



# Grid operations in 2016

T1/T2 workshop – 7<sup>th</sup> edition - Strasbourg

3 May 2017  
Latchezar Betev

# T1/T2 workshops



7-th edition in Strasbourg

\*towers are shorter  
than they appear

# The ALICE Grid today



# New sites

- Basically unchanged since last year
- One more returning site – UPB in Romania
- Altaria is now a Cloud catch-all (see Maarten's talk)

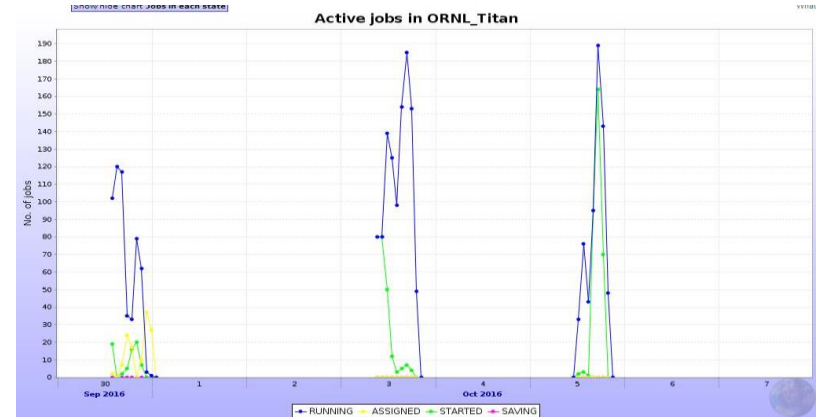


=>



- Still no volunteering sites in Australia/New Zealand

# New Supercomputers



- Set of network-free site services ready
- Substantial amount of work to rid application software of network calls
- Successful test with full chain (Monte-Carlo)
- Several issues remain
  - Access necessitates hand-holding (token generation)
  - Backfill yields many short time (minutes) slots, MC jobs are longer
  - The AMD CPUs are almost 2x slower than the average Grid CPU
  - We can't use the abundant GPUs
- To use effectively, we will need a time allocation
- To be continued...

“  
Do not use a cannon to kill  
a mosquito.”



