Roger Jones

ATLAS Offline Computing Model

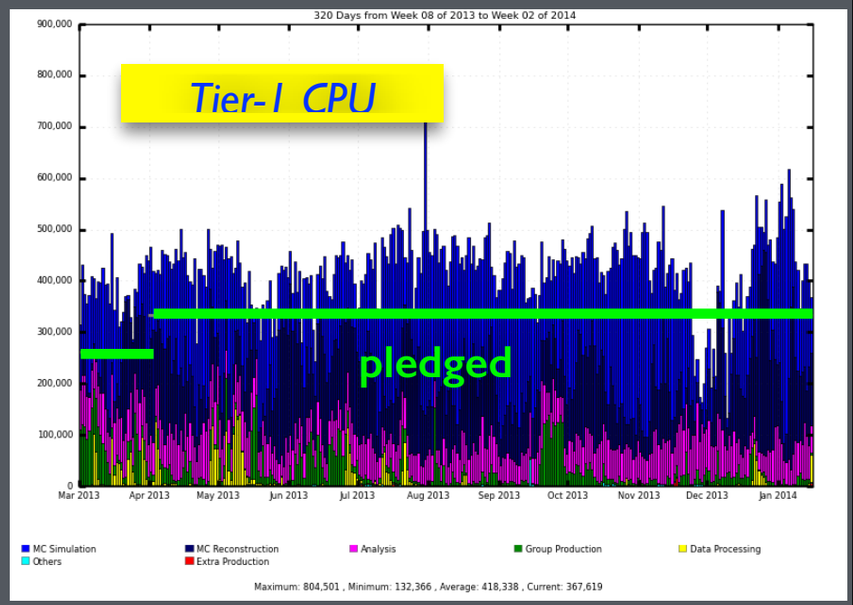
or

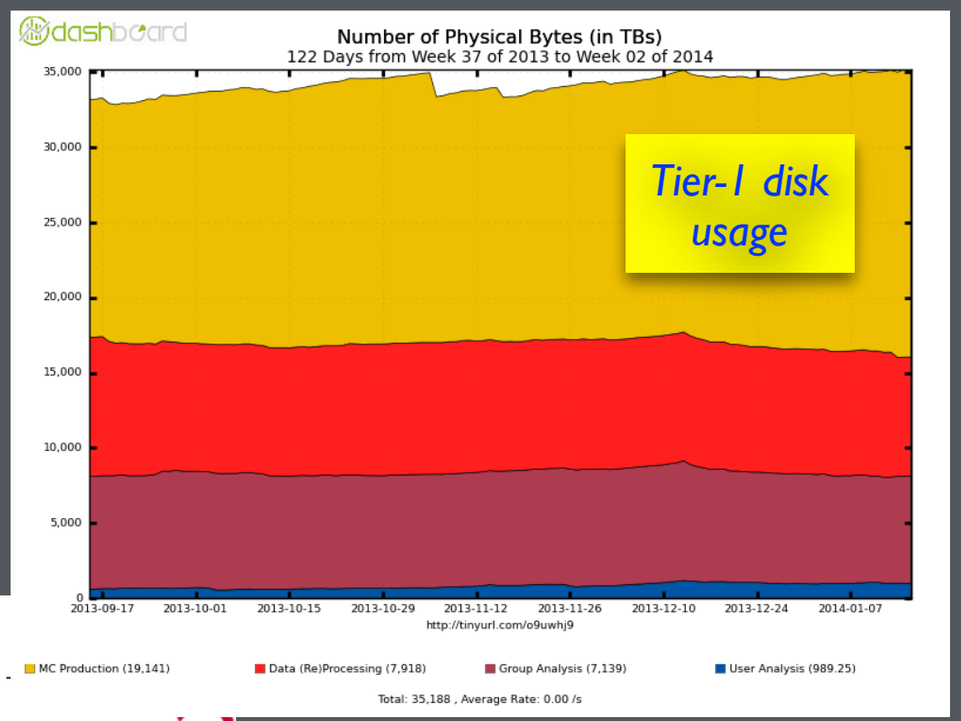
Fitting a quart into a pint pot

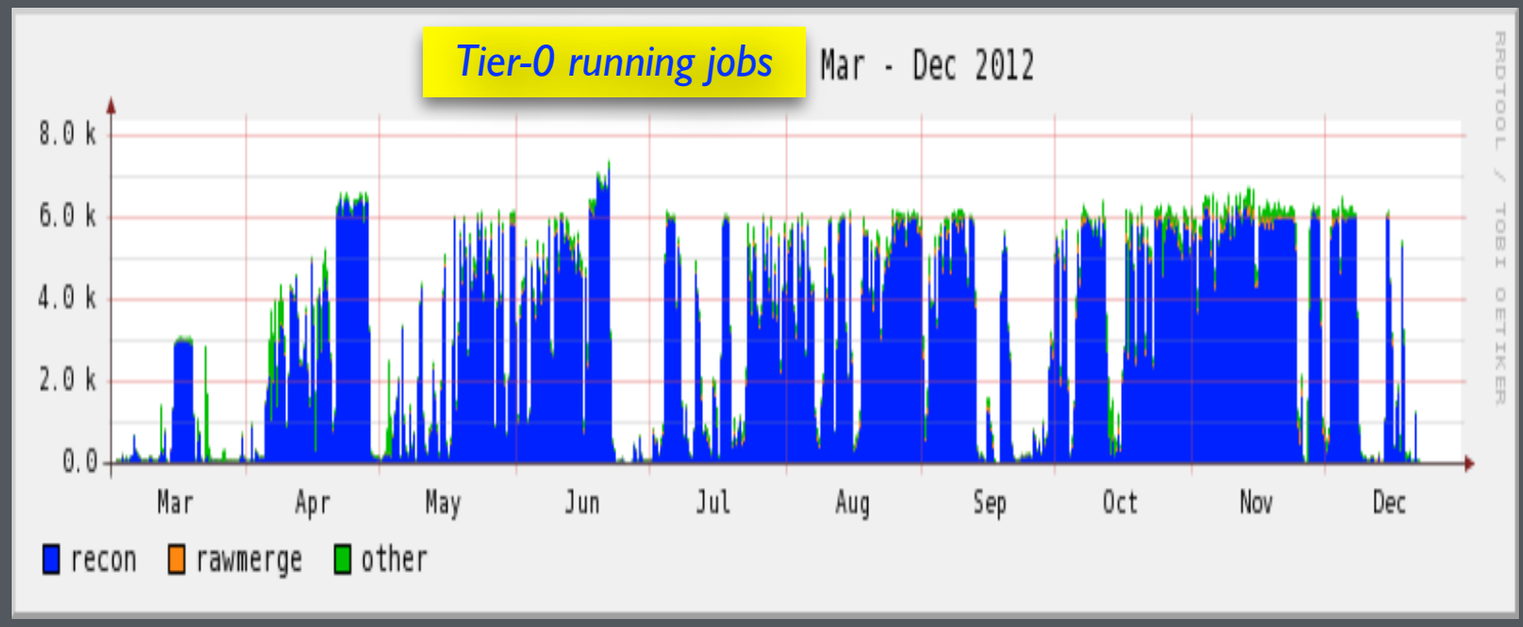
LANCASTER UNIVERSITY

Computing Resource Usage in 2013

- our CPU usage exceeds pledges
  - ➡ opportunistic usage
  - ➡ aim at using further resources (clouds, HPC,...)
- disk resources are fully used
  - ➡ not many opportunistic resources at hand
- Tier-0 managed date rate in 2012
  - ➡ idle between data taking periods
  - ➡ at the limit at the end of 2012

