# Attempt at Estimating Archival Bandwidth Needs

Frank Wuerthwein

UCSD/SDSC

DOMA Access Meeting

September 29th 2020

# Disclaimers

- This is a very crude first attempt at estimating needs.
- It is meant as something for people to think about in preparation for the storage workshop in November.
- It does not include any overprovisioning to arrive at deployed hardware needs to be able to provide the usable bandwidth targets described here.

# Outline of Talk

- Summarize the logic for the needs estimates for ATLAS and CMS
- From the sum of the two experiment’s global numbers calculate the bandwidth for each T1.

# Basic Logic for CMS

- Take the current data volume estimates.
  - Make some arbitrary but reasonable assumption about how the archive will be used, and assess which use cases are likely dominating the needed IO bandwidth.

  => This leads to a total IO aggregate need summed up over all T1s.
- Take the 2020 Tape pledge % for each T1
- Calculate IO per T1 based on the % pledge of 2020

# CMS Volume estimate

- HLT output rate = 7.5kHz
- Total # of RAW events per year = 56 Billions
- Total # of MC events per year = 64 Billions
- RAW evt size = 6.5 MB => 364PB/y
- AOD evt size = 2 MB => 240PB/y RAW+MC
- MINI evt size = 0.25 MB => 30PB/y RAW+MC
- NANO evt size = 0.002 MB => 0.24PB/y RAW+MC

# CMS Annual RAW processing

- Assume it gets done in 100 days
- Coming off the archive: 364PB/100 days = 44 GB/sec ~ 400Gbit/sec
- Going into the archive: 112PB/100 days = 14 GB/sec ~ 130Gbit/sec
- Rounded up generously ~ 550Gbit/sec total