Confucius

# Job record – broken again

April 2016



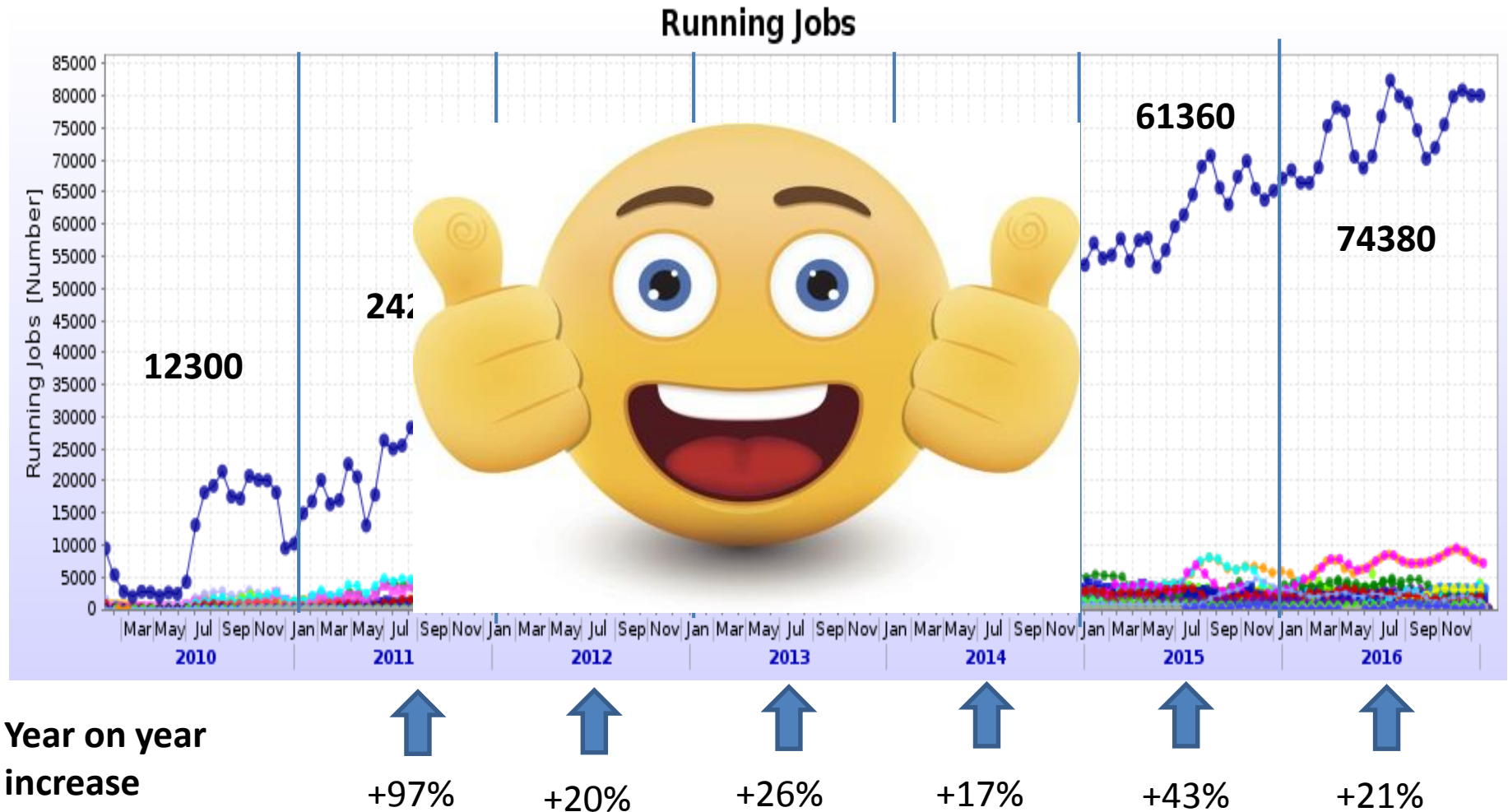
=>

April 2017

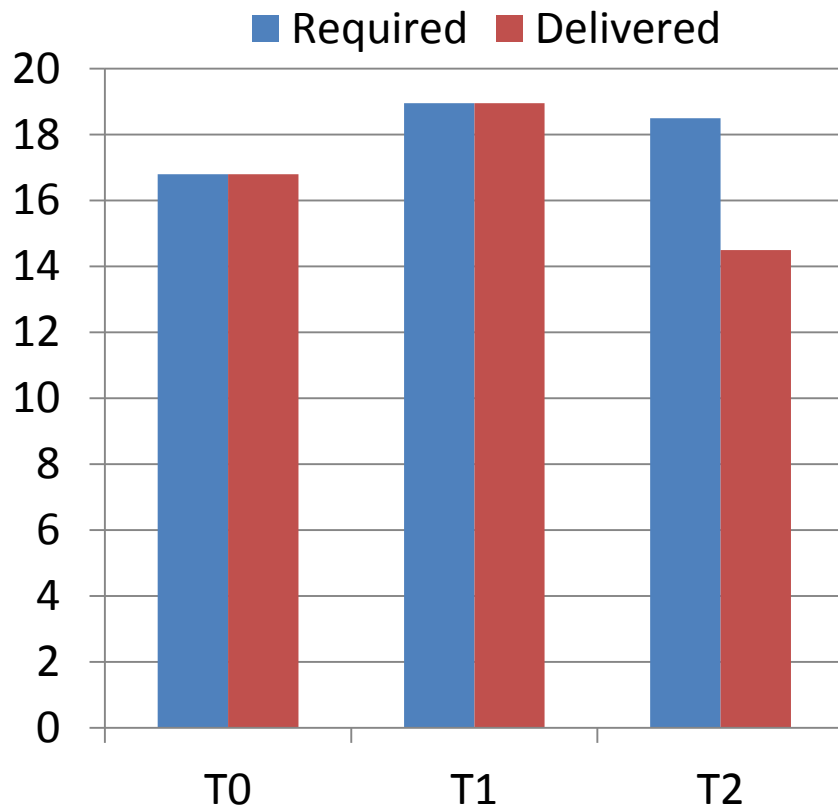


\* This is not the highest we had, but you get the idea

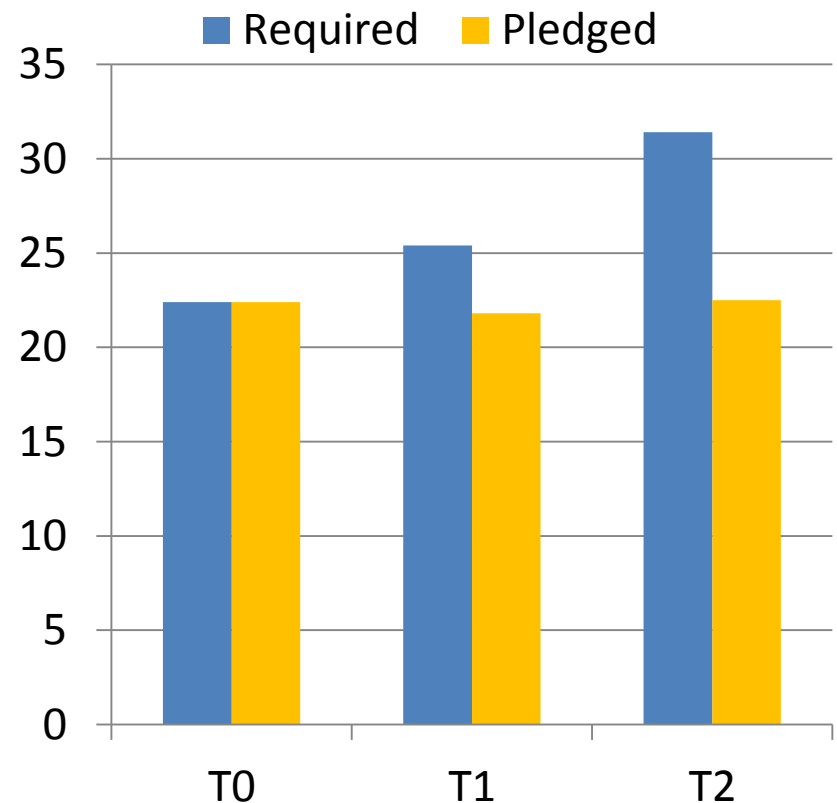
# CPU resources evolution



# Disk Resources evolution



2016: T2 deficit 4PB



2017: T1 deficit 4PB

T2 deficit 9PB

Total 16% disk deficit (if 2016  
T2 deficit is recuperated)