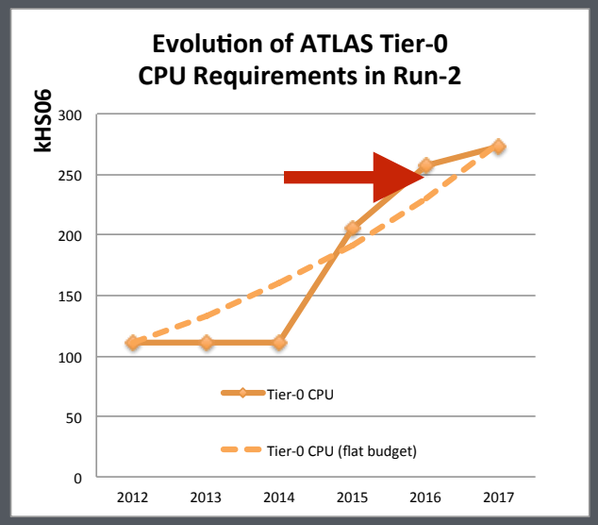




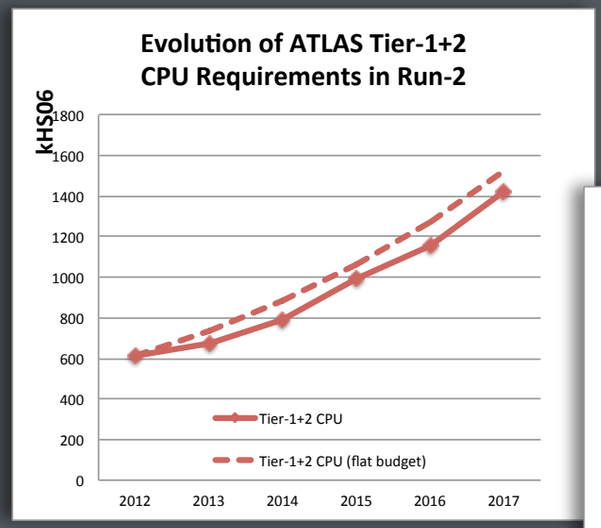


Resource Projections for Run-2

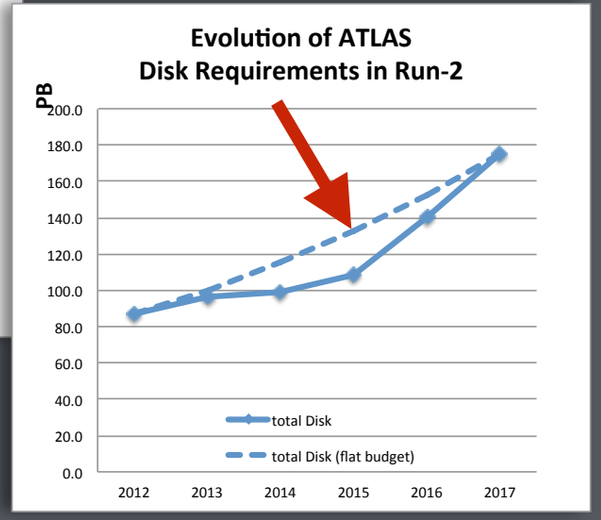
- we have to stay within budget constraints
  - ➡ assumption is we stay with a constant budget
  - ➡ interplay of technology advancement, market price and needed replacements



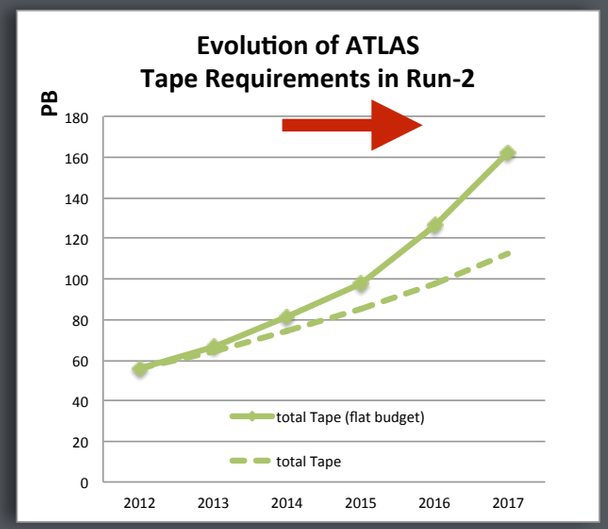
➡ factor 2.5 by 2017



➡ factor 2.3 by 2017

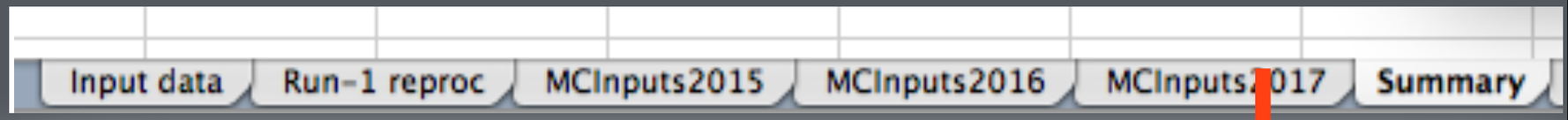


➡ factor 2 by 2017



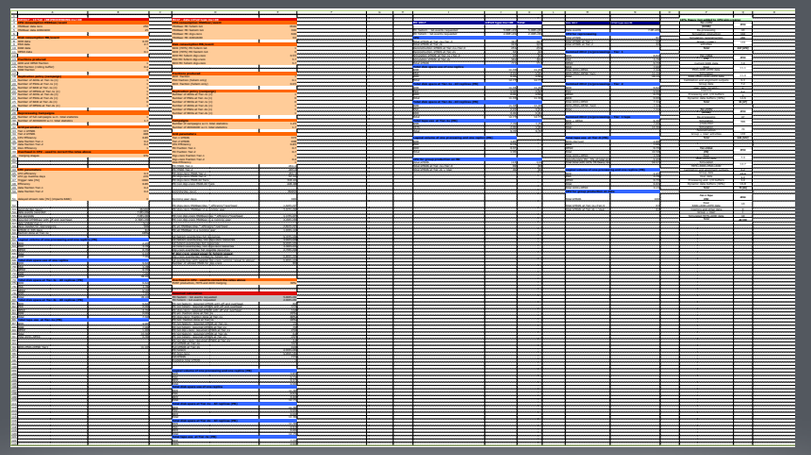
➡ factor 3 by 2017

- requirements follow modelling of needs
  - ➡ assumptions based on 2012, plus LS1 software upgrade goals



