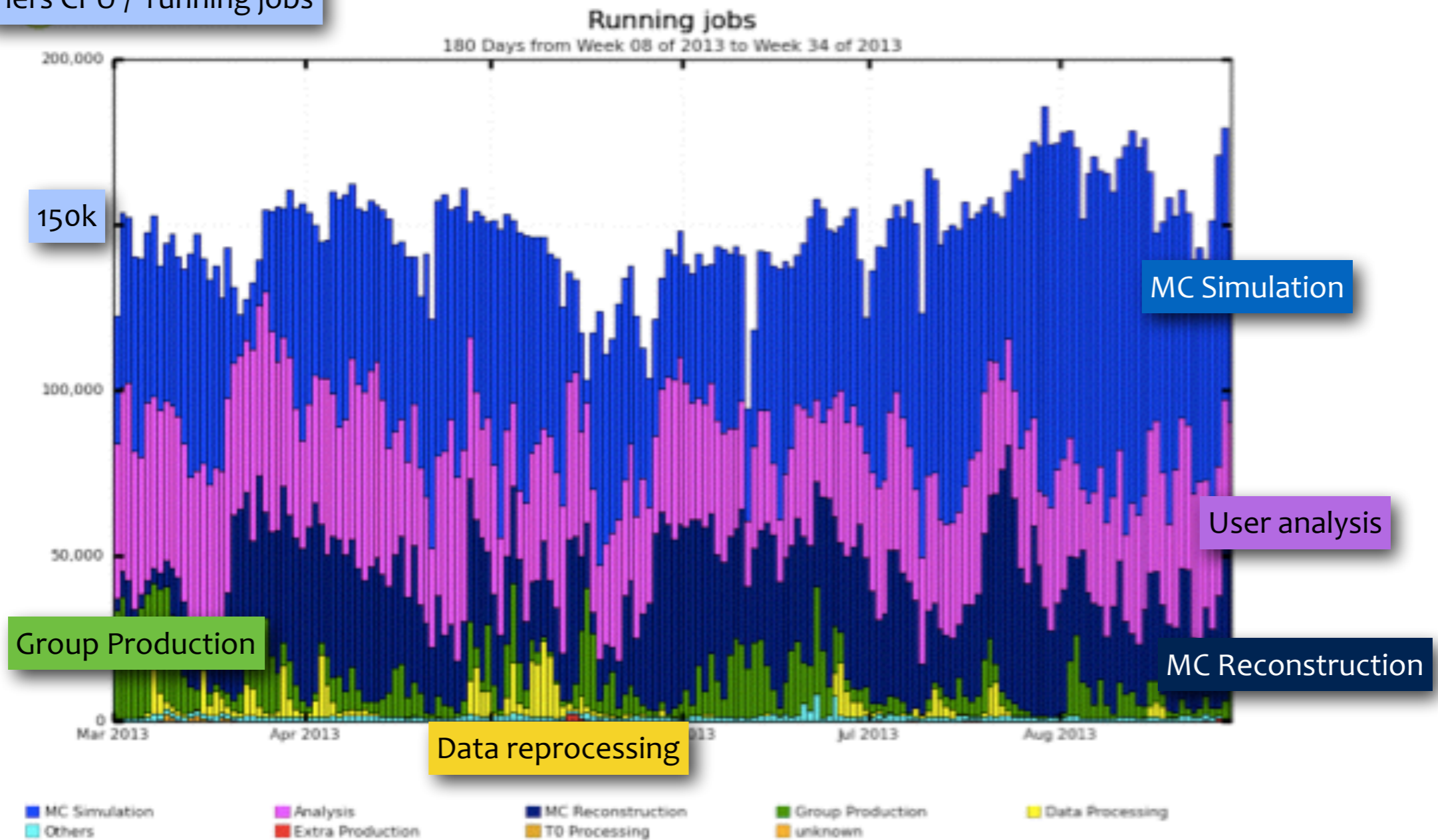


ATLAS Distributed Computing Activities



- Easiest to explain this graphically - example of running jobs on ATLAS distributed resources...

Tiers CPU / running jobs



High Performance Computing!

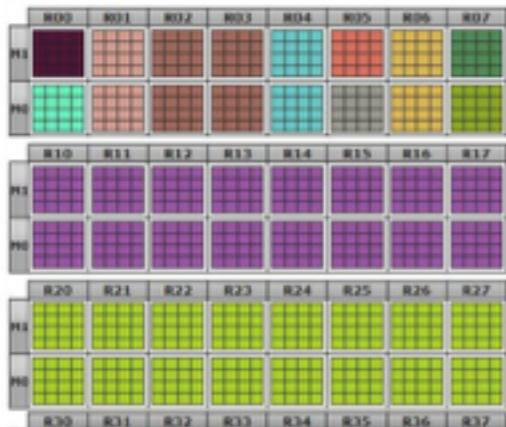


Meet Intrepid



- 40 racks, each with 1024 quad-core compute nodes = 163,840 cores
 - Roughly equivalent to 23,000 x86 cores.
- Attached to several PB of fast data disk
- Operational since 2006 (now that its successor Mira is launching, it's becoming "previous generation")

A Snapshot of Intrepid



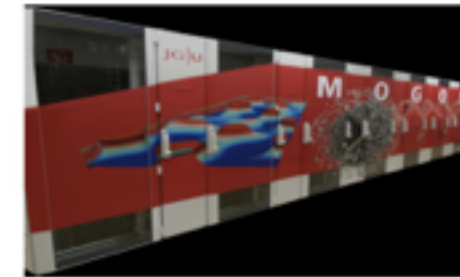
- This is a typical workday view of Intrepid.
- Most of the machine is running big jobs.
- The remainder of the machine is filled by short small jobs.

HPC matches well with two MC simulation steps, simulation of the quantum process and detector response, since they are High CPU & low I/O!

HPC Resource Examples



- SuperMUC, Munich
 - 155,000 Sandy Bridge cores, 2.8M HS06
 - ATLAS 2013 T1/2 pledges ~ 730K HS06
 - Suse Enterprise Linux 11, 2GB/core
 - warm water cooling
 - 40°C inlet. 70°C outlet used to heat building
- Hydra, MPI, Munich
 - 'similar' cluster in spec and scale
 - due Summer 2013. 10k core integration system in place now
- MOGON, Mainz
 - 34k cores SL6



How Might This Fit With ATLAS

- Reminder: ATLAS uses ~800M grid CPU-hours
- 7 billion CPU-hours x 6% for Opportunistic Running = 420M CPU-hours
 - I believe we want a DC level of opportunistic running + scheduled peaks
- One can compete for computer time
 - It's the job of these facilities to support DoE science
 - ALCF: "medium" (average award: ~30 million hours)
 - Implementing and Accelerating Geant4-Based Simulations on Titan (PI R. Mount) 10M hours
 - Grid-enabling High Performance Computing for the LHC (PI T. LeCompte) 18M hours
 - Decisions in May
 - ALCF: "large" (up to 430 million hours)
- Are there other computers out there that look attractive?
 - Supercomputers in Texas (Stampede) and Illinois (Blue Waters)
 - SAKURA at KEK (BlueGene/Q)
 - Others...(some Rod will mention)

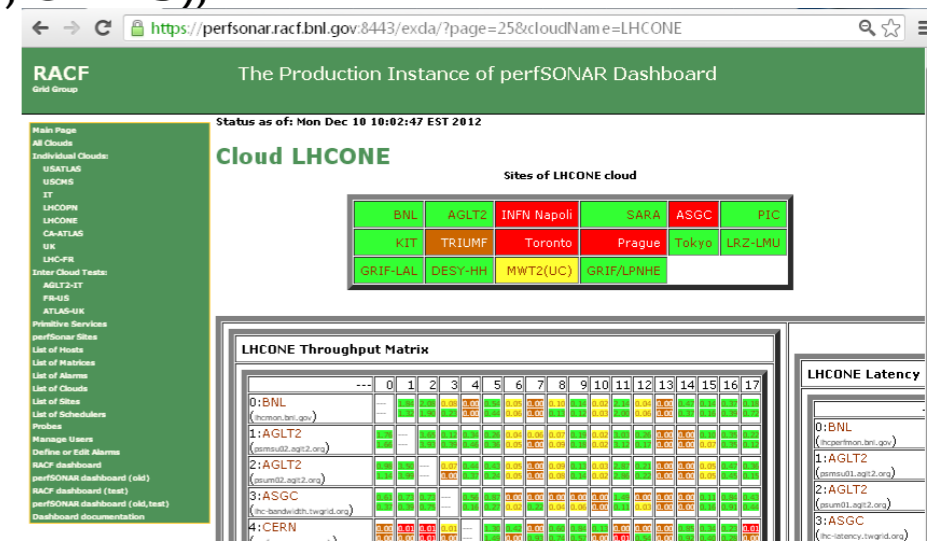
Current Data Processing and Management Tools



- At time of inception no global/commercial solution for the distributed computing (Grid middleware developed) needed for our 'Big Data' handling:
 - Even today, 'Cloud facilities' or HPC centers are not distributed resources, neither in terms of CPU nor storage.
 - We could build a Grid of Cloud facilities but not a Cloud of Grid facilities.
- 'In-house' experiment specific topmost job and data management layers, also tying in different Grid and local (batch/storage..) setup flavors.
 - Distributed Computing Systems (job handling):
AliEn(Alice), PanDA(ATLAS), Crab(CMS), Dirac(LHCb)
 - Distributed Data Management Systems (file placement, replication and access handling):
AliEn(Alice), DQ2/Rucio(ATLAS), PhedEX(CMS), Dirac(LHCb)
- Lower layers generally common (WLCG deliverables/Grid middleware):
 - Computing elements (ARC, Cream, Condor...)
 - Storage (Castor, EOS, dCache, DPM..),
 - File transfer services (CERN FTS2 and FTS3),
 - File/access catalogues (LFC),
 - Virtual machines and remote filesystems for software access (CERNVM, CVMFS),
 - Database caching (Frontier/Squid for ORACLE DB access),
 - Monitoring tools (SAM, PerfSonar, DashBoard),
 - Information infrastructure (BDII).