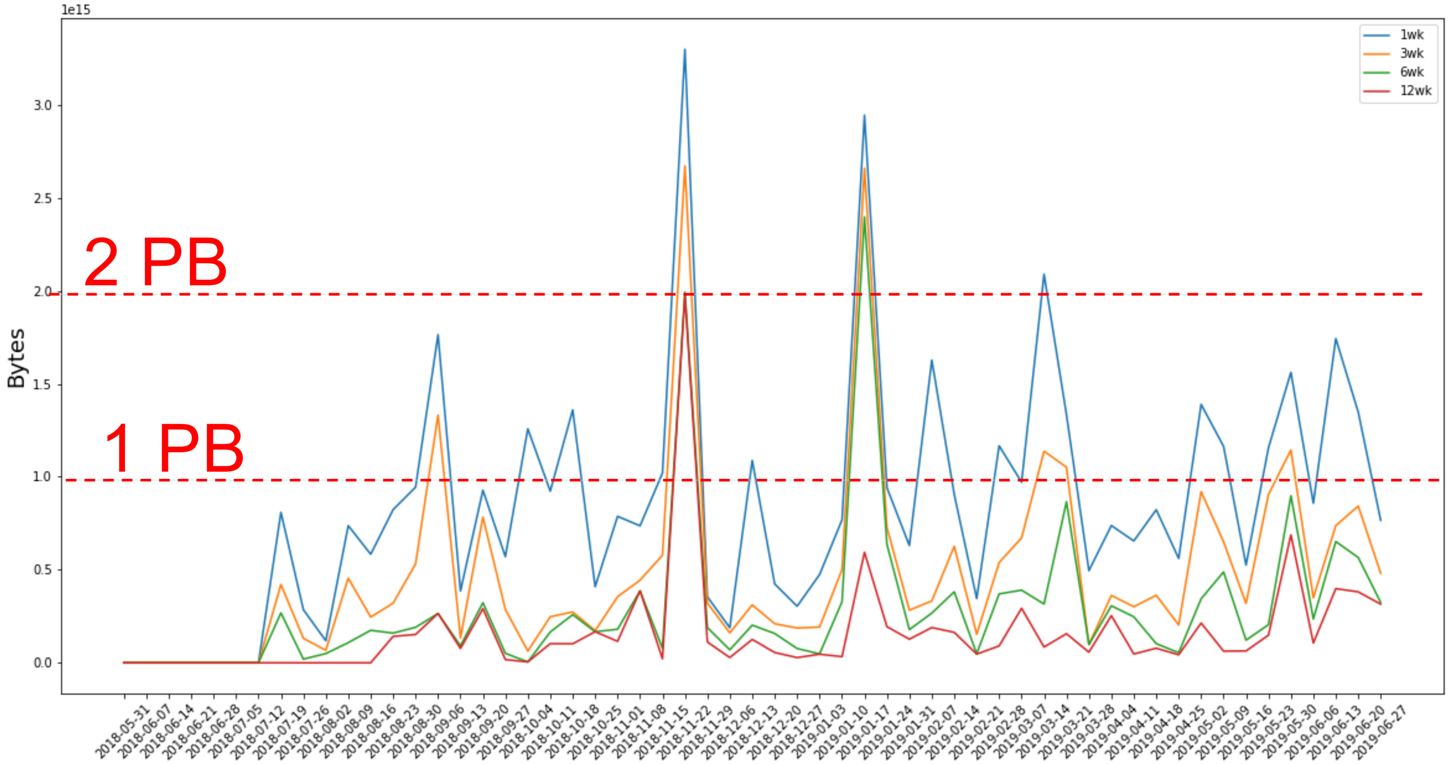
# CMS RAW from T0

- 6.5MB x 7.5kHz ~ 50GB/sec = 400 Gbit/sec
- 6.5M seconds/year data taking = 6.5/31.5 = 20% duty cycle over the year.
- You can pick your number based on how much backlog you are comfortable with.
  - I've picked 50% => 200Gbit/sec archival bandwidth to manage RAW data coming from CERN T0.

# CMS MINI production from AOD

- Assume it gets done in 100 days
- Coming off the archive: 240PB/100 days = 30GB/sec ~ 240Gbit/sec
- Going into archive: 30PB/100 days ~ 30Gbit/sec
- Rounded up generously ~ 300Gbit/sec

# CMS tape recall for analysis



Analyse use data for 1-12 week retention of unused AOD.
Plot recalled data/week as a function of time for each algorithm.

**For 3 week retention and a max of 1PB tape recall capacity per week, we expect processing delays of up to 3 weeks.**

# CMS Tape recall for analysis

- 1PB per week on Run2 => ~ 23PB/week HL-LHC to stay within the same processing delays.
  - 23 = ratio in AOD volume/year HL-LHC/Run2
- 23PB/week = 40GB/sec = 320Gbit/sec