Model Assumptions for projections

B.Kersevan



- model uses
  - ➡ basic assumption: operation at 25 nsec and 1 kHz
    - μ=25 in 2015, μ=40 in 2016/7
  - ➡ event numbers (in Billions)

using LHC seconds .....▸

	2015	2016	2017
Real Data	3	5	7
Full Simulation	2	2	2
Fast Simulation	5	5	5

ad hoc assumption .....▸

- ➡ event sizes and CPU per year, e.g. for 2017:

size (MB)	HITS	RAW	ESD	AOD
data 2017	-	1	2.7	0.35
MC 2017	1	(3.7)	3.7	0.55

MC truth .....▸ (RAW)

not much kept ◂..... (3.7)

plus generators ▸..... (fast sim, full sim)

AOD2AOD fix ▸..... (fixes)

CPU	fast sim	full sim	reco	fixes	group	user
data	-	-	250	25	3	0.4
MC	300	3500	600	60		

incl. trigger on MC ▸..... (reco)

used 170 in 2012, plus overheads ▸..... (reco)

times ~ 50 teams ▸..... (group, user)

- ➡ plus additional samples and processing
  - DESD (mostly CP groups), calibration and alignment data, cosmics, misc. user data
- ➡ technical overheads
  - processing and I/O buffers
  - ~ 20% dynamic data placement buffers (PD2P) – in reality the disks are full

do we need all of that ? .....▸

➡ How realistic are these numbers ?

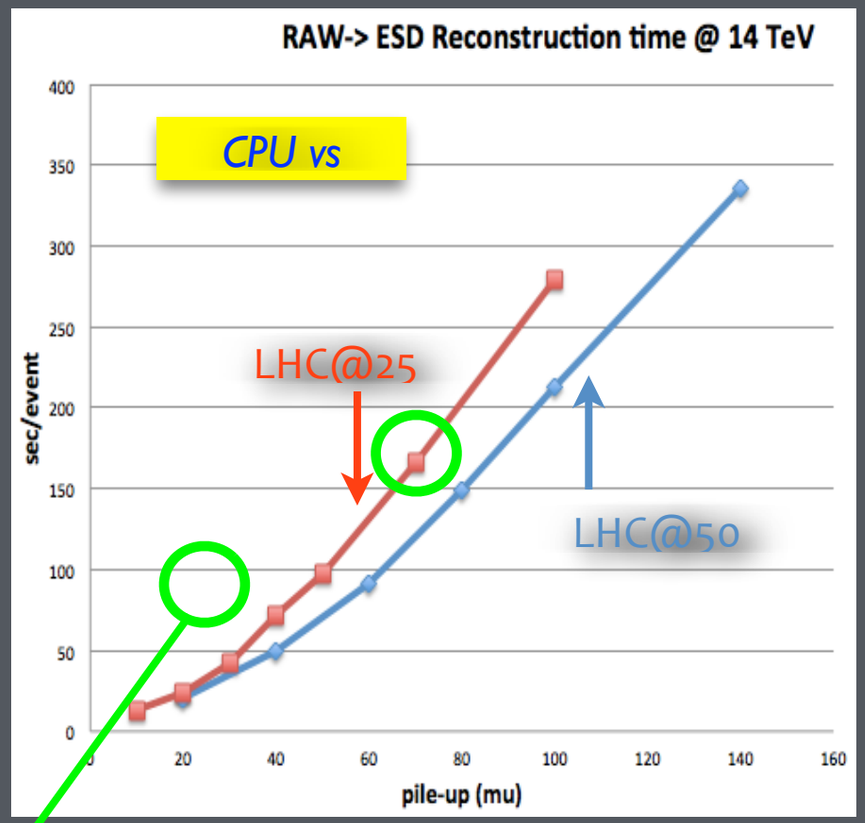
The CPU Explosion

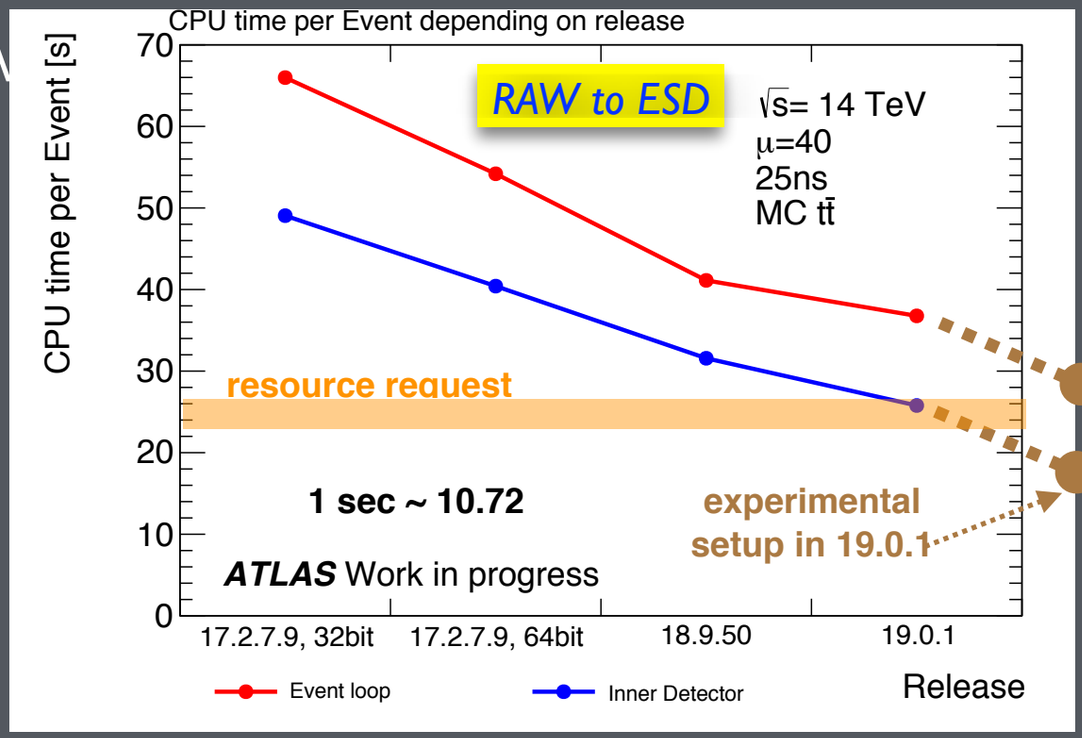
- tracking is driver in CPU vs pileup
  - ➡ combinatorial explosion in hit combinations

- strong tracking LS1 upgrade project
  - ➡ migration to Eigen and simplified EDM (>1000 pkg)
  - ➡ new track seeding strategies (exploring IBL)
  - ➡ squeeze technical performance ...