Processing millions of jobs and PB of data weekly



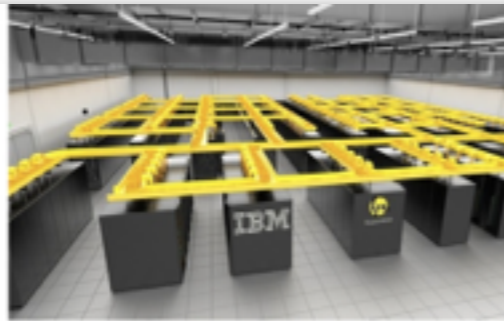
What Will our Computing Sites Look Like?



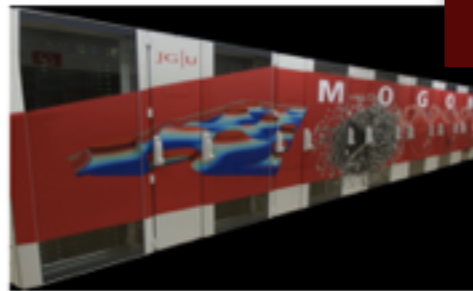
- The answer to this has strong political/financial components, which are hard to predict. Still..

R. Walker

HPC Resource Examples



- SuperMUC, Munich
 - 155,000 Sandy Bridge cores, 2.8M HS06
 - ATLAS 2013 T1/2 pledges ~ 730K HS06
 - Suse Enterprise Linux 11, 2GB/core
 - warm water cooling
 - 40°C inlet. 70°C outlet used to heat building
- Hydra, MPI, Munich
 - 'similar' cluster in spec and scale
 - due Summer 2013. 10k core integration system in place now
- MOGON, Mainz
 - 34k cores SL6



Even today, technically our CPU capacities could fit into one Exa-Scale super-computing center.
Will we get fractions of HPC CPUs? What about our HTC needs?

What will be the impact of IaaS (Cloud) technologies?
Will we get cheques for commercial cloud use (or, again, super-computers..)?

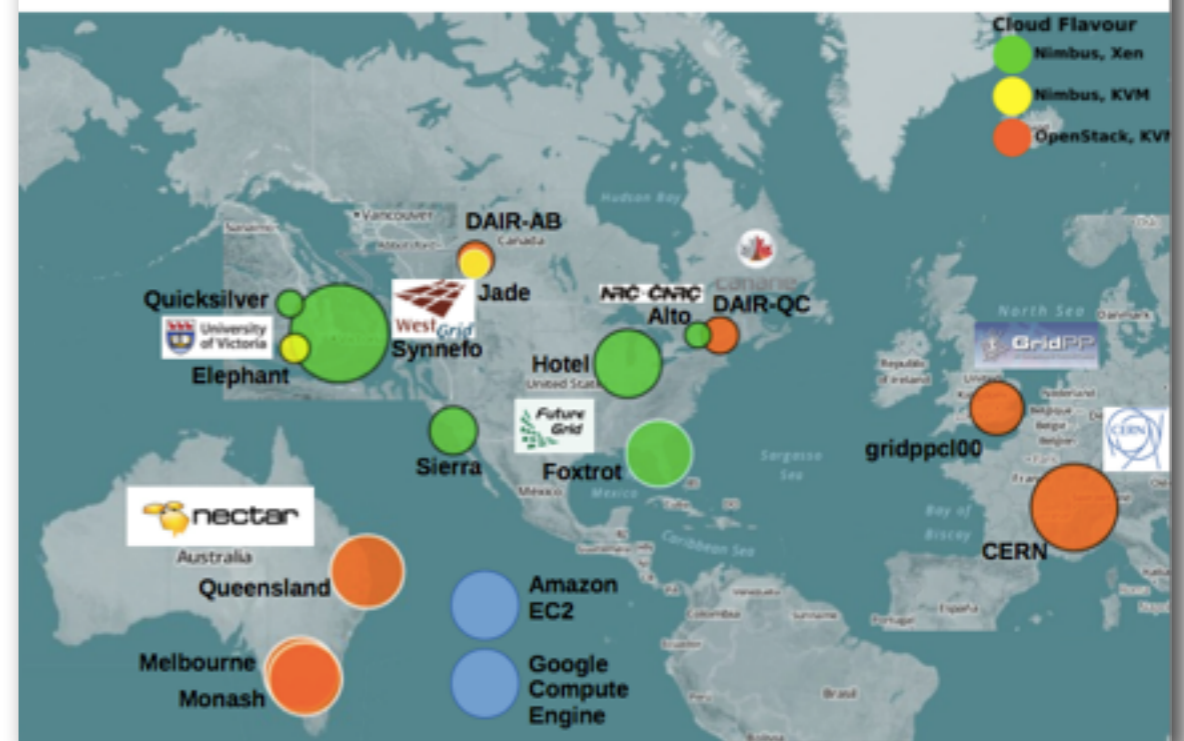
The main item that does not have many solutions and gives a severe constraint is our data storage:

We need reliable and permanent storage under ATLAS control.

From another perspective, with the network evolution (and federated storage, event service..) 'local' becomes re-defined (again).

No need for local storage?
Consolidate to a few (1|10|20|..?) main storage points? Cheaper?

The "Grid of Clouds"



Ian Gable, Ryan Taylor - Sep. 2013

Future Computing Resources

- The ‘global community’ did not really buy into Grid technologies, which were very successful for us:
 - We have a dedicated network of sites, using custom software and serving (mostly) the WLCG community.
 - This does not bode well for the future in terms of funding (not a global solution...).
- From the material presented, it is clear that the amount of data and CPU processing requirements **will not become ‘trivial’ in a global context and will not be feasible in a ‘cellar’ cluster in an institution** but will continue to require the participation of distributed resources, big computing centres, leadership facilities (e.g. HPC) and opportunistic computing. As such they will continue to **need a global workload and data management system such as the current Grid technologies.**
 - We need to make sure a distributed computing solution exists, is sustainable (financially) and has evolved to address the future HL-LHC needs.
- Several venues to explore:
 - Optimizing/changing our workflows, both in analysis and on the grid.
 - It will necessarily involve also a change in the ways people analyze the data!
 - Incorporating diverse/opportunistic/common resources:
 - High Performance Computing centres have a lot of CPU available, we could use the available idle cycles for (a subset of) our activities, e.g. MC simulation.
 - Cloud resources: Again, for a subset of our activities, similar to HPC.. commercial resources?
 - Opportunistic/limited time offers of big computing centers:
 - The experiments need to be able to simply and quickly integrate such resources into their distributed computing environment.
 - Volunteer computing resources: exploiting virtualization (CernVM), BOINC..
- Furthermore, the scientific community will remain globally distributed and will need a managed access to the dedicated resources with **appropriate security features included (!)**.

What can we do for Run-3 and beyond?



- The ATLAS Computing Model for Run-2 is already quite austere:
 - Assuming our CPU/event in reconstruction and AOD size will be the same in Run-2 and Run-4 is **again very optimistic, requires a lot of work..**
 - savings in numbers of data replicas, data retention on disk ... cannot be pushed much further before impacting accessibility - **we could gain fractions, not factors.**
 - **Options:**
 - Work on software improvements (non trivial...), compromising Physics very (more) expensive.
 - Find additional resources, adapt to using anything ‘on the market’ in an optimal way:
 - new CPU architectures (many-core/MIC, GPGPU...),
 - profit from parallelism wherever possible (memory savings..).
 - Opportunistic access to any resource available (**HPC, Cloud, BOINC**):
 - Very fine job granularity control (per record processing, ‘event service’)
 - Optimal use of the WAN/LAN/.. for data access and management.
- In any case:
 - **Anticipate computing evolution and work on adapting ...**
 - A strong assumption on High Performance and High Throughput Computing converging!
 - Use our experience in distributed computing and use it to adapt!

Summary

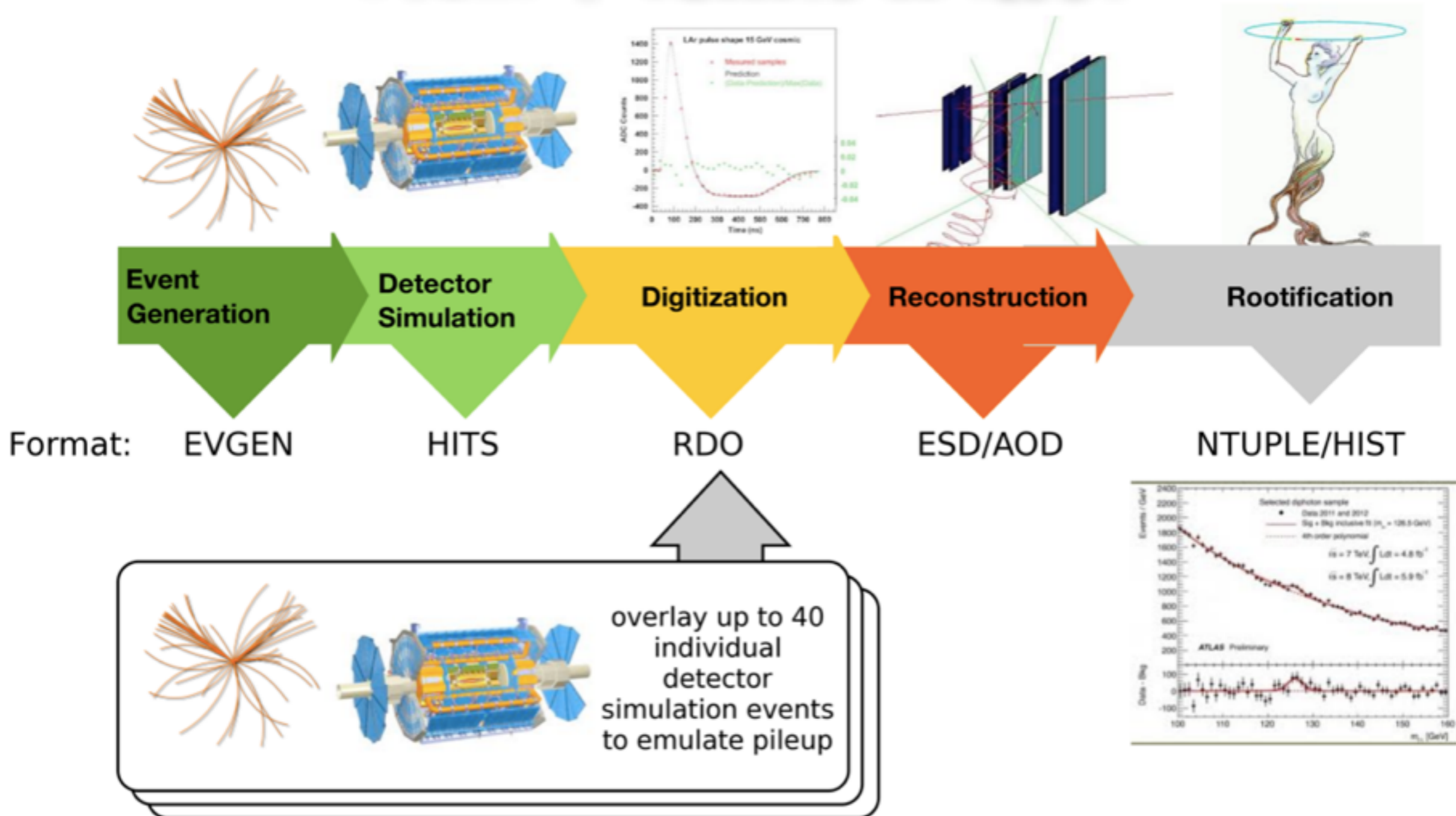


- This talk just a bit more than a collection of vague thoughts for further discussion and planning.
 - **It is however clear that with an ambitious ATLAS Physics program for the future and the (computing) world changing around us we need to be prepared and invest time and effort!**