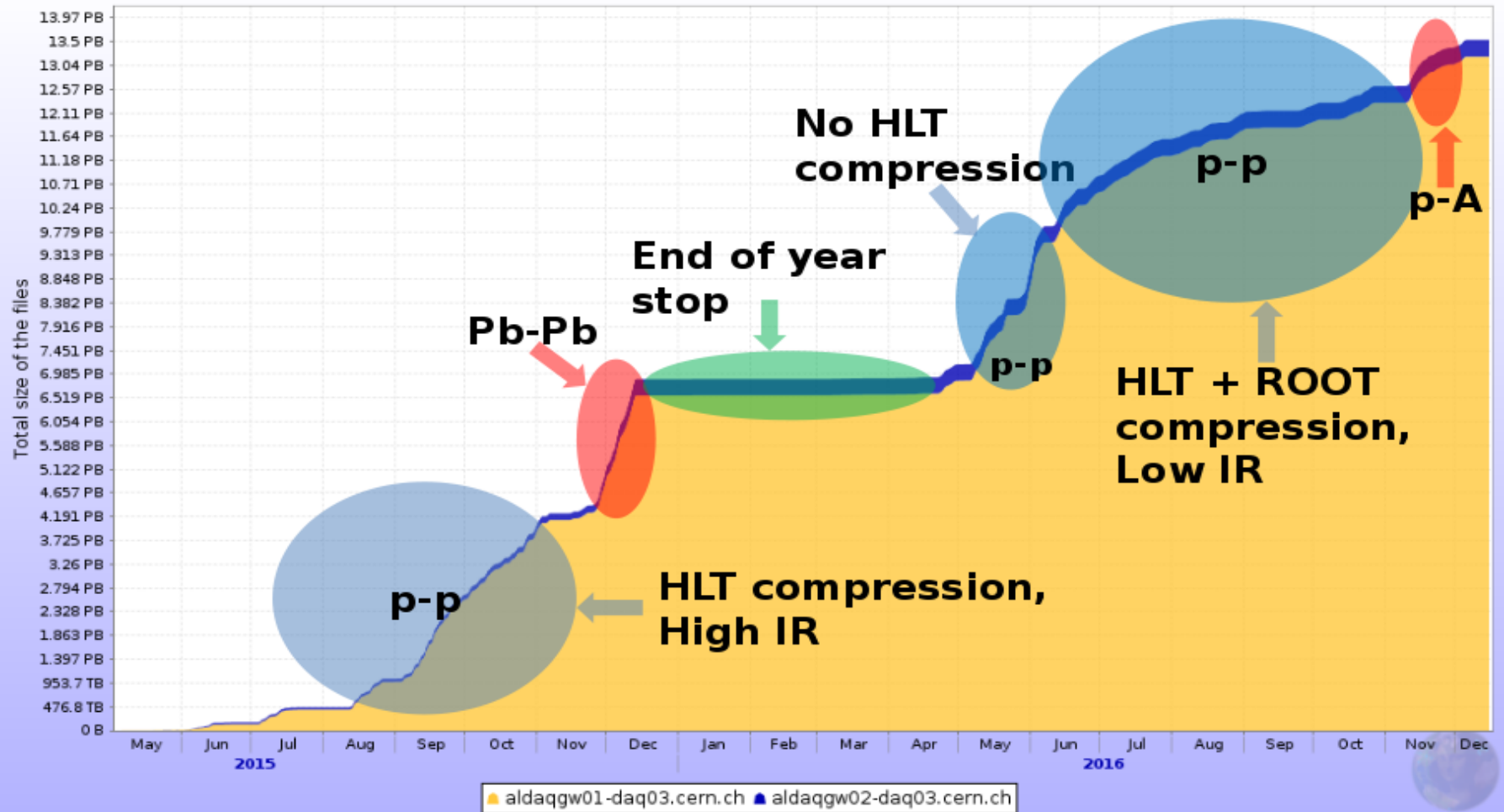
More details – Yves' talk on Friday

# Reminder on replicas

- Tapes – only RAW data (2 copies)
- Derived data of active datasets
  - ESDs – 1 copy
  - AODs – 2 copies
- Inactive datasets – 1 copy
- Regular cleanups (unpopular and dark data) done and no more freedom to reduce replicas
  - => Disk deficit will be increasingly difficult to compensate with ‘technical’ means

# 2015+2016 RAW data collection

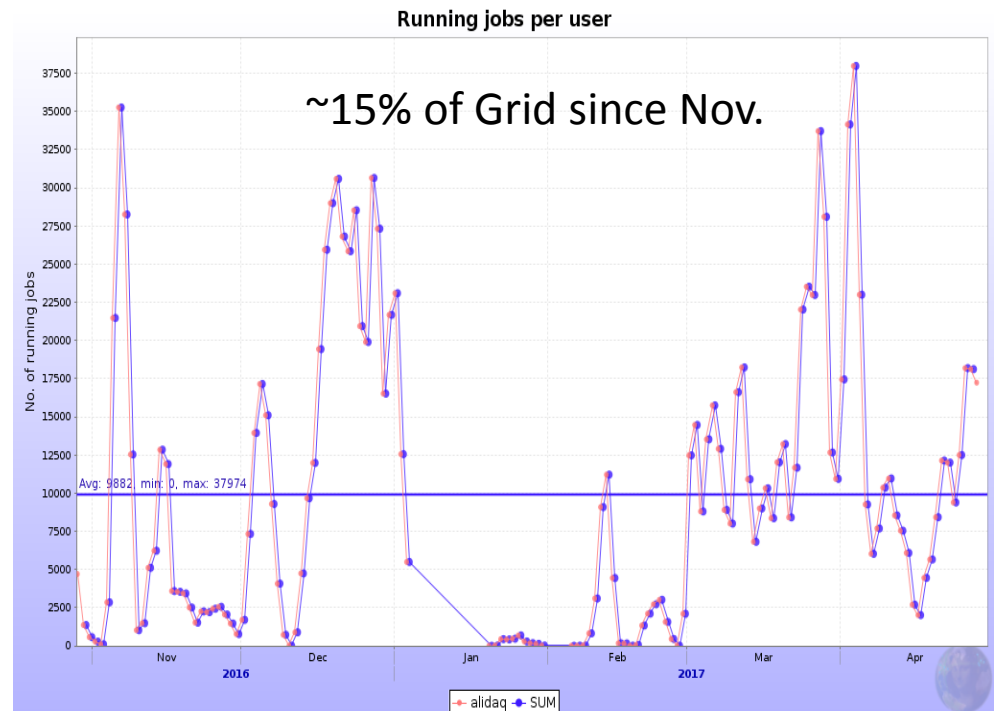
Total size of the files



# Status of RAW data processing

- Substantial IR-induced distortions in the TPC
- Offline correction algorithms (embedded in CPass-0/CPass1) developed and validated

- Reco of 2016 RAW data @90%, 2015 data @60%
- Completion expected by June 2017



# Status of MC processing

- Numerous general-purpose and specialized MC cycles anchored to 2015 and 2016 data
  - Total of 135 (~same as in 2015)
  - 2,897,020,717 events
- RAW and MC productions are now handled by the Data Preparation Group (DPG)