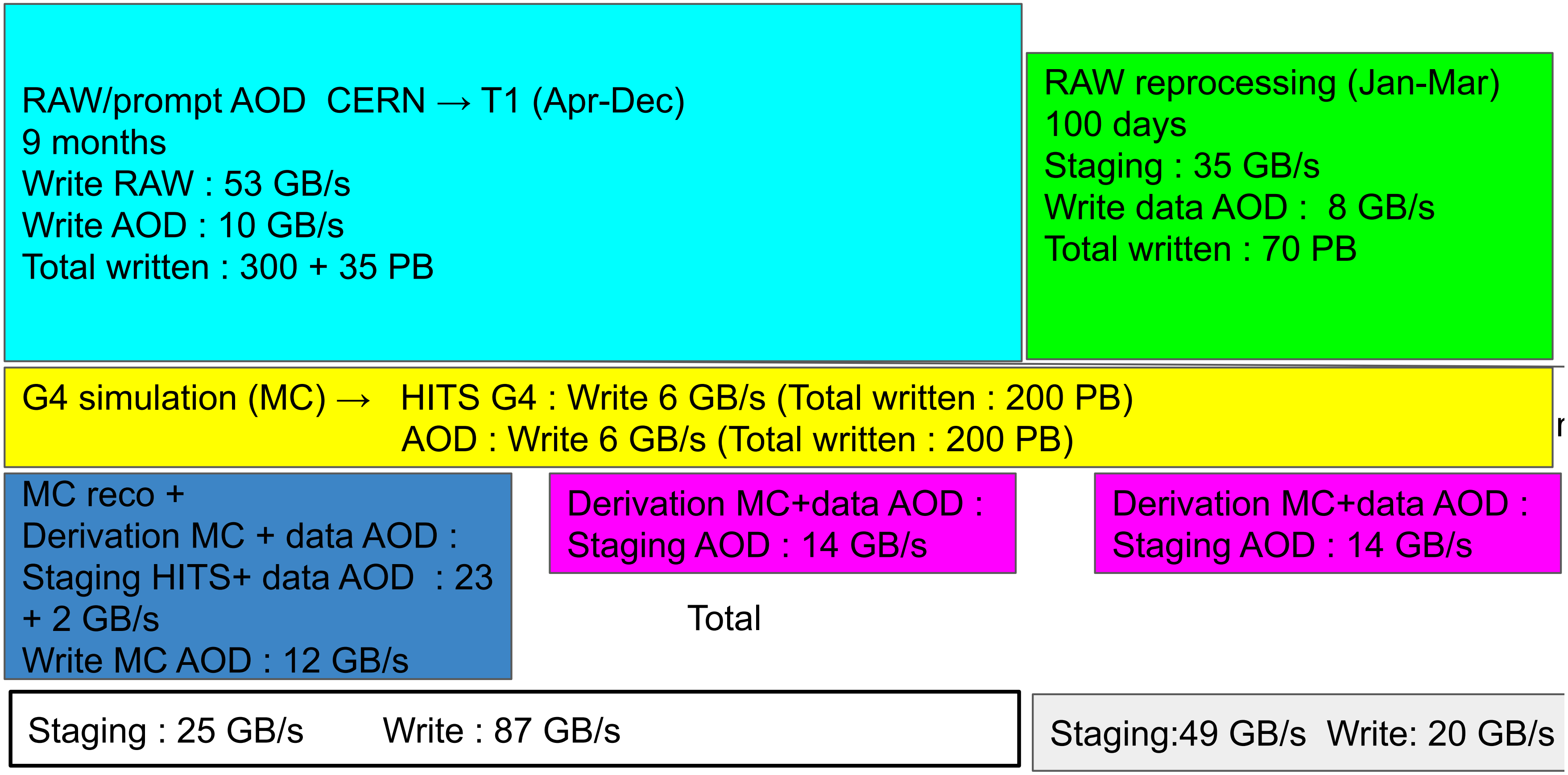
**Will use 250Gbit/sec as planning number.**
**Assumes that we do better in HL-LHC than Run2 with avoiding reliance on AOD**

# CMS Total

- RAW processing ~ 550Gbit/sec
- RAW from T0 ~ 200 Gbit/sec
- MINI production ~ 0
  - It falls in the shadow of the RAW processing
- AOD recall for Analysis ~ 250Gbit/sec
  - It better be small otherwise we are in trouble.
- **Total ~ 1000 Gbit/sec aggregate archival bandwidth needs across all CMS T1s.**

# ATLAS Scenario 1

## Scenario 1 : Maximise TAPE usage



RAW/prompt AOD CERN → T1 (Apr-Dec)
9 months
Write RAW : 53 GB/s
Write AOD : 10 GB/s
Total written : 300 + 35 PB

RAW reprocessing (Jan-Mar)
100 days
Staging : 35 GB/s
Write data AOD : 8 GB/s
Total written : 70 PB

G4 simulation (MC) → HITS G4 : Write 6 GB/s (Total written : 200 PB)
AOD : Write 6 GB/s (Total written : 200 PB)

r

MC reco +
Derivation MC + data AOD :
Staging HITS+ data AOD : 23 + 2 GB/s
Write MC AOD : 12 GB/s