Simulation Flow in ATLAS



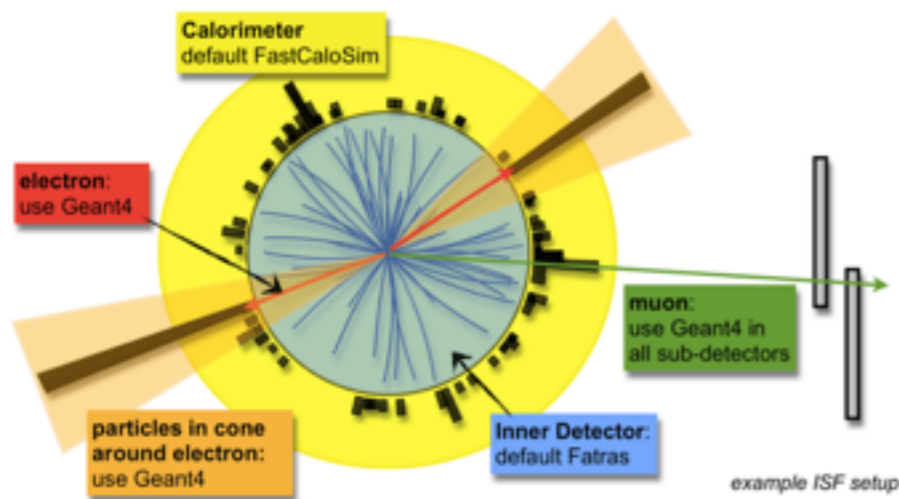
From 4-vectors to ROOT



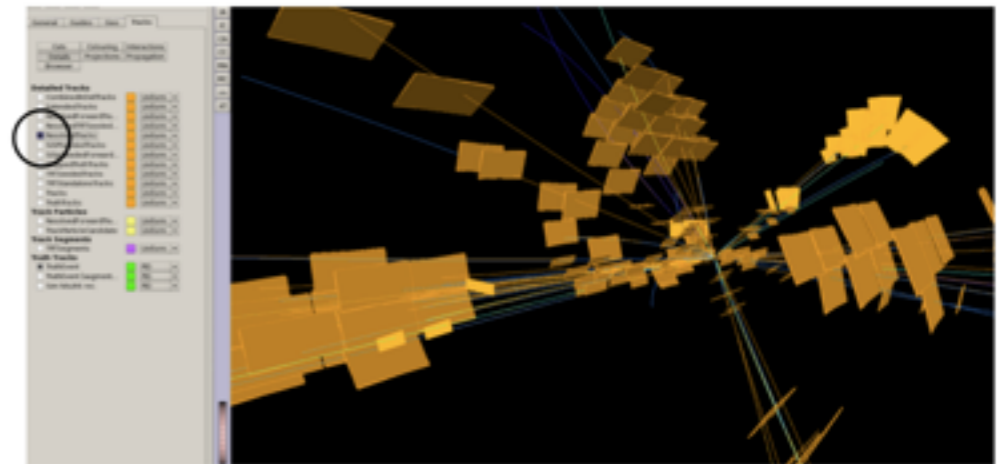
Faster Solutions: ISF



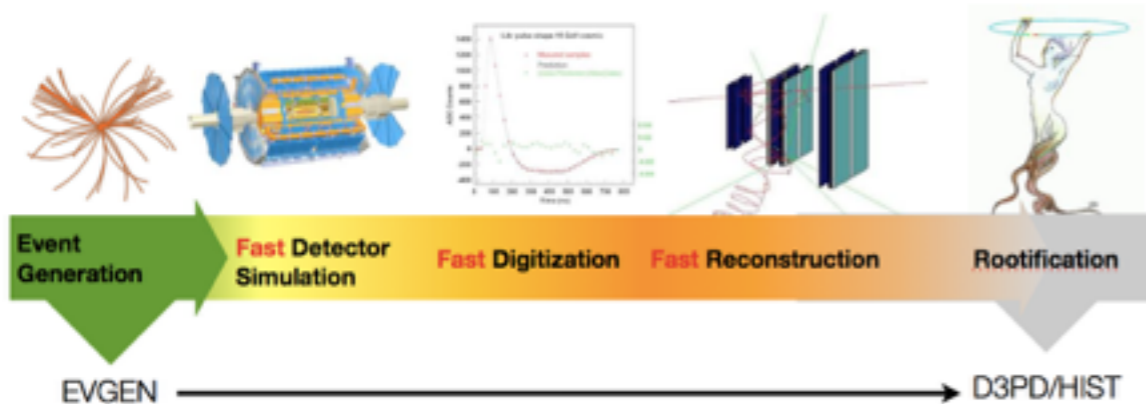
- Do people really need to use the (very CPU consuming) G4 full simulation chain for all cases?
 - It is the 'safe' solution, **but it does not scale** with respect to the resources we will have available for Run 2..
- Conceptualizing and developing 'Fast simulation/digitization/reconstruction chain'.. for 2016++