- preliminary look at CPU
  - ➡ CPU time on 14 TeV, ttbar, mu=40:
    - new compiler, 32→64 bit,
      newest TcMalloc, 4 layer seeding,
      new b-field service, Eigen, simpler Tracking EDM
      and Intel math lib (no auto-vectorising yet, ...)
  - ➡ ~300 HS06/event within reach
    - compared to 250 HS06/event in budget

saved nightly





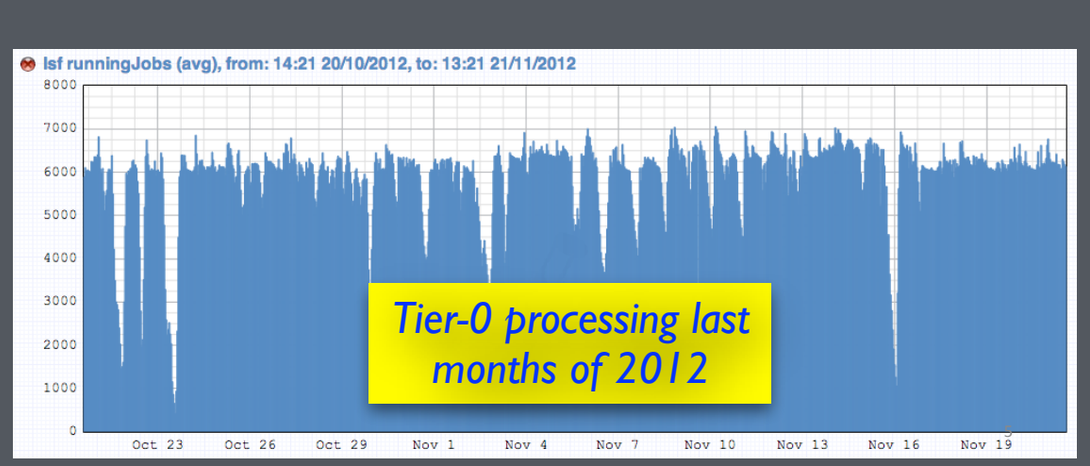
R.Mandrysch

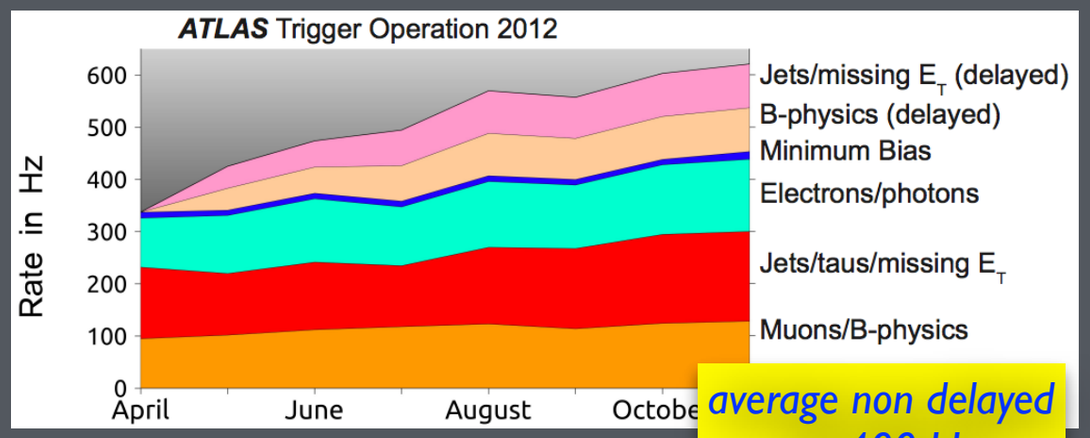
Tier-0 Processing and L4

- 1kHz processing at Tier-0 seems ok with LS1 improvements
  - ➡ not much safety margin, 1.5 kHz would require to add Tier-1 resources
  - ➡ at 50 nsec bunch spacing, we can only afford ~500 Hz

- main gains from new software trigger level L4 are in pileup control on MET and τ
  - ➡ requires full scan tracking
    - New fast tracking in trigger, FTK, may fill in, to be demonstrated
    - fast primary vertex z-scan algorithms?
- We will also consider Deferred Triggers
  - ➡ inter-fill/run processing
  - ➡ MC@Point-1 for inter-fill use of farm ?