Derivation MC+data AOD :
Staging AOD : 14 GB/s

Derivation MC+data AOD :
Staging AOD : 14 GB/s

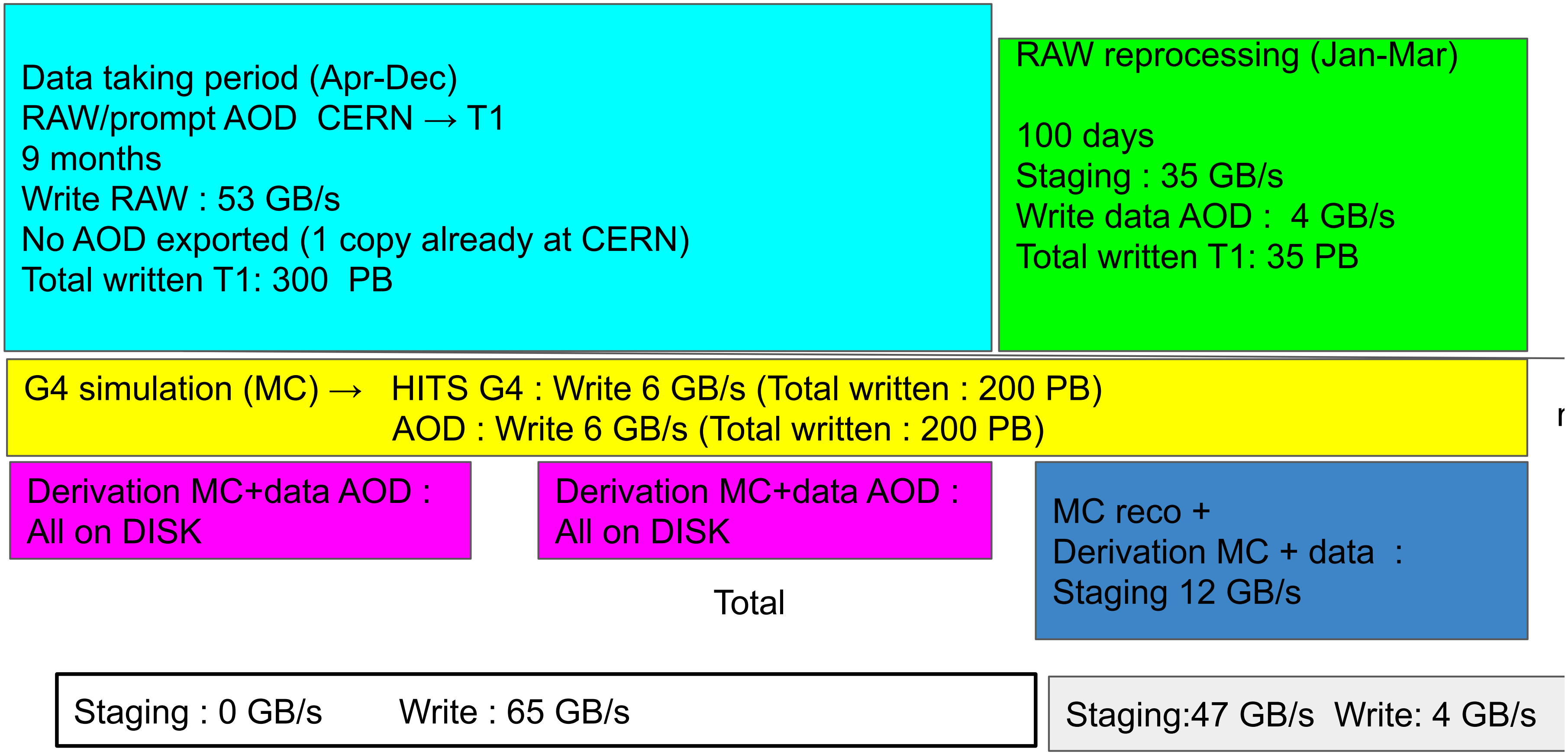
Total

Staging : 25 GB/s Write : 87 GB/s

Staging:49 GB/s Write: 20 GB/s

# ATLAS Scenario 2

## Scenario 2 : Minimise TAPE usage



Data taking period (Apr-Dec)
RAW/prompt AOD CERN → T1
9 months
Write RAW : 53 GB/s
No AOD exported (1 copy already at CERN)
Total written T1: 300 PB

RAW reprocessing (Jan-Mar)