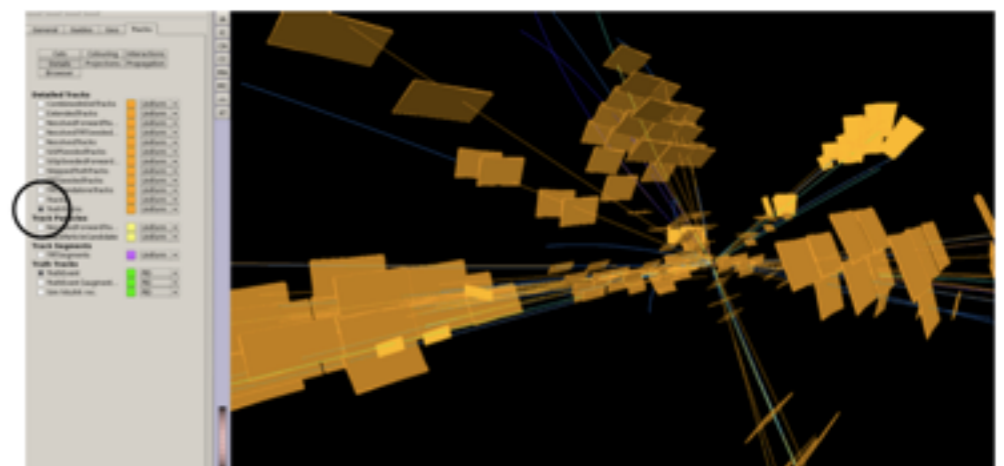
reconstructed tracks



An all-in-one chain for fast MC



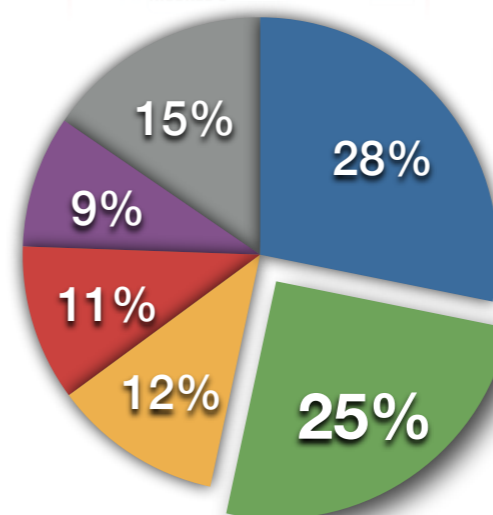
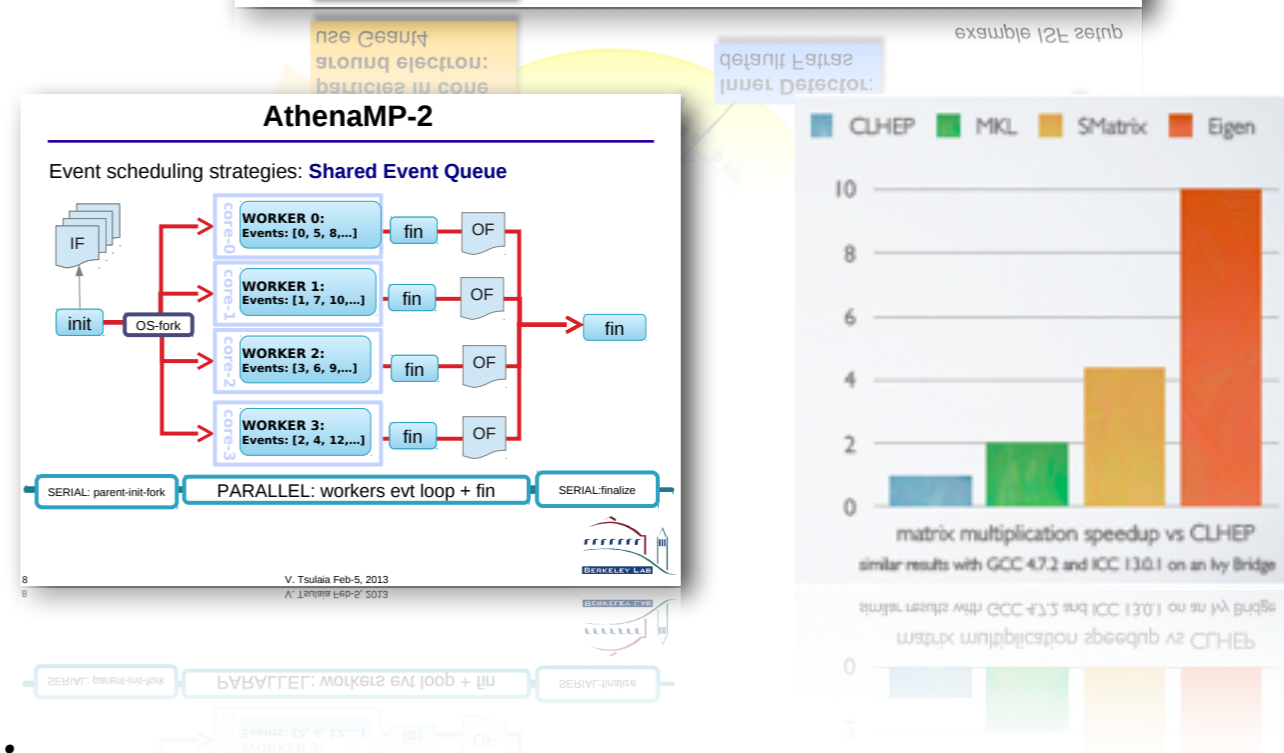
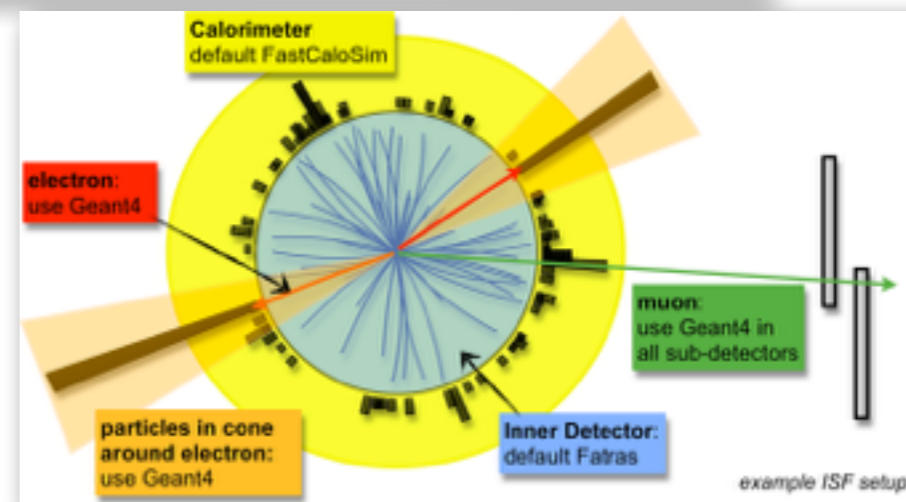
truth tracks



Working towards Solutions



- **Simulation : CPU**
 - Integrated Software Framework.
- **Reconstruction : Memory & CPU**
 - Parallelism, code speedup.
 - MP solution to reduce memory footprint.
- **Analysis Model : multiplication of data formats**
 - Common analysis data format, xAOD.
 - Streamlining the analysis flow.



Disk usage at T1s & T2s

- AOD
- ESD
- RAW
- ntuple
- HITS
- others

500k jobs on Google



Cloud computing

On going R&D on academic clouds and Amazon or Google (AUS,CA, US,...)

Issues with long jobs and I/O

Plan for use 'academic' clouds and opportunistic use of 'cheap' commercial is possible

Some cloud computing providers start to propose cost-competitive offers (with some limitations)



SuperMUC a PRACE Tier-0 centre :
155,000 Sandy Bridge cores, 2.8M HS06

WLCG 2013 T0/1/2 pledges ~2.0M HS06

HPC (High-Performance Computing) resources

- ◆ Large investments in many countries : from Peta to Exa scales initiatives[1]
- ◆ Latest competitive supercomputers are familiar Linux clusters
- ◆ Large number of spare CPU cycles are available at HPCs which are not used by ‘standard’ HPC applications
- ◆ Projects to use idle CPU cycles at HPC centers in US, China & DE
- ◆ Demonstrators working for simulation & event generation
- ◆ Difficult to use HPC centers for I/O intensive applications
- ◆ Outbound connectivity of HPC centers may also be an issue
- ◆ Some T2s plan to provide pledges resources on shared HPC facilities

Might endanger traditional HEP computing budget

Use of Opportunistic Resources

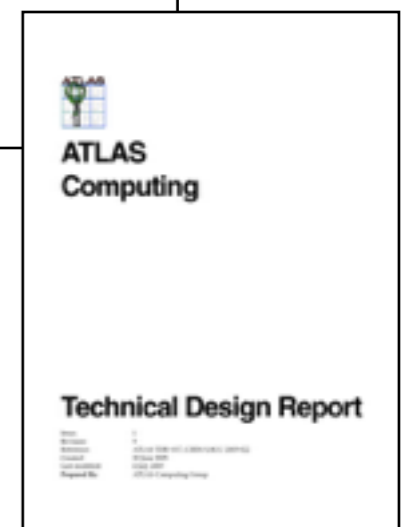
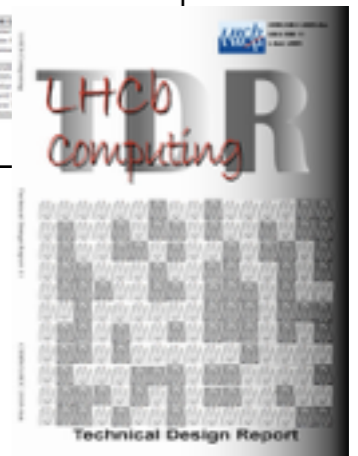


- Opportunistic use of cloud resources development:
 - **Google Compute Engine (GCE) preview project.**
 - Google allocated additional resources for ATLAS for free
 - ~5M cpu hours, i.e. 4000 cores for about 2 months (original preview allocation 1k cores).
 - Powerful machines with modern CPUs Resources.
 - Organized/integrated as HTCondor based PanDA
 - Centos 6 based custom built images, with SL5 compatibility libraries to run ATLAS software.
 - Condor head node, proxies are at BNL
 - Output exported to BNL SE
 - Work on capturing the GCE setup in Puppet
 - *The idea was to test long term stability while running a cloud cluster similar in size to Tier 2 site in ATLAS*
 - Use-case are CPU intensive Monte-Carlo simulation workloads.
 - Also, work in progress in integrating **opportunistic HPC (Super Computer) resources into the Grid.**
 - Several interested participants in US and EU.



Experiment Computing Models

- The LHC experiment Computing Technical Design Reports were produced in ~2005, with the best knowledge available at the time.
 - **No plan survives the reality, in this case the arrival of the data:**
 - Operational experience introduced significant modifications and improvements in Run 1.
 - E.g. moving away from the Monarc model, all Tiers perform similar activities and pass the data between them.
 - Significant technological evolution until today also impacted (and continues to impact) the optimal operational models:
 - For example network bandwidths increased more than anticipated, one can make better use of storage resources with more dynamic data movement.
 - **We made it work! A big success in Run-1.**
 - Still, expensive to maintain and develop in terms of manpower.
 - Awareness that searching for common use cases between experiments and global community could boost the activities and economize manpower is becoming crucial in view of the current financial climate.
 - Updated computing models made in a document requested by the LHCC:
 - <http://cds.cern.ch/record/1695401>
 - **This will get us through the next years (Run-2), but we need to look beyond!**



WLCG tiered structure

- The LHC experiments rely on **distributed computing resources**:

- **WLCG - a global solution, based on the Grid technologies/middleware.**

- distributing the data for processing, user access, local analysis facilities etc.
- at time of inception envisaged as the seed for global adoption of the technologies.