- new Analysis Model
  - ➡ staged Tier-0 conflicts with stable L4



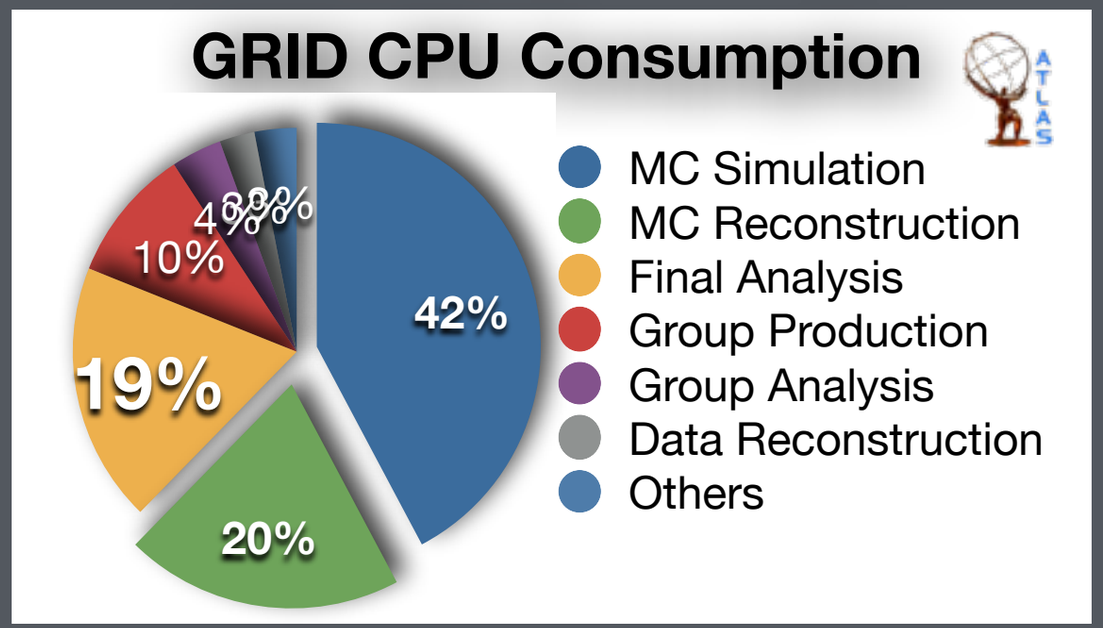


J.Boyd

Resources for Simulation

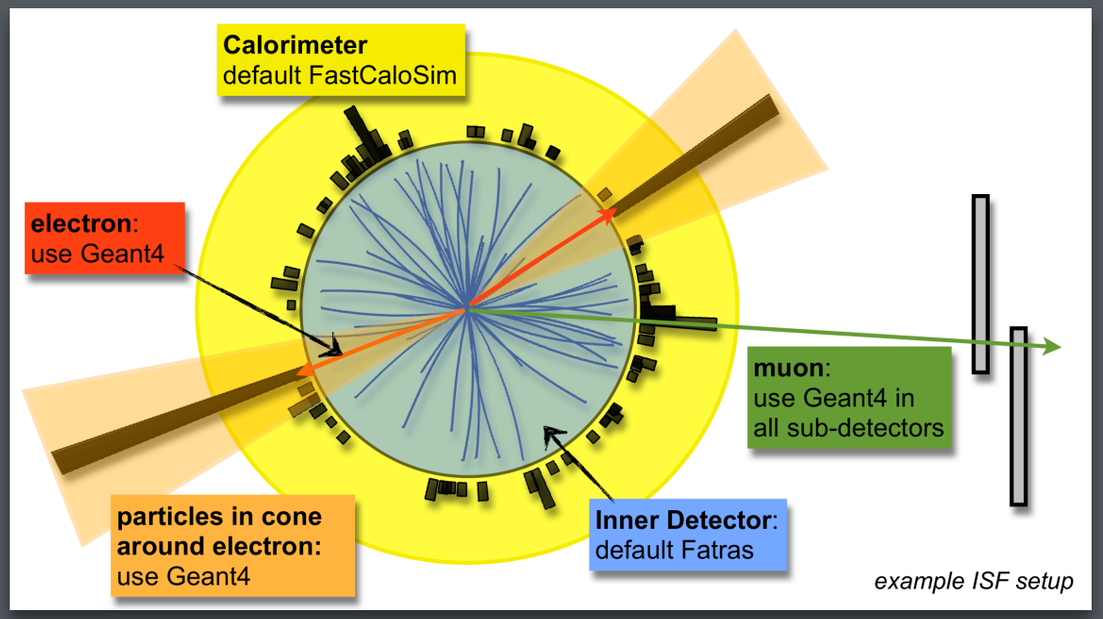
- full Geant4 is a CPU driver
  - ➡ so are many physics generators !
  - ➡ model allows for limited MC statistics
    - full G4 sample of $2*10^9$ events/year
    - plus $5*10^9$ events/year fast simulation
  - ➡ speed improvements in Geant4 ?
    - disk space for more simulation ?

- Integrated Simulation Framework
  - ➡ better/mixed fast simulation options



Tracker	Calo.	Muons	speedup
full	fast	full	~20
fast	fast	fast/full	>100
RoI guided fast/full			~100

- ➡ needs fast digitisation and fast reconstruction (e.g. truth guided tracking)
- ➡ no disk space to save factors more MC
  - only save DxAOD from production



Fast tracking in trigger FTK

- FTK emulation – possible resource driver
  - ➡ currently it takes 9800 HS06/event at pileup of 40!!

- clearly, insufficient resources to include FTK on all MC!
  - ➡ 600 HS06/event budgeted for MC reco. (incl. trigger)
  - ➡ factor 16, if we neglect rest of trigger and offline reco.

- several options
  - ➡ data driven trigger acceptance corrections
  - ➡ use FTK only for very restricted set of triggers
    - This looses a lot of the advantage of the FTK
  - ➡ have a scheme for FTK fast simulation !
    - not yet underway

FTK Emulation chain			
Pileup	Input	Emulation	Reco
40	3200	23900	47?0
60	5750	25300	8900
80	8900	30514	4000

- Units HS06 hours in a H→tautaulh sample with 10k events
- PU 80 sample running on high memory slots

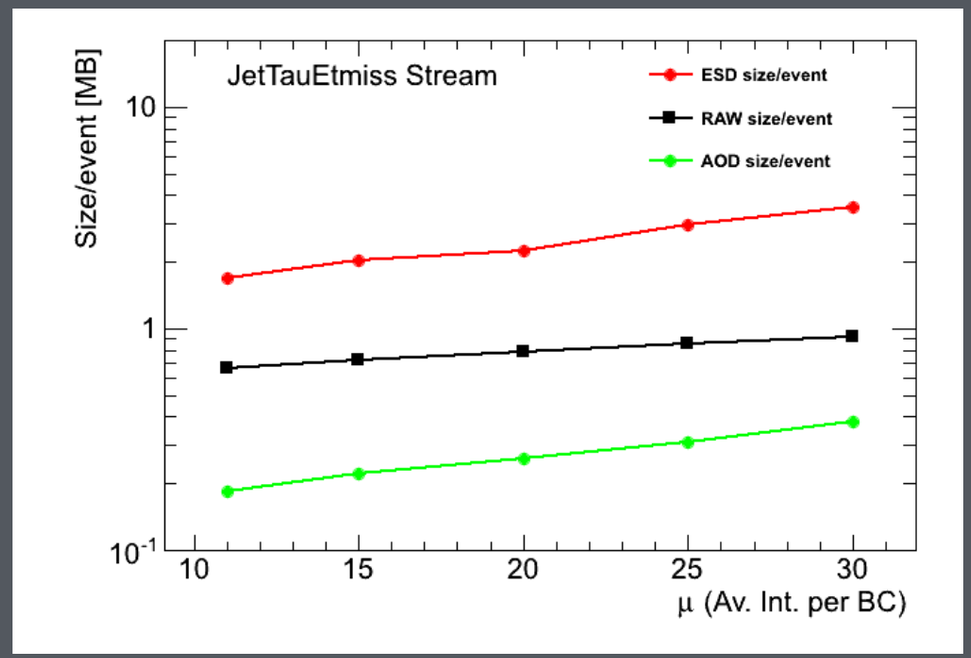
G.Volpi, Marrakech

Event Size in 2012 vs Model

L.Fiorini

- experience from 2012
  - ➡ Naive scaling to 13 TeV, μ=40 @ 25 nsec
    - uses JetTauEtmiss stream (20% larger than average !)

size (MB) at 40 pileup	RAW	ESD	AOD
"guess"	~ 1.1	~ 5.0	~ 0.55
model	1.0	2.7	0.35



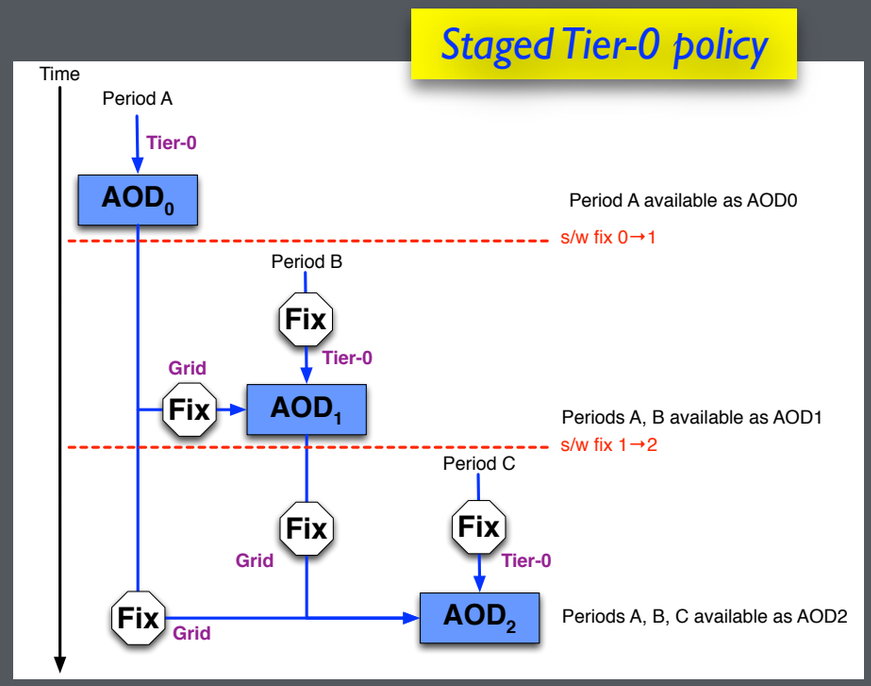
- ➡ Not quite matching model yet
  - full reprocessing/analysis formats not yet available
- ➡ MC: need to add truth and Trigger
- What analysis (xAOD) sizes to expect ?
  - ➡ we see ~same size for same content as before
    - affected by changes in tracking settings
    - some xAOD objects with ESD type of extensions
    - possibly additional jet collections ?
  - ➡ (x)AOD size is currently a worry