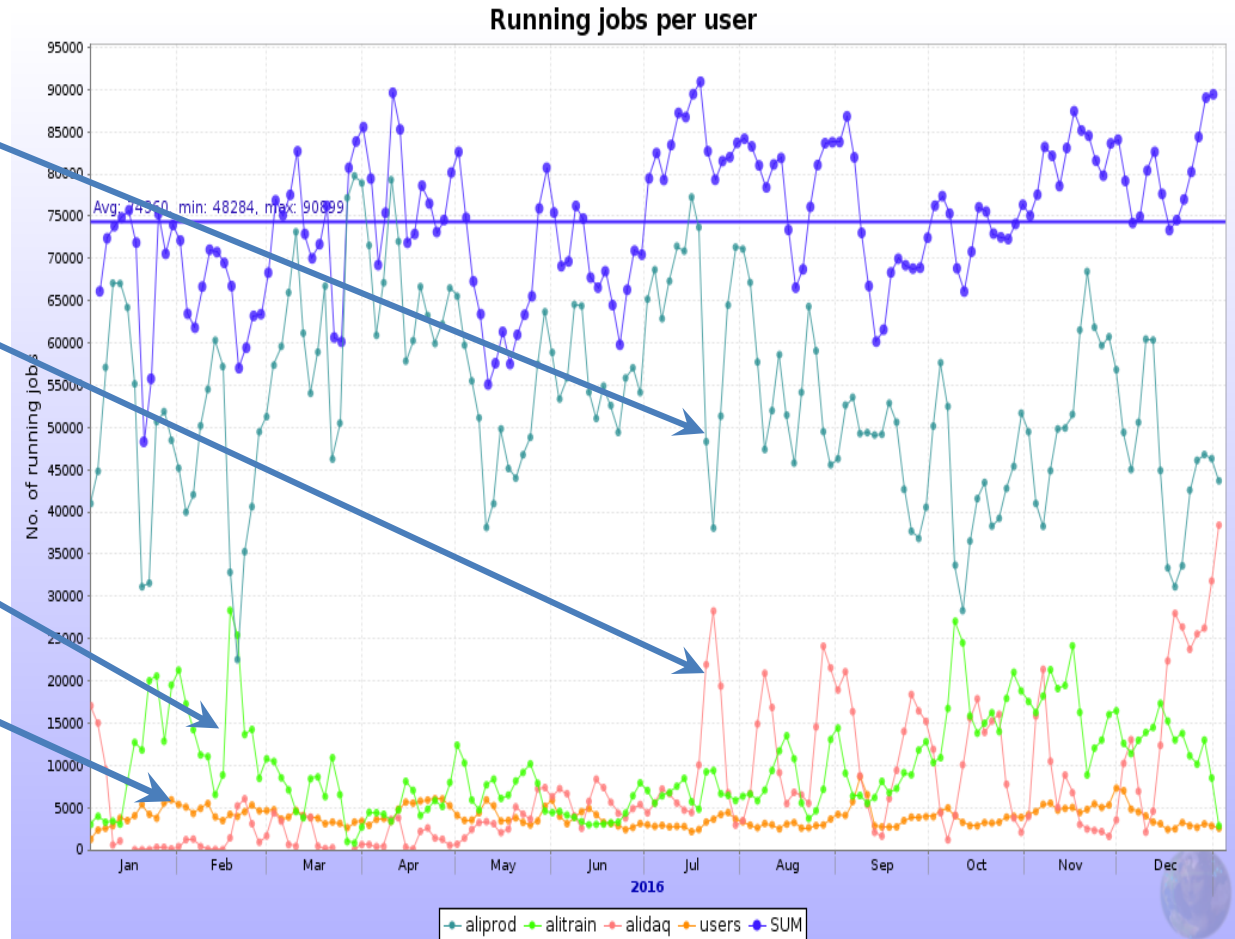
# Resources sharing

**MC productions: 72%  
@all centres**

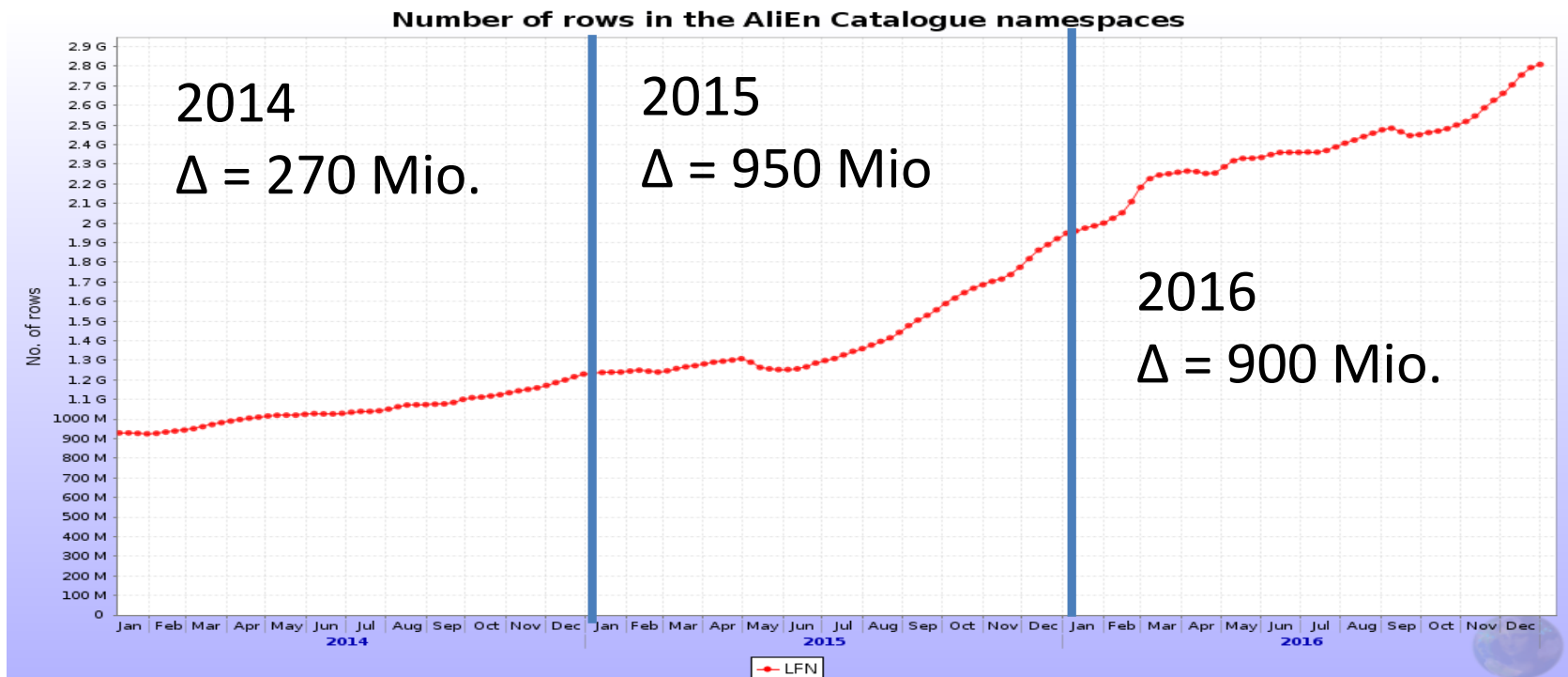
**RAW data processing:  
10% @ T0/T1s only**