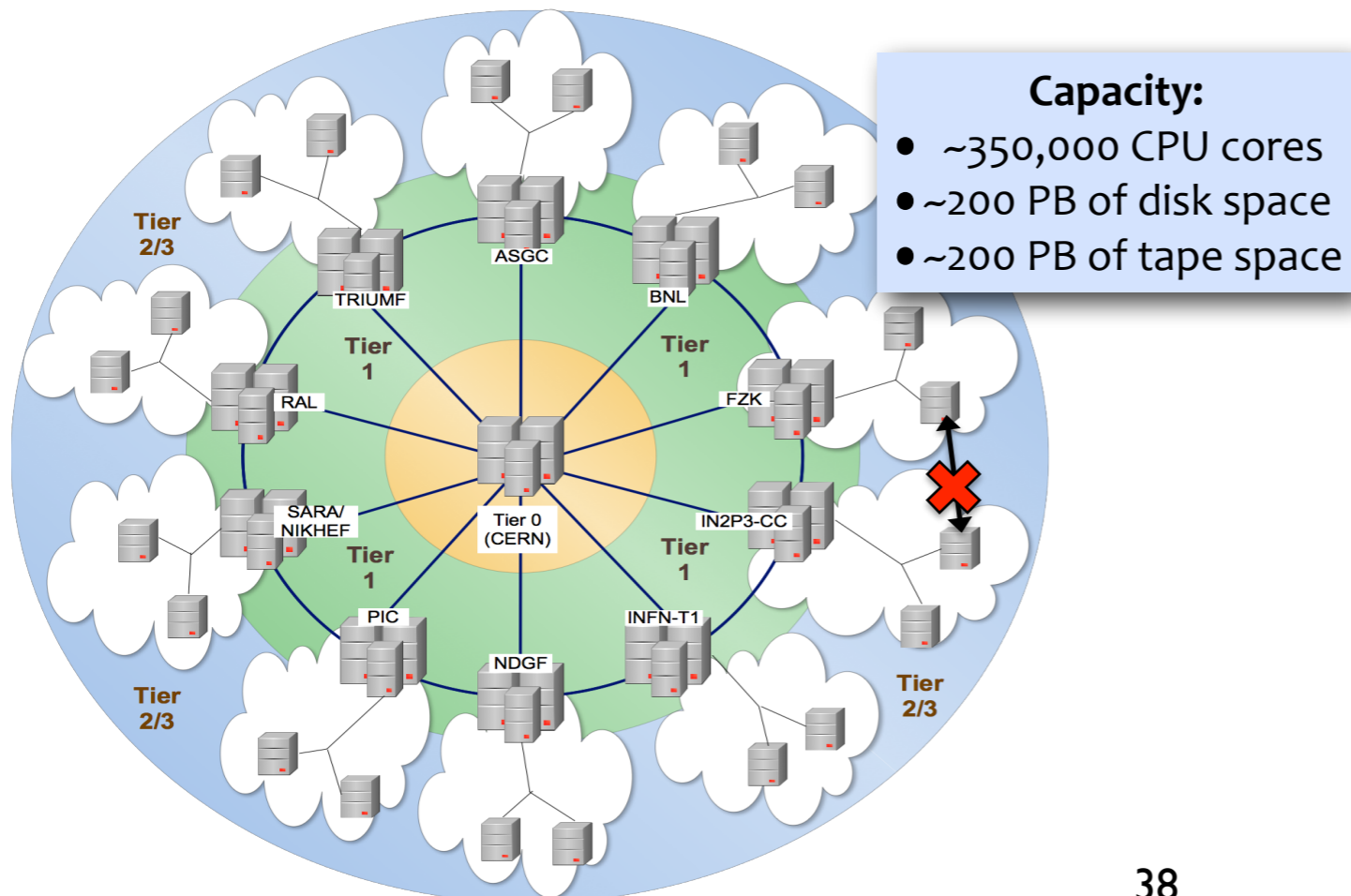


- **Tiered structure:**

- Tier-0 at CERN: the central facility for data processing and archival,
- 11 Tier-1s: big computing centers with high quality of service used for most complex/intensive processing operations and archival,
- ~140 Tier-2s: computing centers across the world used primarily for data analysis and simulation.

- **WLCG and LHC computing a big success in Run 1!**

- **Computing was not a limiting factor for the Physics program of the LHC experiments.**
- **Thanks to our developers, operators and Grid sites for their excellent performance and contributions!**

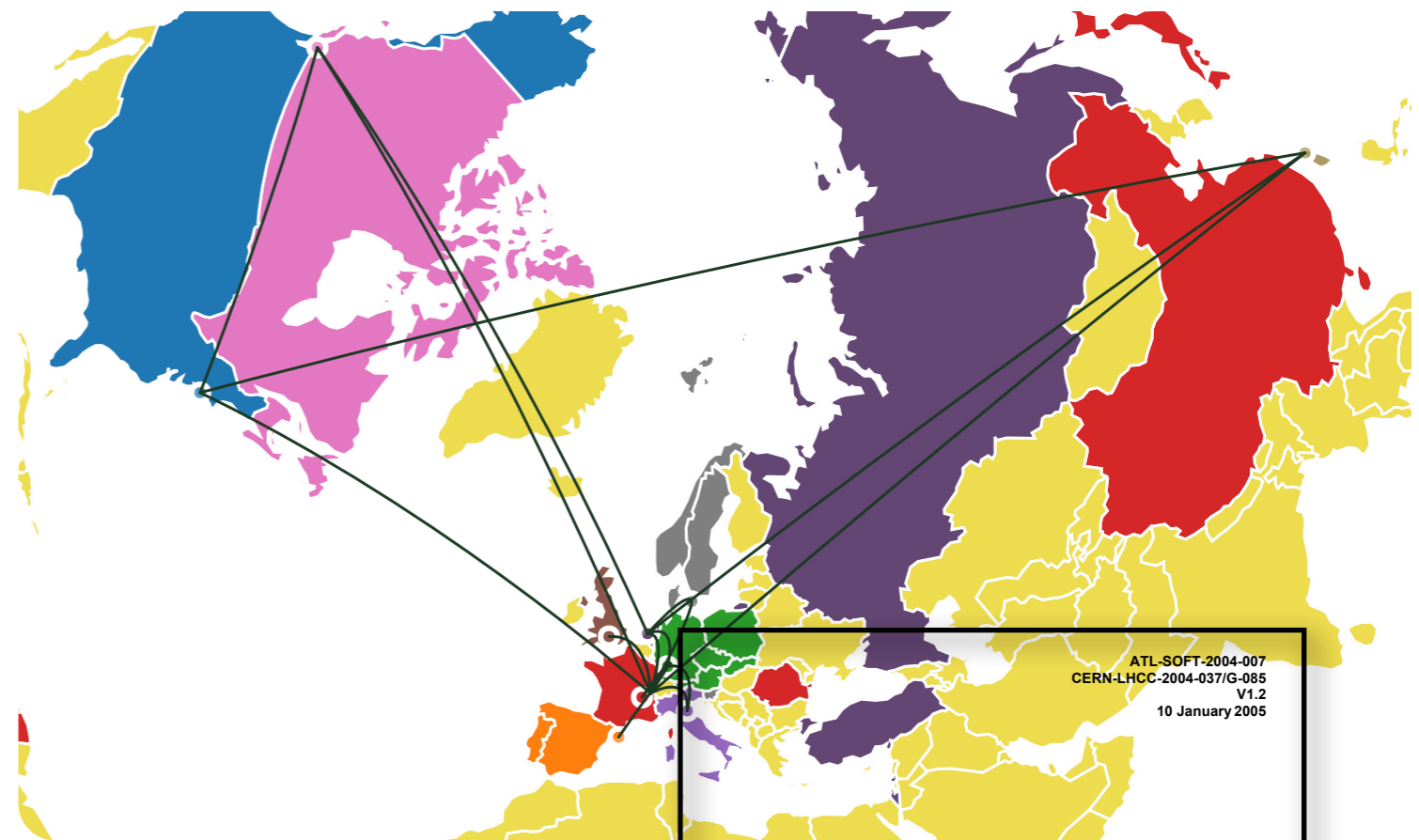


- **Hierarchical tier organization based on MONARC (MODELS OF NETWORKED ANALYSIS AT REGIONAL CENTERS) network topology**
- In **ATLAS** sites are grouped into **clouds** for organizational reasons
- Possible communications:
 - Optical Private Network
 - T0-T1
 - T1-T1
 - National networks
 - Intra-cloud T1-T2
- Restricted communications: General public network
 - Inter-cloud T1-T2
 - Inter-cloud T2-T2

Initial Computing Model (2005)



- Derived from MONARC ('99) model
- CERN-To the center
- 10 T1s connected by dedicated 10Gb/s links (LHCOPN)
- O(100) T2s each attached to a T1
- The data flows along the hierarchy
- Insufficient networking assumed
- Hierarchy of functionality and capability



ATL-SOFT-2004-007
CERN-LHCC-2004-037/G-085
V1.2
10 January 2005

THE ATLAS COMPUTING MODEL

Prepared by: D. Adams, D. Barberis, C. Bee, R. Hawkins, S. Jarp, R. Jones¹,
D. Malon, L. Poggioli, G. Poulard, D. Quarrie, T. Wenaus

on behalf of the ATLAS Collaboration

Abstract: The ATLAS Offline Computing Model is described. The main emphasis is on the steady state, when normal running is established. The data flow from the output of the ATLAS trigger system through processing and analysis stages is analysed, in order to estimate the computing resources, in terms of CPU power, disk and tape storage and network bandwidth, which will be necessary to guarantee speedy access to ATLAS data to all members of the Collaboration. Data Challenges and the commissioning runs are used to prototype the Computing Model and test the infrastructure before the start of LHC operation.

The initial planning for the early stages of data-taking is also presented. In this phase, a greater degree of access to the unprocessed or partially processed raw data is envisaged.