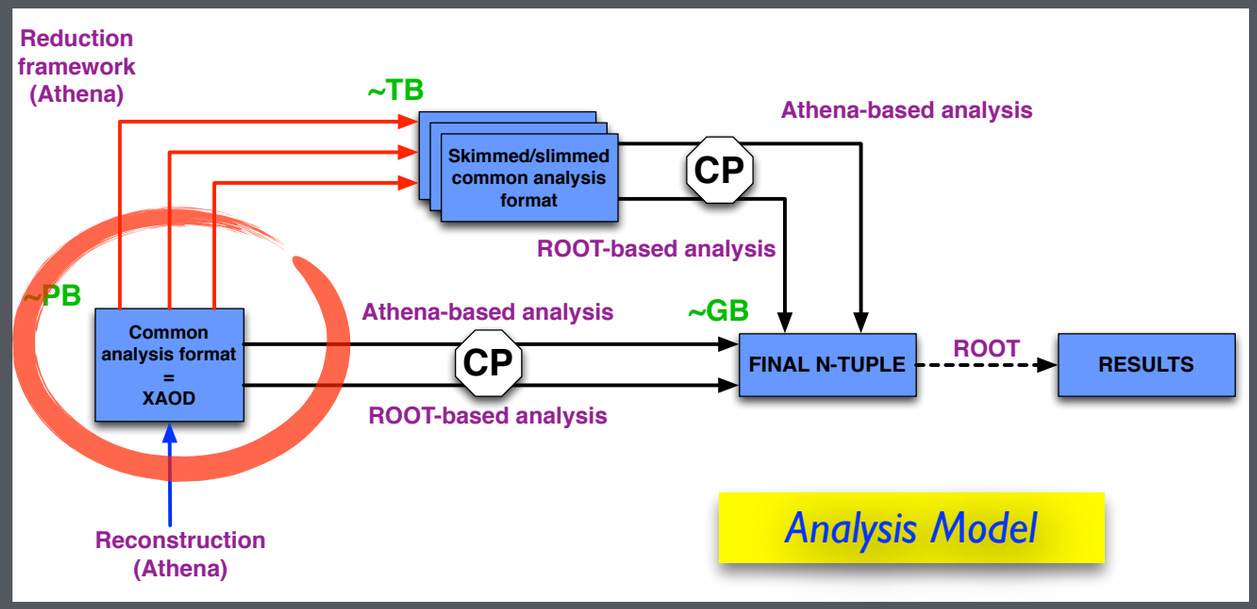
size (MB) vs pileup	RAW	ESD	AOD
11	0.66	1.69	0.185
15	0.72	2.04	0.222
20	0.78	2.26	0.258
25	0.85	2.96	0.309
30	0.92	3.56	0.380

L.Fiorini

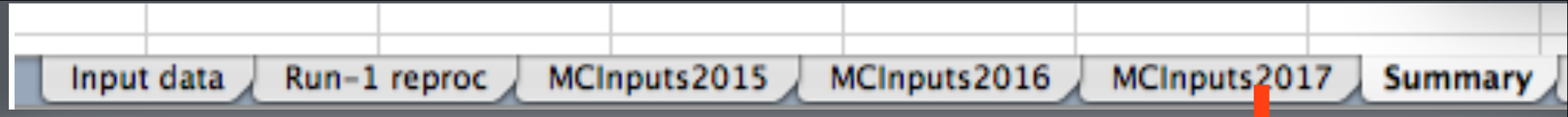
The Analysis Model "Revolution"

- based on AMSG report last year
  - ➡ staged Tier-0 policy with AOD→AOD reprocessing
    - fix data in Tier-0 and not in group production
    - final AOD plus previous versions for old periods
    - inflates the AOD size by factor 1.7 (+1 for reproc.)
  - ➡ new Analysis Model
    - xAOD format directly root readable (eliminate D3PD)
    - reduction framework produces DxAOD in trains
    - analysis releases with Combined Perf tools, examples...
- D3PD overheads in 2013
  - ➡ factor 3.5 (!) in size per AOD version
    - multiplied by revision (fix version) kept
    - sitting on group disks
    - model has even larger group disks (!)
  - ➡ D3PD added a lot to total CPU needs
    - was heavy on jets reconstruction, ...
    - staged Tier-0 addresses this



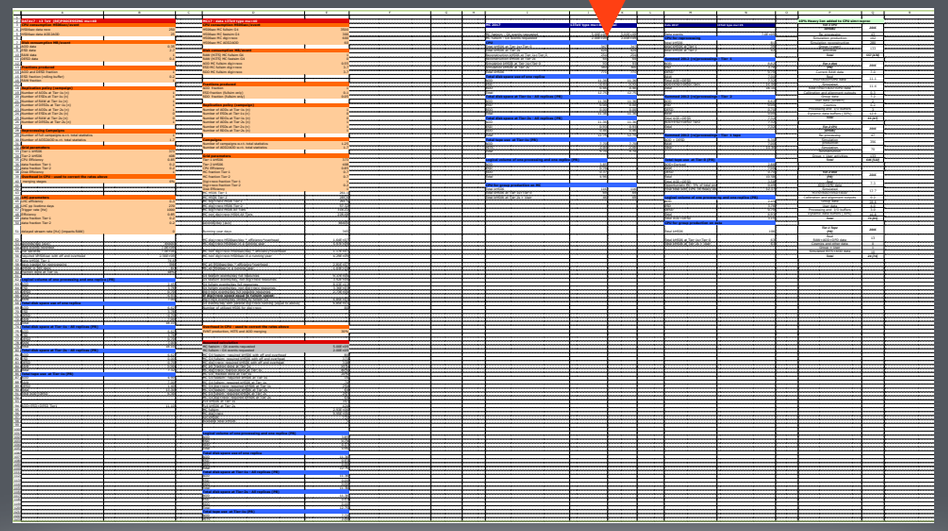


➡ success on new Analysis Model is vital (DC-14)