**Organized analysis:  
13% @all centres**

**Individual analysis: 4%  
@all centres, 480 users**

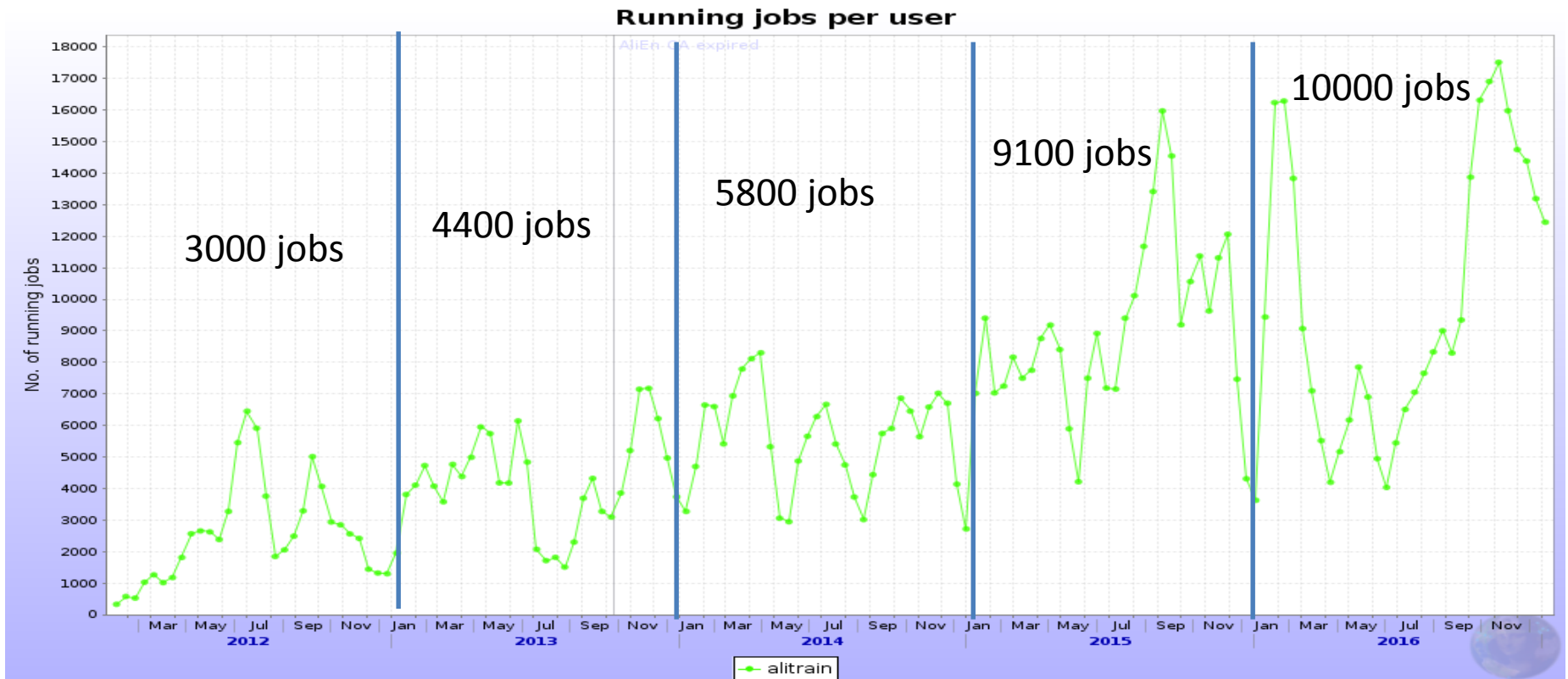


# Catalogue stats



- A more dramatic increase mitigated by cleanup
  - That said – January-April 2017: +500 Mio new files
  - RAW data reconstruction is a big generator
  - New catalogue in the works (Miguel's presentation)