¹ Chair and contact person: Roger.Jones@cern.ch

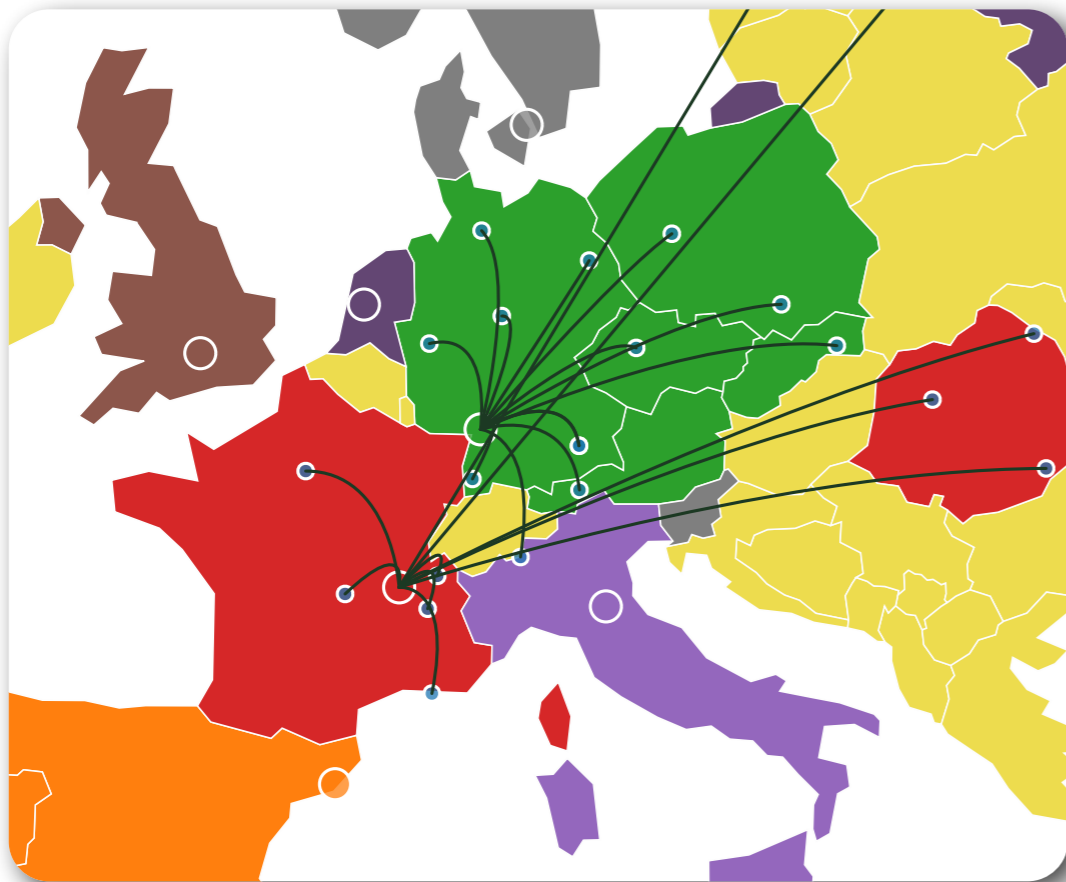
2010

Planned data distribution

Jobs go to data

Multi-hop data flows

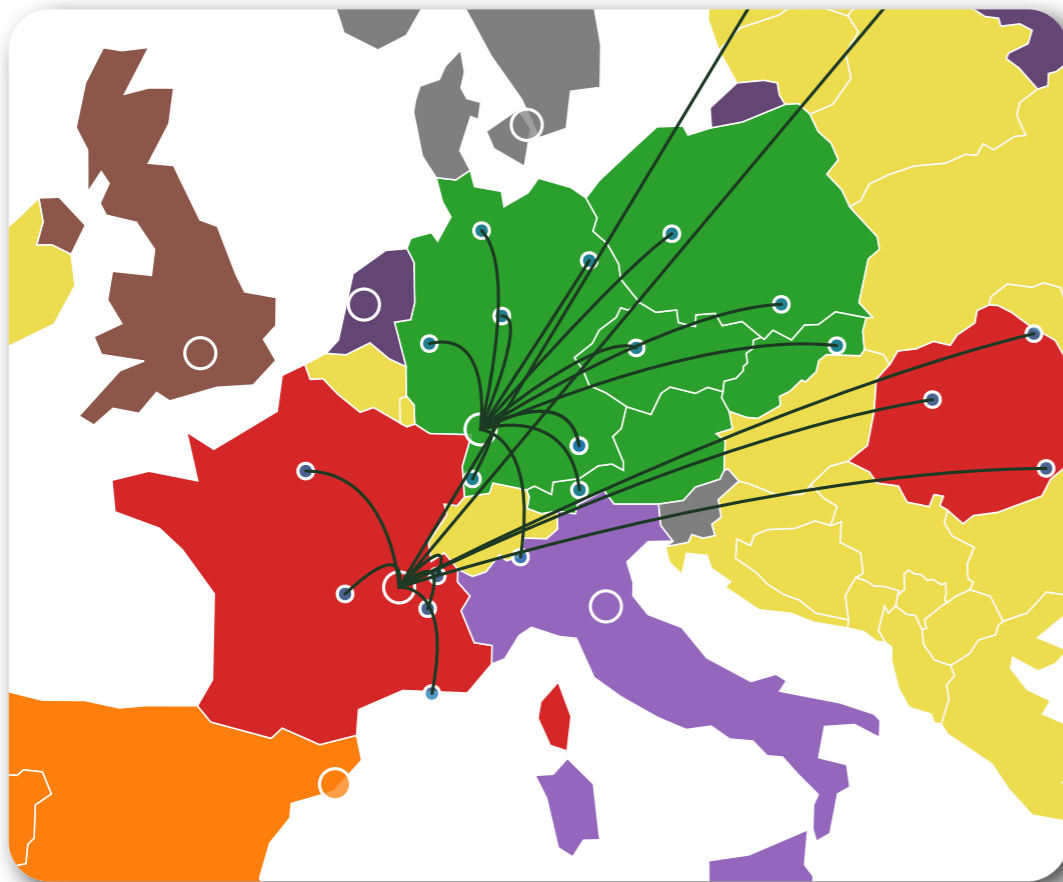
Poor T2 networking across regions



~20 AOD copies distributed worldwide

2010

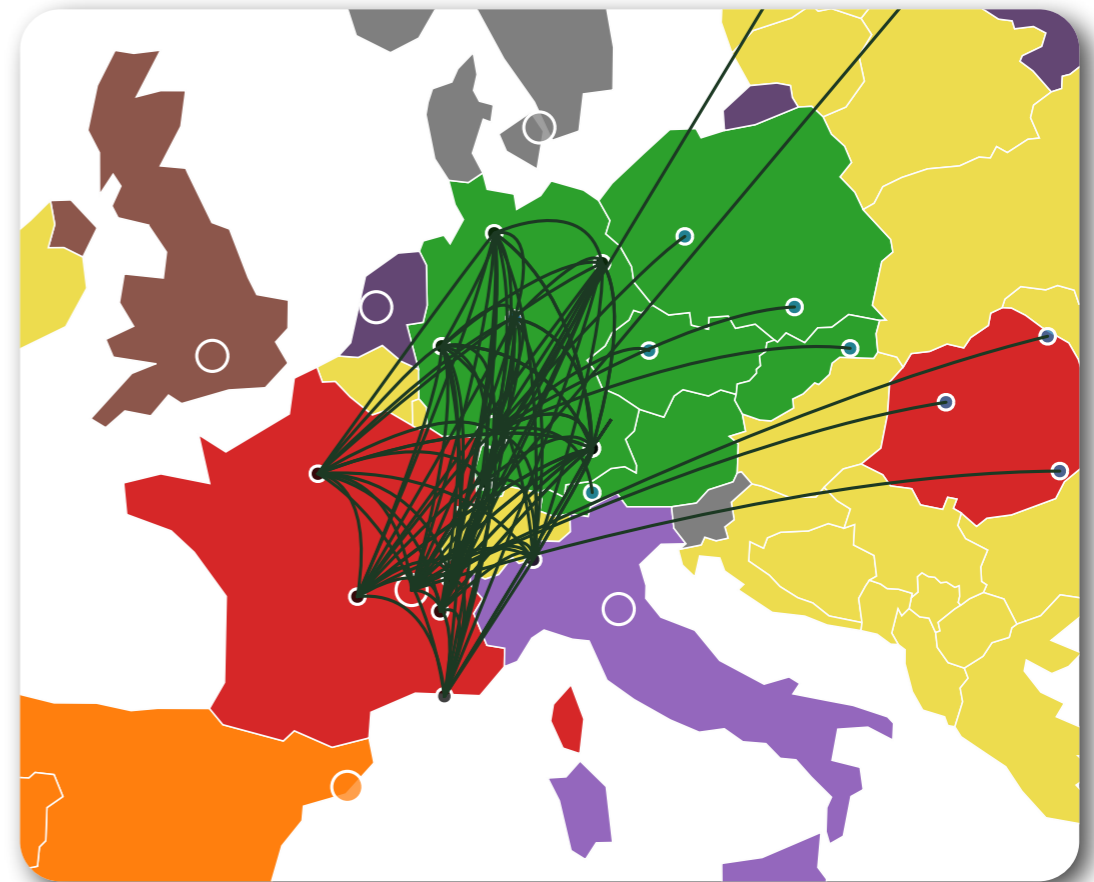
Planned data distribution
Jobs go to data
Multi-hop data flows
Poor T2 networking across regions



~20 AOD copies distributed worldwide

2013

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



4 AOD copies distributed worldwide

Model Parameters



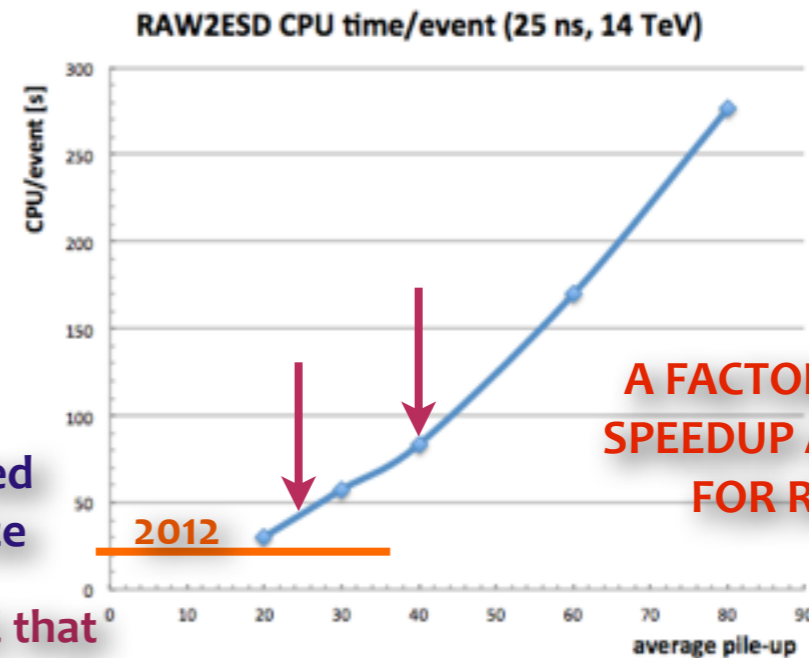
- **FUTURE** planning resource usage for all the activities:
 - **A lot** of work needed to reach the target CPU times/event, event sizes etc..
 - **need** to achieve a reasonable match to available resources!