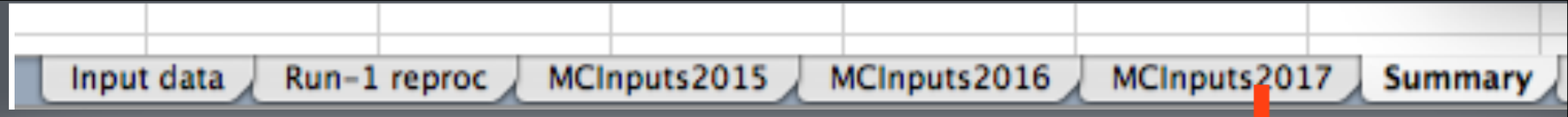
Model Assumptions for projections

B.Kersevan



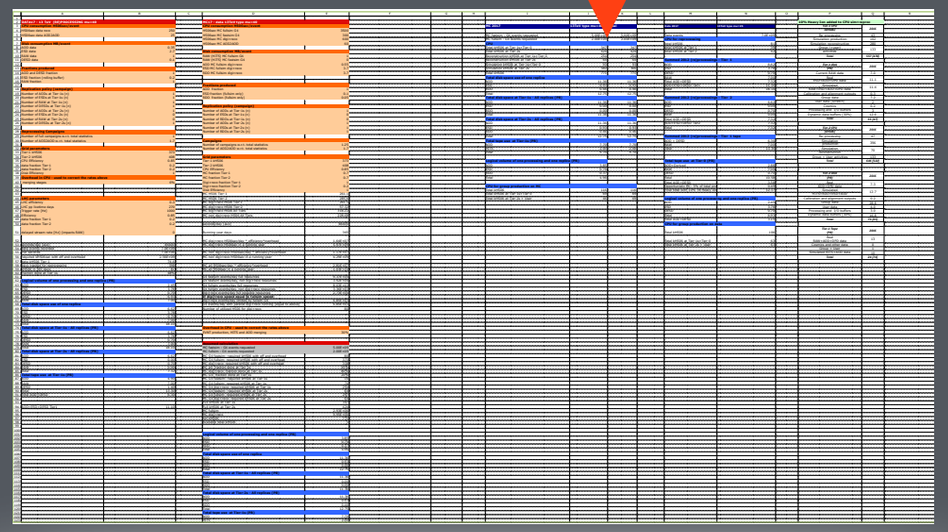
- CPU and event sizes not far off
  - ➡ Tier-0 processing at 1 kHz is not a problem
  - ➡ (x)AOD sizes are a worry, but time to improve
  - ➡ assumed MC sample sizes seem small compared to past experience
    - ISF may give us technology to do much more
    - and FTK emulation is an unknown
  - ➡ new Analysis Model needs to be a success

- now one needs to add choices to the model:
  - ➡ 1 reprocessing per year, 2 AOD→AOD fixes
  - ➡ one can increase/reduce replication, move data to tape ...
    - in a sense, the 20% PD2P disk buffer is a choice (so far not realised)
  - ➡ one needs to define numbers of versions and replicas on disk all data formats



Model Assumptions for Projections

B.Kersevan



- model for versions and replicas
  - ➡ emphasis is on current year data
    - very much like in Run-1, still true in Run-2 ?
  - ➡ older data archived to tape, i.e.:
    - raw data only kept 1 year
    - no AOD on disk from Run-1 after 2015, no 2015 AOD on disk in 2017(!!)
    - but: analysis trains in central production could run off tape
  - ➡ group data as well gets archived to tape after 2 years

would not be by choice

well, not easy

	data RAW	data ESD	data AOD	data DESD	MC HITS	MC RAW	MC ESD	MC AOD	Group DxAOD
current year	1	0.2	2 * 2.7	2	-	0.06	0.125	2 * 2.9	~ 5
previous year	-	-	1	1	-	0.06	0.125	1	1.5-3

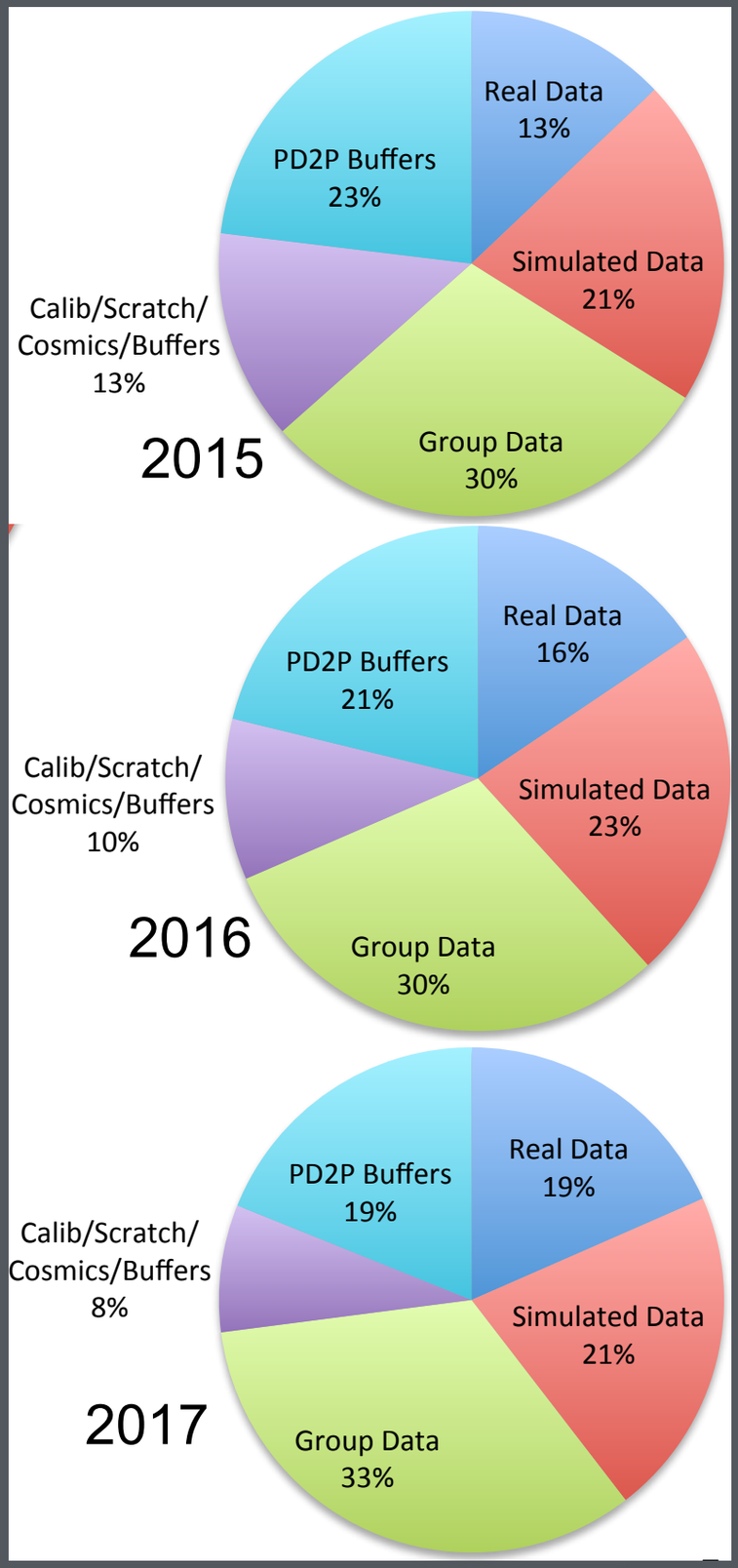
- model is not cast in stone, room and great need for optimisation