# Organized analysis



Year on year increase

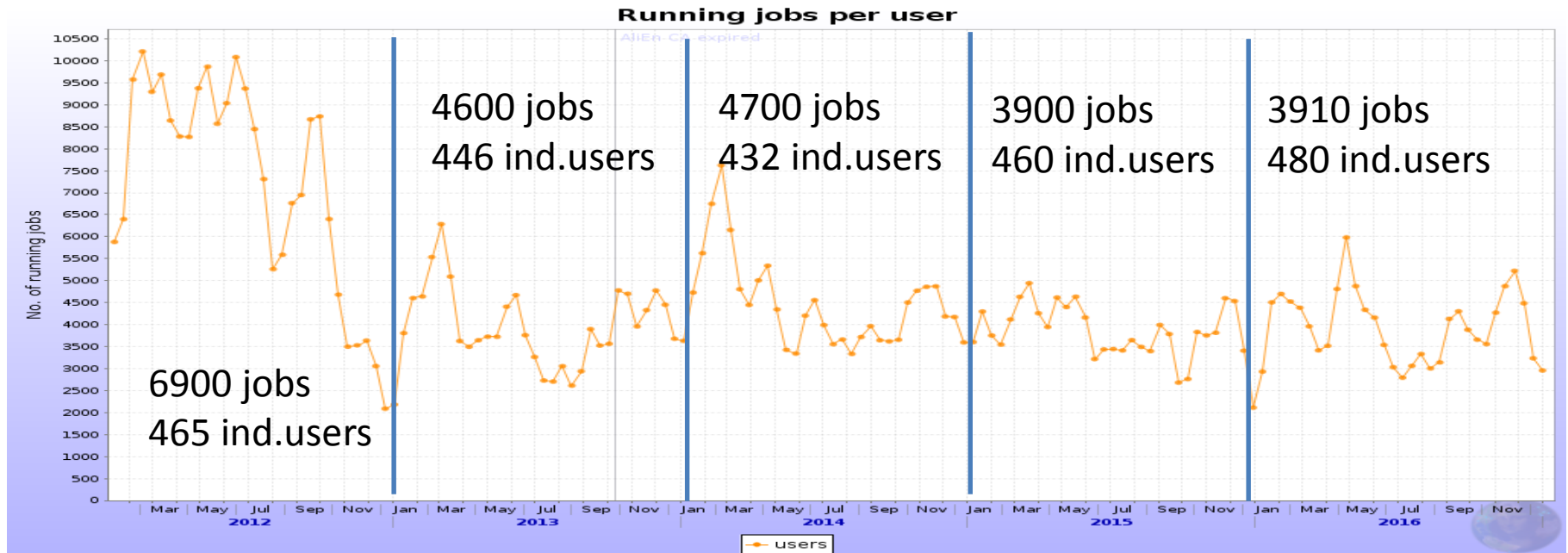
+47%

+32%

+57%

+10%

# Individual analysis



Year on year increase  
Individual analysis

**-50%**

+3%

**-17%**

0%

Year on year increase  
organized analysis

+47%

+32%

+57%

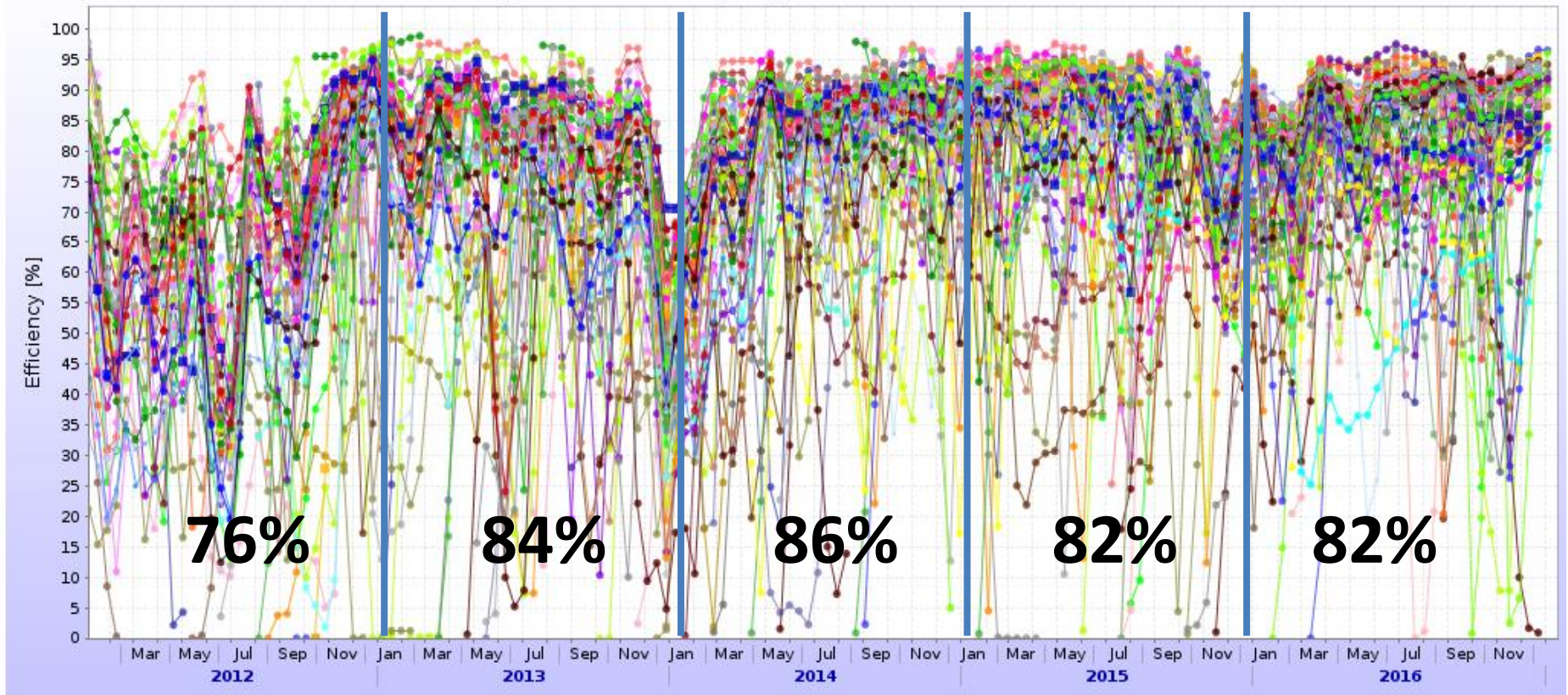
+10%

# Analysis evolution

- In the past 2 years
  - Stable number of users
  - Resources used by individual analysis are also stable
- The organized analysis share keeps increasing
  - Not surprising – more data and more analysis topics
  - This also brings increased load and demand on the storage

# Grid efficiency

Jobs efficiency (cpu time / wall time)