LHC and data taking parameters		2012 pp actual	2015 pp mu=25 @ 25 ns	2016 pp mu=40 @ 25 ns	2017 pp mu=40 @ 25 ns
Rate [Hz]	Hz	400 + 150 (delayed)	1000	1000	1000
Time [sec]	MSeconds	6.6	3.0	5.0	7.0
Real data	B Events	3.0 + 0.9 (delayed)	3.0	5.0	7.0
Full Simulation	B Events	2.6 (8 TeV) + 0.8 (7 TeV)	2	2	2
Fast Simulation	B Events	1.9 (8 TeV) + (7 TeV)	5	5	5
Simulated Data					
Event sizes					
Real RAW	MB	0.8	0.8	1	1
Real ESD	MB	2.4	2.5	2.7	2.7
Real AOD	MB	0.24	0.25	0.35	0.35
Sim HITS	MB	0.9	1	1	1
Sim ESD	MB	3.3	3.5	3.7	3.7
Sim AOD	MB	0.4	0.4	0.55	0.55
Sim RDO	MB	3.3	3.5	3.7	3.7
CPU times per event					
Full sim	HS06 sec	3100	3500	3500	3500
Fast sim	HS06 sec	260	300	300	300
Real recon	HS06 sec	190	190	250	250
Sim recon	HS06 sec	770	500	600	600
AOD2AOD data	HS06 sec	0	19	25	25
AOD2AOD sim	HS06 sec	0	50	60	60
Group analysis	HS06 sec	40	2	3	3
User analysis	HS06 sec	0.4	0.4	0.4	0.4

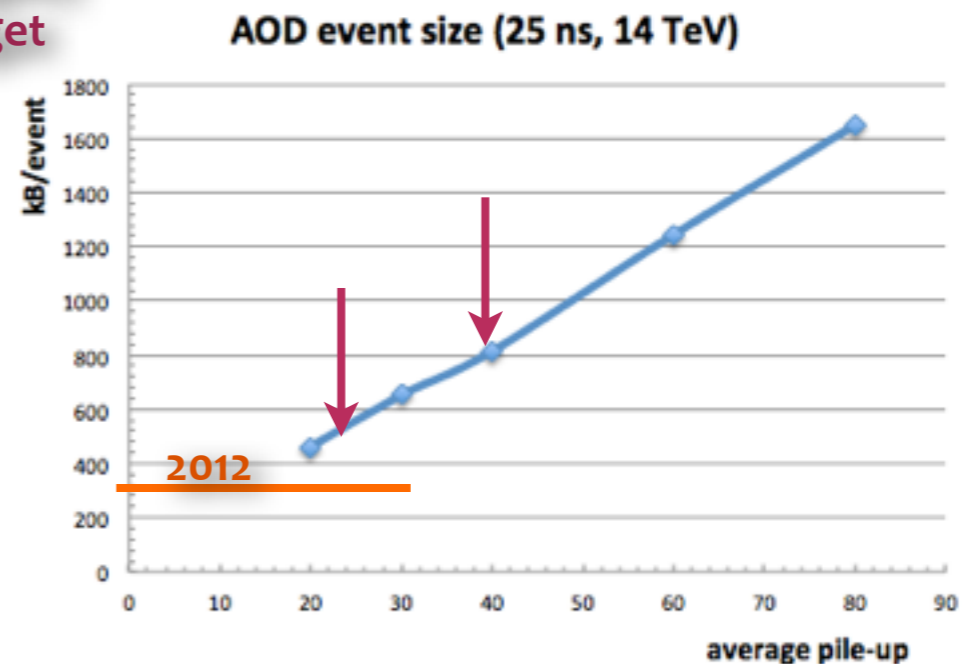
Expected data size

The MC that fits within the (pledged) budget

The processing times will have to get much faster than Run-1!



A FACTOR 3 CODE SPEEDUP ACHIEVED FOR RUN-2!



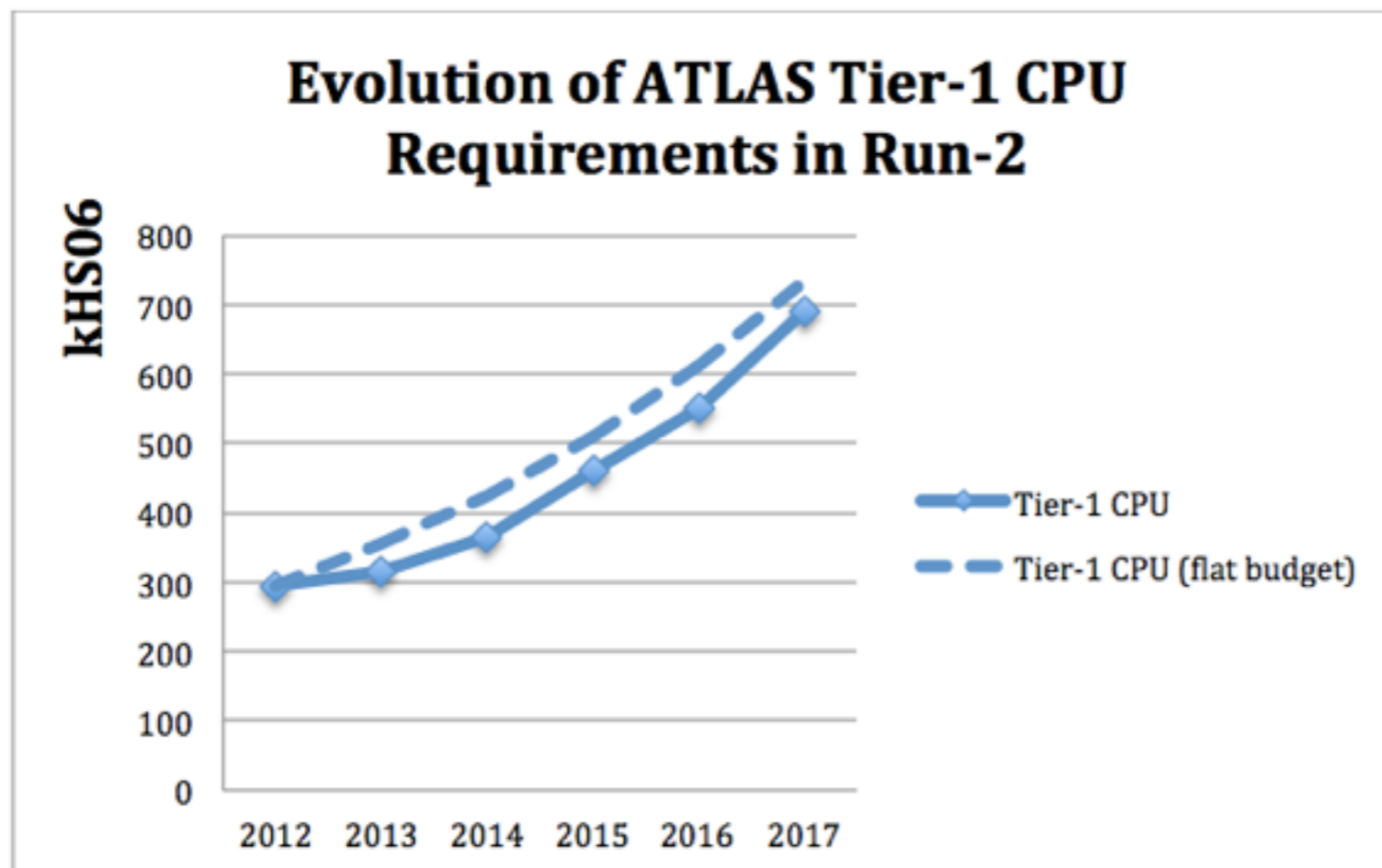
Resource Evolution (planning)



Tier-1 CPU (kHS06)	2015	2016	2017
Re-processing	38	30	43
Simulation production	154	89	102
Simulation reconstruction	194	245	280
Group (+user) activities	76	187	267
Total	462 [478]	552	691

Centrally managed activities, done by expert teams and of general interest.

The CPU consumption of Group activities (Reduction framework) and final user analysis jobs.



Flat CPU budget: factor 1.2/year

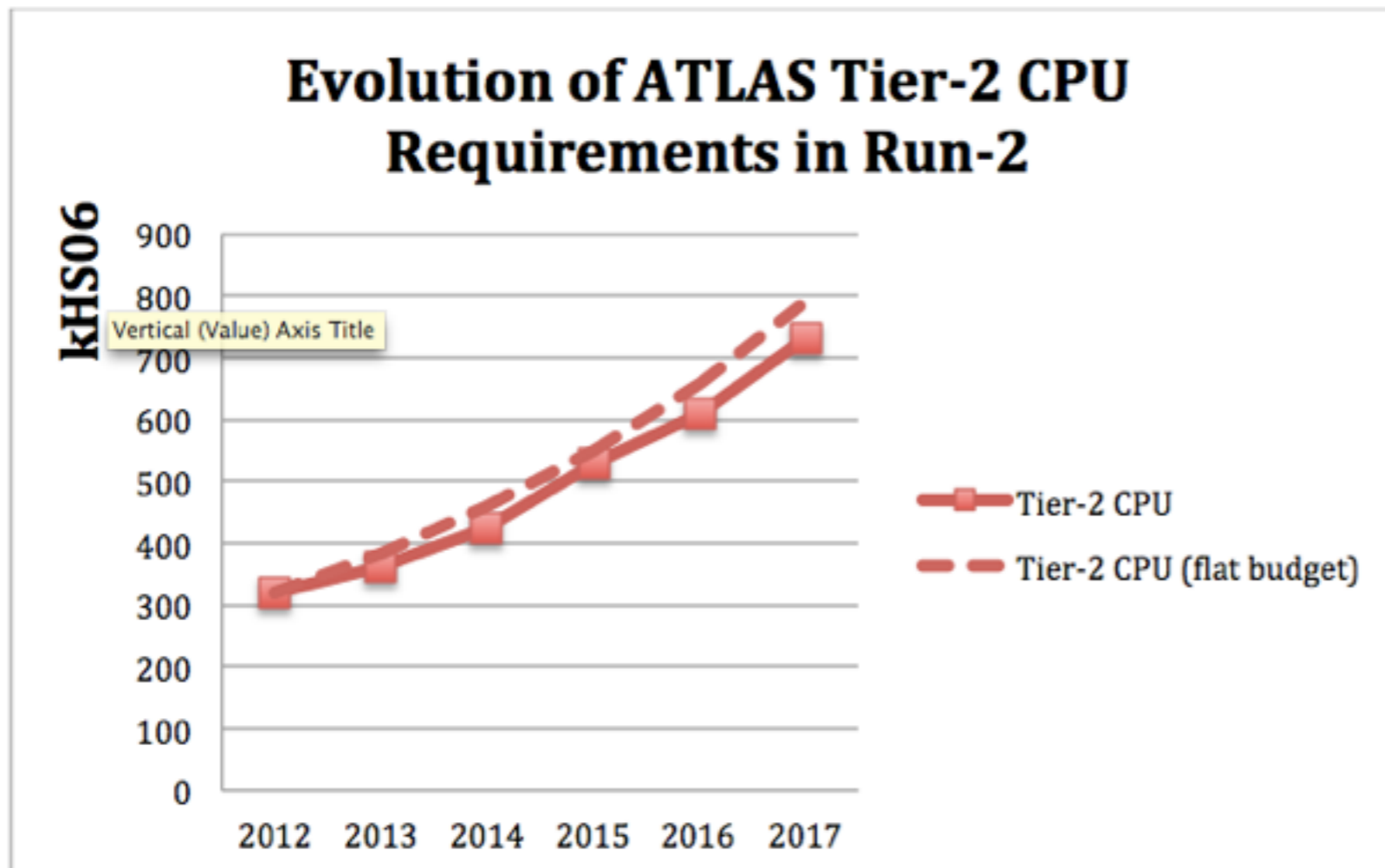
Resource Evolution cont'd



Tier-2 CPU (kHS06)	2015	2016	2017
Re-processing	20	33	47
Simulation production	338	347	396
Simulation reconstruction	77	61	70
Group + User activities	96	166	219
Total	530 [522]	608	732

Central production activities will be balanced between Tier-1s and Tier-2s.