Increasing caching, less placement

Disk Usage Summary

- predicted total disk sizes per data type:

in PB	2015	2016	2017
Real Data	12.2	19.2	28.7
Simulated Data	19.8	28.0	32.9
Group Data	27.8	37.3	52.0
Calib/Scratch/ Cosmics/Buffers	12.7	12.7	12.7
PD2P Buffers	21.8	26.1	29.4



- ➡ relative shares don't change much
  - real data increases following integrate lumi. per year
  - PD2P buffers and additional samples slowly reduced
- ➡ group data in model is huge
  - real and simulated data are 35–40% of the total

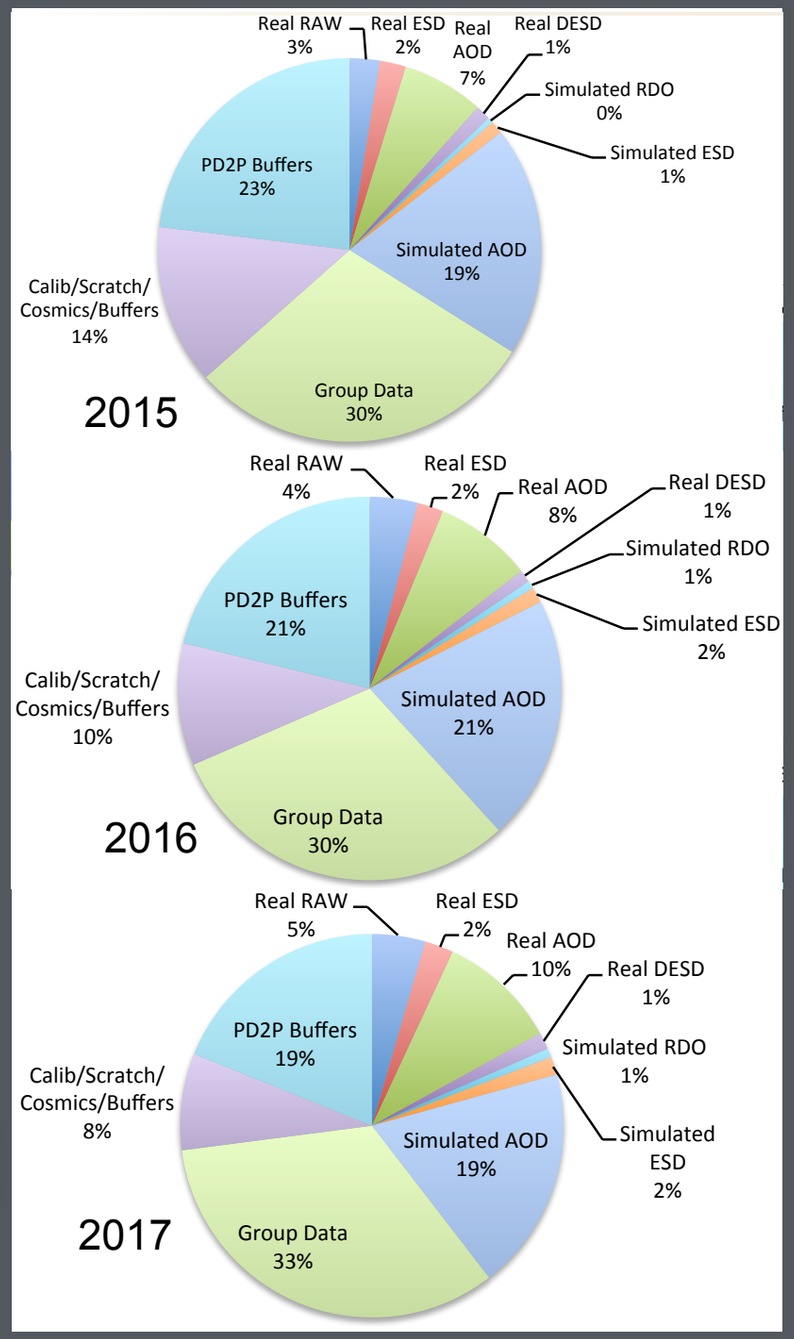
This requires event picking, increased use of network

R.Mount, B.Kersevan

Disk Usage Details

- predicted disk usage per format:

	2015	2016	2017
Real RAW	2.4	5.0	7.0
Real ESD	2.1	2.7	3.8
Real AOD	6.5	10.2	15.7
Real DESD	1.2	1.3	2.2
Simulated RDO	0.4	0.8	1.2
Simulated ESD	1.0	1.6	2.4
Simulated AOD	18.3	25.5	29.4
Simulated HITS			
Group Data	27.8	37.3	52.0
Calib/Scratch/Cosmics/Buffers	12.7	12.7	12.7
PD2P Buffers	21.8	26.1	29.4



➡ Implied network requirements to be clarified in DC14
➡ Expect largest Tier 2s to reach 40Gbps later in run2
➡ Expect average network doubling ~2years for largest sites

R.Mount, B.Kersevan

Disk Usage Details

- predicted disk usage per format:

	2015	2016	2017
Real RAW	2.4	5.0	7.0
Real ESD	2.1	2.7	3.8
Real AOD	6.5	10.2	15.7
Real DESD	1.2	1.3	2.2
Simulated RDO	0.4	0.8	1.2
Simulated ESD	1.0	1.6	2.4
Simulated AOD	18.3	25.5	29.4
Simulated HITS			
Group Data	27.8	37.3	52.0
Calib/Scratch/Cosmics/Buffers	12.7	12.7	12.7
PD2P Buffers	21.8	26.1	29.4
total	94.2	123.2	155.8

- single replica:

size (PB)	7 TeV	8 TeV
AOD	0.3	1.0
MC AOD	0.9	0.9
sum	1.2	1.9

size (PB)	2015	2016	2017
RAW	2.4	5.0	7.0
ESD*0.2	1.5	2.7	3.8
AOD	0.8	1.8	2.5
DESD	0.3	0.5	0.7
MC RDO	0.4	0.4	0.4
MC ESD*0.125	0.7	0.8	0.8
MC AOD	2.8	3.9	3.9
sum	8.9	15.1	19.1
plus AOD	-	18.7	28.4
plus Run-1	12.0	21.8	31.5

➡ data+MC of 2017 is 19.1 PB total, (x)AOD of all other years is 12.4 PB together

➡ discussion in OAB starting on what can be optimised

Summary

- Software and Computing preparation for Run-2
  - ➡ baseline is to prepare the offline for 1 kHz data taking rate

- "strawman" model stays within expected resources
  - ➡ numerous software upgrades to make it reality:
    software speedup, new Analysis Model, ISF simulation framework, ...
  - ➡ CPU looks good, event size is currently a worry but early to say

- overall there is room for optimisation
  - ➡ "strawman" model is putting emphasis on data of the year
    - was true in Run-1, same fast turnaround in Run-2 ?
  - ➡ technical improvements will give us more freedom (e.g. MC)
    - but manpower is short

- new Analysis Model needs to be a success
  - ➡ we can not afford group disk for old D3PD production model

- Overall model tested Summer/Autumn 14 in DC14