Year on year change

↑  
+8%

↑  
+2%

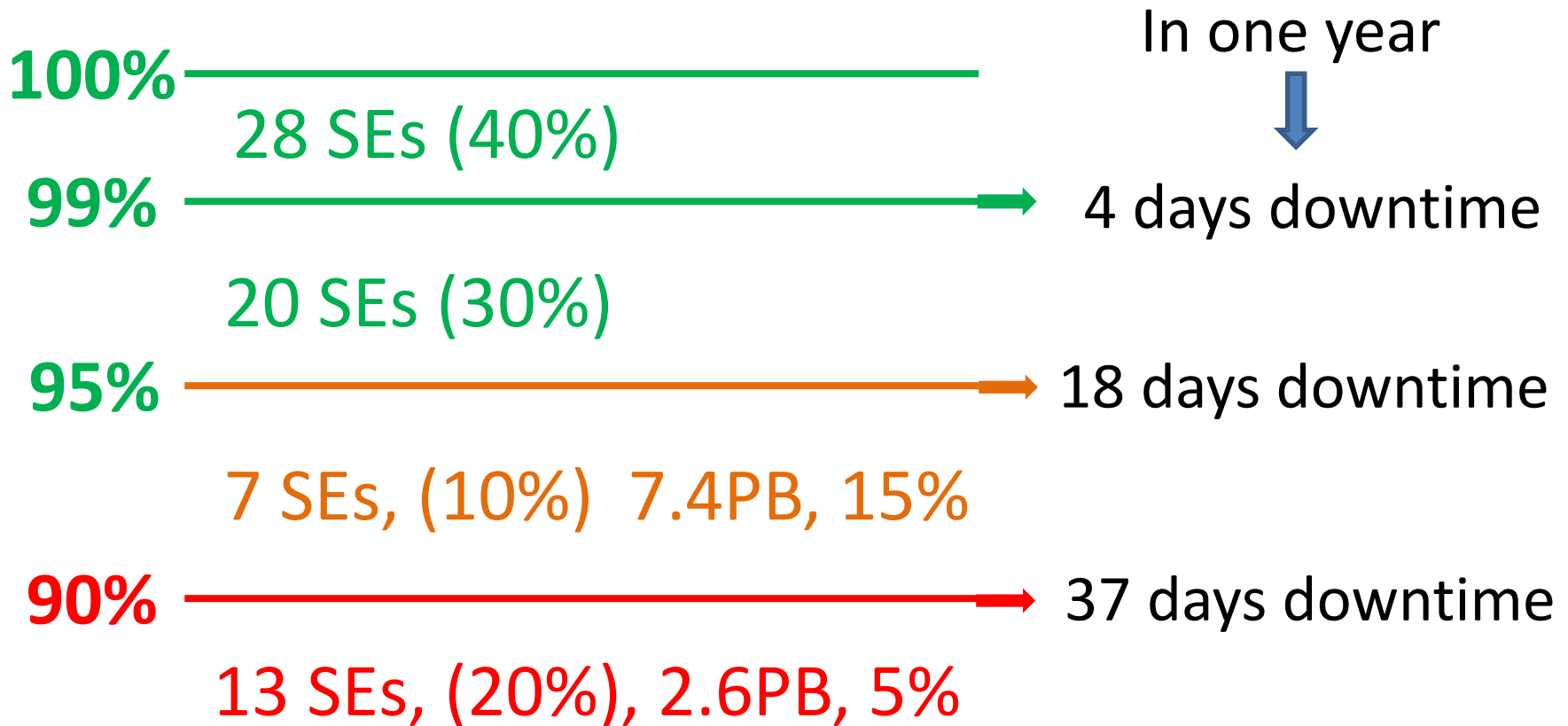
↑  
-4%

↑  
0%

# Grid efficiency evolution

- Efficiency is ~flat
- Specific effort to increase the efficiency of calibration tasks
  - Part of the new calibration suite development
  - With substantial increase of analysis, the efficiency is not degraded!
- In general, ~85% efficiency is perhaps the maximum we can expect
  - No manpower to work on this (increasingly difficult to improve) topic

# Storage availability



- Total number of SEs = 68 (including tape),
- Only disk = **48PB** total available space

# Storage availability evolution

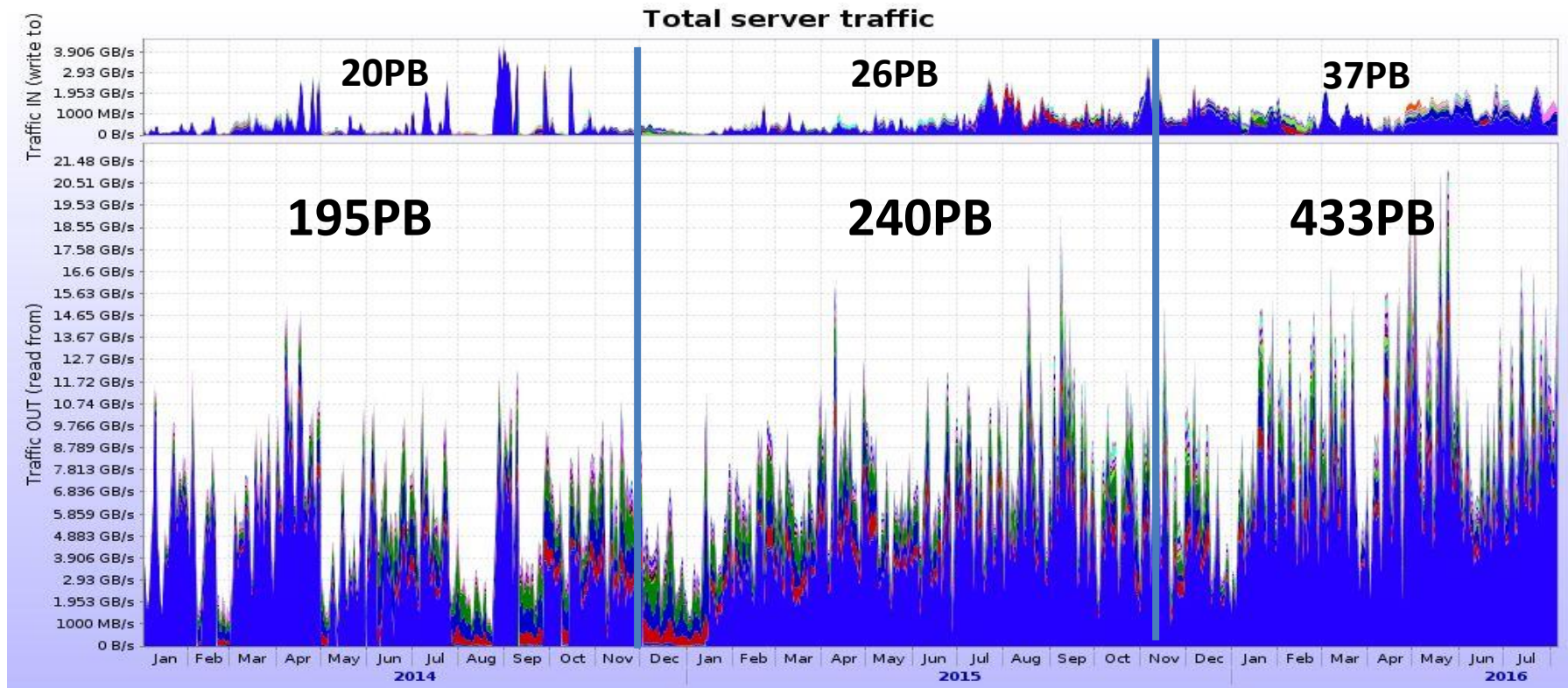
- Significant capacity still below the target 95% availability
  - 20% of the current 48PB
  - Tools for the sysadmins exist to follow up on the problems
  - Despite that, we have to send messages ‘look at your storage’
  - Question on what to do with <90% SEs – should these be ‘transferred’ to other sites?
- Reminder - replica model is 1 ESD copy only
  - Thanks to fully distributed model we have ~5% of data loss
  - This is at about the ‘pain’ threshold for analyzers