Group and User activities will be balanced between Tier-1s and Tier-2s, all groups and users get a share!



Flat CPU budget: factor 1.2/year

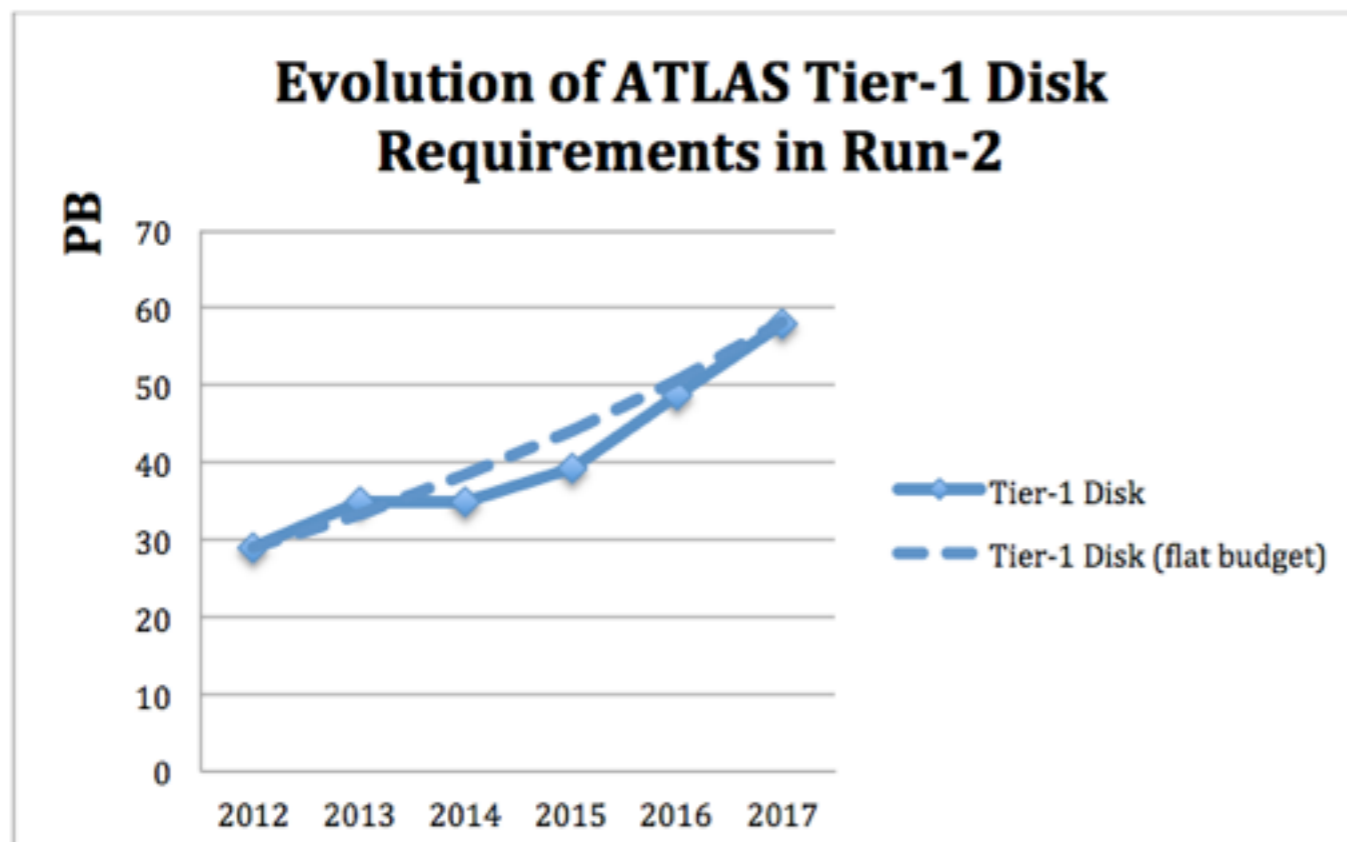
Resource Evolution cont'd



Tier-1 Disk (PB)	2015	2016	2017
Current RAW data	2.4	5.0	7.0
Real ESD+AOD+DPD data	5.6	7.9	11.1
Simulated RAW+ESD+AOD+DPD data	9.2	11.4	11.4
Calibration and alignment outputs	0.3	0.3	0.3
Group data	7.5	8.0	10.4
User data (scratch)	2.0	2.0	2.0
Cosmics	0.2	0.2	0.2
Processing and I/O buffers	3.0	3.0	3.0
Dynamic data buffers (30%)	9.0	10.9	12.6
Total	39 [47]	49	58

Centrally managed and stored data of interest to everyone.

Group data is stored in dedicated locations, managed by the ATLAS (physics, detector,...) groups. User data is meant to be downloaded to your laptop eventually.



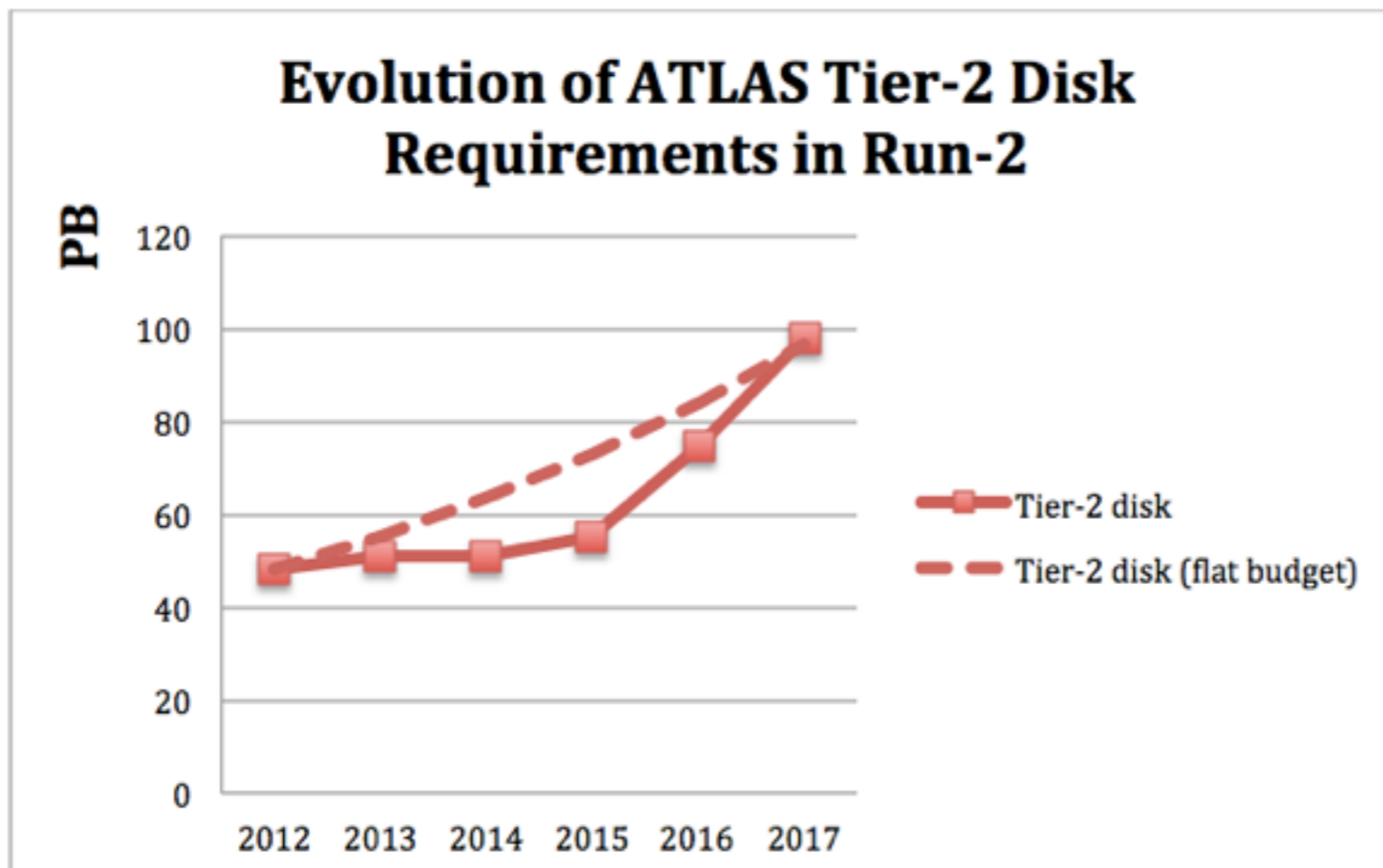
Flat disk budget: factor 1.15/year

Resource Evolution cont'd



<i>Tier-2 Disk (PB)</i>	2015	2016	2017
Real AOD+DPD data	4.1	6.3	10.6
Simulated HITS+RDO+ESD+AOD	10.6	16.6	21.6
Calibration and alignment outputs	0.2	0.2	0.2
Group data	20.4	29.3	41.6
User data (scratch)	4.0	4.0	4.0
Processing and I/O buffers	3.0	3.0	3.0
Dynamic data buffers (30%)	12.7	15.3	16.8
Total	55 [65]	75	98

... however we build on Run-1 experience: there will be an yearly accumulation of Group data throughout Run-2.



Flat disk budget: factor 1.15/year