100 days
Staging : 35 GB/s
Write data AOD : 4 GB/s
Total written T1: 35 PB

G4 simulation (MC) → HITS G4 : Write 6 GB/s (Total written : 200 PB)
AOD : Write 6 GB/s (Total written : 200 PB)

r

Derivation MC+data AOD :
All on DISK

Derivation MC+data AOD :
All on DISK

MC reco +
Derivation MC + data :
Staging 12 GB/s

Total

Staging : 0 GB/s Write : 65 GB/s

Staging:47 GB/s Write: 4 GB/s

# Some Comparisons

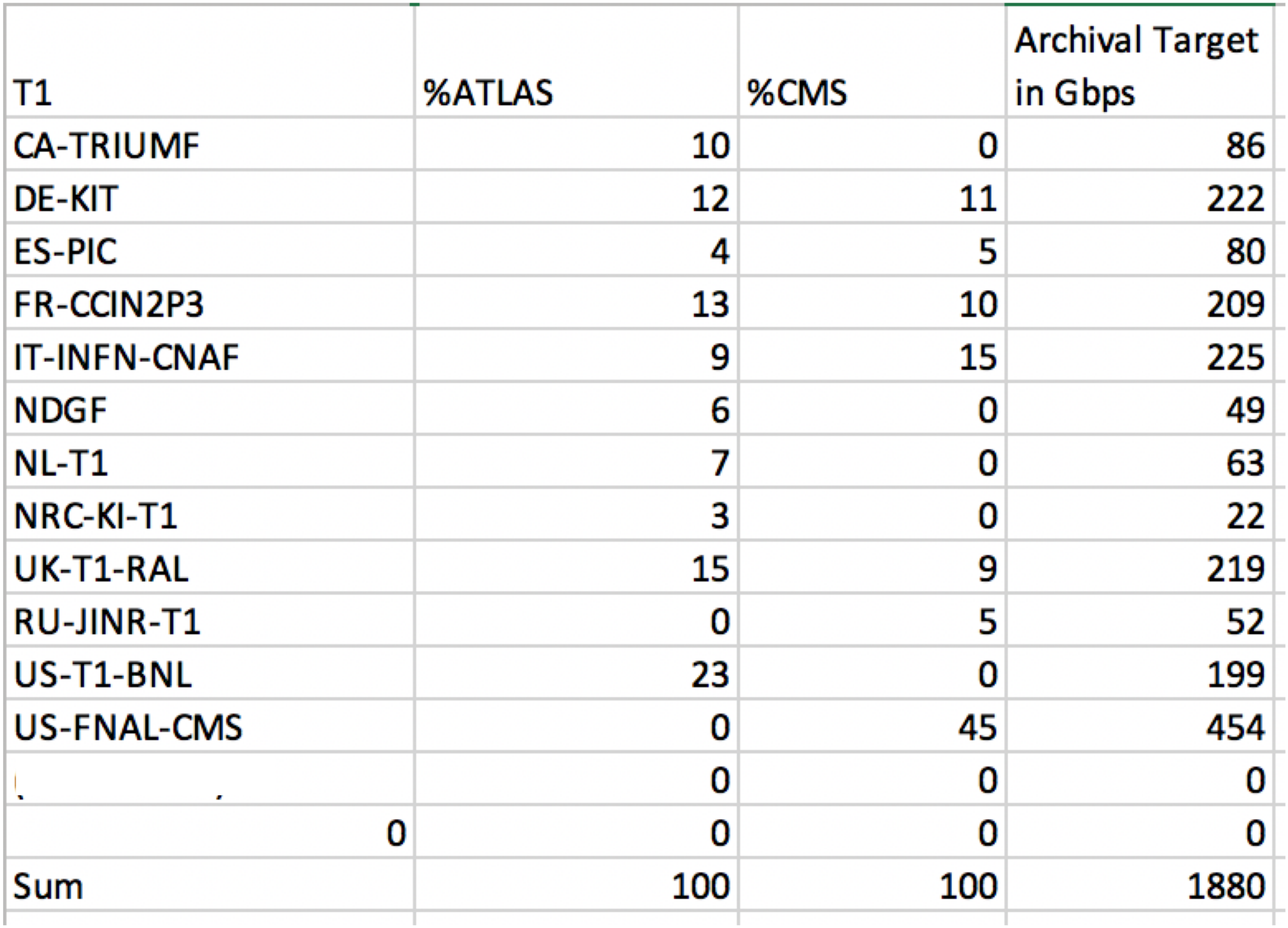
- ATLAS
  - Write RAW = 424Gbps
  - RAW Processing = 344Gbps
  - Higher data tier processing = 296 Gbps
- ATLAS sustained peak needs ~ 880 Gbps