# New storage capacity in 2017

Size	Used	Free	Usage
185.2 TB	183.4 TB	1.801 TB	99.03%
7.563 PB	7.479 PB	86.35 TB	98.89%
2.267 PB	2.239 PB	28 TB	98.79%
20.01 TB	19.75 TB	266.2 GB	98.7%
1.972 PB	1.945 PB	27.15 TB	98.66%
3.405 PB	3.334 PB	72.88 TB	97.91%
1.128 PB	1.101 PB	27.83 TB	97.59%
829.2 TB	808.8 TB	20.4 TB	97.54%
1.446 PB	1.402 PB	45.27 TB	96.94%
669.3 TB	647.2 TB	22.11 TB	96.7%
263.3 TB	252.3 TB	11 TB	95.82%
418.4 TB	398.4 TB	19.98 TB	95.23%
1.591 PB	1.509 PB	84.05 TB	94.84%
1.1 PB	1.039 PB	62.16 TB	94.48%
319.8 TB	299.3 TB	20.49 TB	93.59%
421.5 TB	394.4 TB	27.11 TB	93.57%
203.9 TB	189.6 TB	14.31 TB	92.98%
298.3 TB	275 TB	23.33 TB	92.18%
12.2 PB	11.11 PB	1.091 PB	91.06%
289 TB	262.2 TB	26.78 TB	90.73%
37.3 TB	33.81 TB	3.487 TB	90.65%
265.6 TB	238.8 TB	26.78 TB	89.92%
270.7 TB	242.8 TB	27.95 TB	89.68%
48.85 TB	43.75 TB	5.101 TB	89.56%
1.339 PB	1.191 PB	151.7 TB	88.94%
593 TB	524.3 TB	68.71 TB	88.41%
180.1 TB	152.4 TB	27.71 TB	84.61%
617.5 TB	501.2 TB	116.3 TB	81.17%
30.92 TB	25 TB	5.92 TB	80.86%
27.99 TB	22.05 TB	5.937 TB	78.79%
76.39 TB	59.12 TB	17.27 TB	77.4%
894.1 TB	691.4 TB	202.7 TB	77.32%
1.012 PB	800.7 TB	235.6 TB	77.27%
61.84 TB	43.87 TB	17.98 TB	70.93%

- 43/48 PB disk used (90%)
- Cleanup ongoing, but productions are filling faster...
- We need the 2027 capacity installed, especially on SEs which are  $\geq 90\%$  full
- Presently, these are 24 of 58 and many large >1PB SEs

# Storage use



**Year on year change**

Write +30%  
Read +23%  
Ratio 9.2

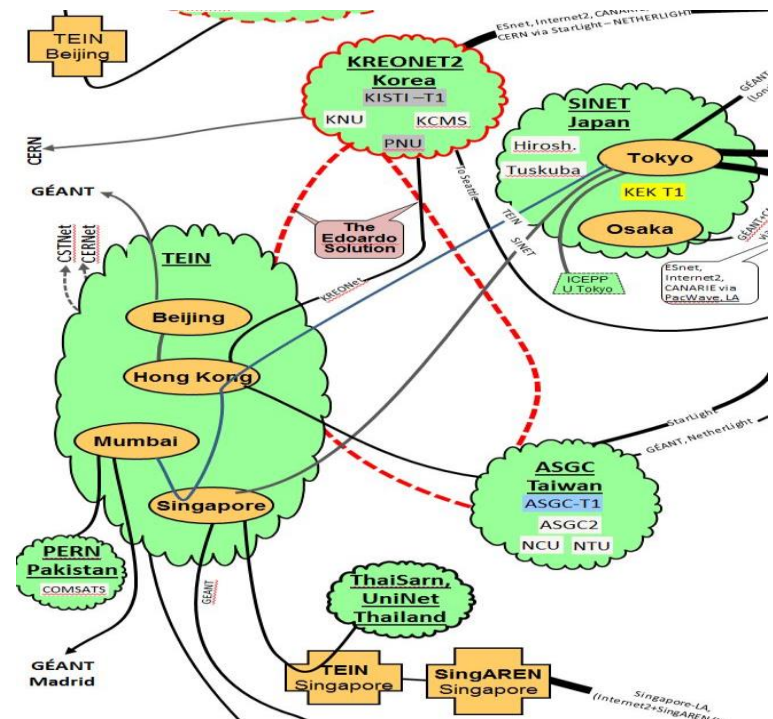
Write +42%  
Read +80%  
Ratio 12

# Storage use evolution

- Read volume increase
  - Preparation for QM'2017 (passed) and now SQM'2017
  - Increased number of analysis trains + read data volume
- Positive aspects and points to worry about
  - No perceptible degradation of overall efficiency
  - The storage is not (yet) saturated, but there will be an inflection point
    - Large capacity disks => lower i/o per unit – something to worry about
    - Effect seen on loaded SE units – tape buffer
- Cleanup
  - 'Orphan files' < 1%, verified periodically
  - Deletion of files and replica reduction – constant activity

# The Fifth Element - Network

- Still one step ahead (Europe/USA)
  - Within the present computing model
- Large growth of computing capacities in Asia
  - Regional network - routing to be improved
  - Now followed at the Asian Tier Center Forum (ACTF)
- South Africa – same issue and active follow-up



# Summary

- 2016 was a great year for Grid operations
- The impressive data taking continues, so does the pressure on the computing resources
- Efficiency remains high, site stability remains high
  - Storage stability of few sites must be improved
- The ALICE upgrade is one year closer
  - Development of new software is accelerating
- ***On track for the third year of LHC Run2***

**Thank you** to all who contributed to 2016  
being another great Grid year!

Thanks to all speakers

For those who still did not do it – please upload your  
presentations