- CMS
  - Write RAW = 200Gbps
  - RAW processing = 550Gbps
  - Higher data tier processing = 300 Gbps
- CMS sustained peak needs ~ 750-1000 Gbps

**No attempt made yet to understand and reconcile differences**

# Now lets fit it into a spreadsheet

# The % number in collums are rounded
# #s before rounding are used for target calculation



| T1 | %ATLAS | %CMS | Archival Target in Gbps |
|---|---|---|---|
| CA-TRIUMF | 10 | 0 | 86 |
| DE-KIT | 12 | 11 | 222 |
| ES-PIC | 4 | 5 | 80 |
| FR-CCIN2P3 | 13 | 10 | 209 |
| IT-INFN-CNAF | 9 | 15 | 225 |
| NDGF | 6 | 0 | 49 |
| NL-T1 | 7 | 0 | 63 |
| NRC-KI-T1 | 3 | 0 | 22 |
| UK-T1-RAL | 15 | 9 | 219 |
| RU-JINR-T1 | 0 | 5 | 52 |
| US-T1-BNL | 23 | 0 | 199 |
| US-FNAL-CMS | 0 | 45 | 454 |
|  | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 |
| Sum | 100 | 100 | 1880 |

# Comments & Questions

# Backup

# ATLAS Estimates (1)

## Inputs

- Assumptions :
  - RAW reprocessing do not overlap with data taking period (RAW export)
  - HI should not require more bandwidth
  - Very rough assumption for extrapolation in 9 years : Will evolve with Data Caroussel experience

- RAW + AOD/DAOD (prompt processing) export CERN → T1s (Table 8 of CDR)
  - RAW : 53 GB/s (during stable beam)
  - AOD/DAOD : 10 GB/s

- RAW staging at T0+T1s for reprocessing
  - T0 TAPE will be used but considered as safety margin
  - 300 PB in 100 days : 35 GB/s
- Write data AOD (35 PB) at T1s (output of RAW reprocessing)
  - Scenario 1 (higher TAPE load) :
    - 2 copies → 70 PB in 100 days : 8 GB/s
  - Scenario 2 (lower TAPE load) :
    - 1 copy → 35 PB in 100 days : 4 GB/s

# ATLAS Estimates (2)

Inputs : First simulation campaign

- HITS (after G4) produced over year
  - 1 copy on TAPE
  - (50 B evts fullsim + 150 B evts fastsim) * 1 MB/evt = 200 PB → 6 GB/s
- Write MC AOD (200 PB produced spread over 1 year) at T1s :
  - 100 % on TAPE : 6 GB/s

# ATLAS Estimates (3)

## Inputs : MC reco and derivation with existing input

- **Reprocess G4 HITS + derivation MC + derivation AOD**
  - Scenario 1 :
    - Process 100 % HITS in 100 days : Staging : 23 GB/s (No staging of MC AOD)
    - Write 50%/50% MC AOD on TAPE/DISK : Write 12 GB/s (MC).
    - Derivation 100 % data AOD in 100 days (35 PB with 50% on TAPE) : Staging 2 GB/s
  - Scenario 2 :
    - Process : 50 % HITS in 100 days during shutdown : 12 GB/s (No staging of MC AOD)
    - All data AOD processed from DISK copy
    - Write all on DISK

- Read data+MC AOD for derivation (No MC reco campaign) :
  - 3 repro of 100 days each year : ~Permanent derivation activity
  - Most often reprocessed (benchmark, important channel) on DISK : 50 %
  - Scenario 1 : 100% of AOD data+MC
    - 50% read from TAPE → staging 14 GB/s
  - Scenario 2 : 50% of AOD data+MC
    - All accessed from DISK