# Anticipated Analysis Workflows from the Facilities Perspective

Jim Cochran
Iowa State

Outline:

- Analysis Model First Year (AMFY) Status
- Expected Workflows
- Resources (including Strawman for US groups)
- How to manage group storage ?

# Analysis Model First Year (AMFY) Status

Report is due to be released in next few days (?)

Most recent presentations at NYU meeting in August & Barcelona in October (with rather different points of view)

Committee's mandate is very broad:

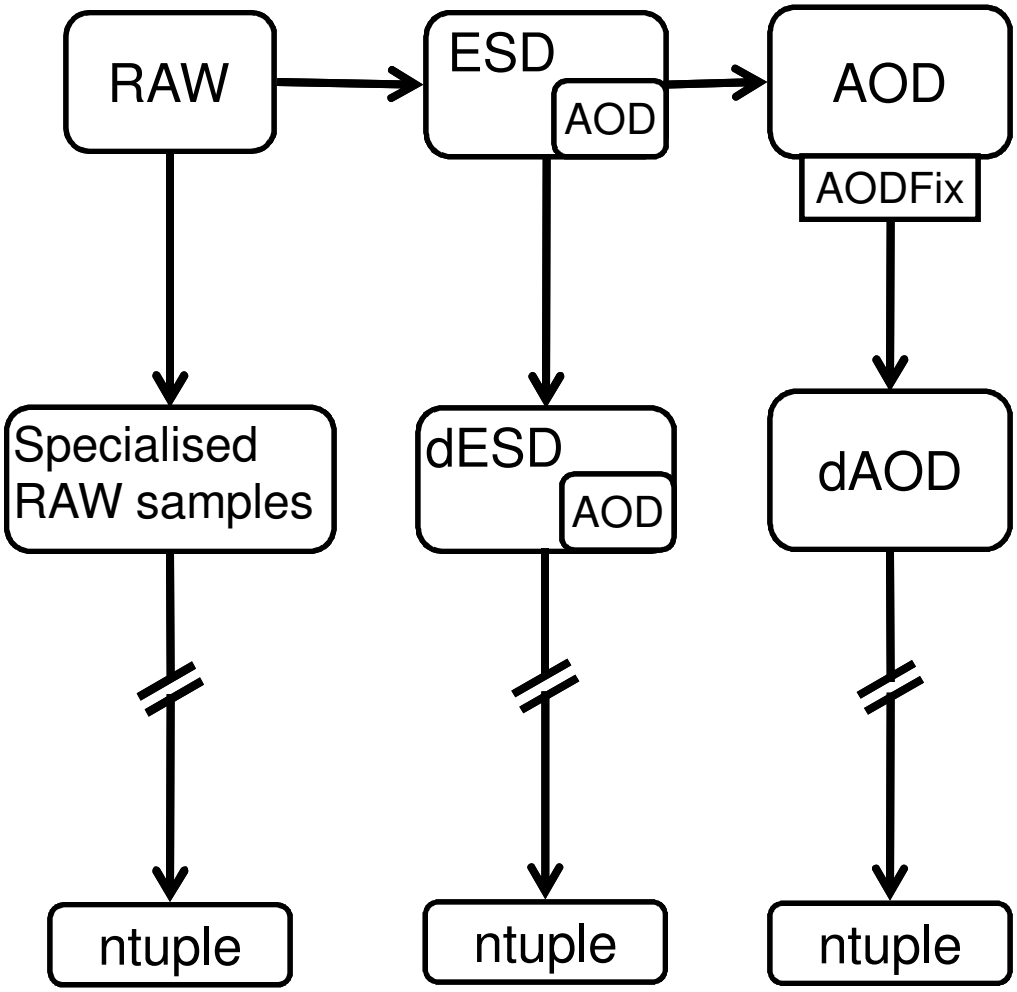
- define actual model
- develop recipes for commissioning, performance, physics, …
- identify needed data sets (including how much RAW, ESD, etc.)
- determine whether existing tools are sufficient
- what should be the role of tag data ?
- every step of analysis chain should be validated - how will this be done ?
- scrutinize our ability to perform distributed analysis
- match needs above with available resources

Considerable overlap with readiness walkthroughs

whew! no wonder it's taken so long to get this report out

# Analysis data formats

From T. Wengler's Talk at Barcelona




- Produced in ProdSys
- Main emphasis 2009/10: Detector studies → dESD
  (AOD branch can equally well be started from AOD in dESD)
- dAOD production by group manager

---

- Further derived formats/ ntuple production (PAT tools) by individual/group
  (Exceptions: Commissioning ntuples produced on Tier-0, and within ProdSys on Tier-1s)

Storage: RAW, ESD, AOD → ATLAS space, dAOD … group space / local

From T. Wengler's talk at Barcelona

# Analysis data formats

- Detector studies and commissioning
  - Online and offline monitoring
  - Locally saved data (ROD/ROS dumps) under detector control
  - About 10% of RAW and ESD available on CAF for regular checks (life time a few weeks)
  - Commissioning ntuples produced at Tier-0, or, preferably, at Tier-1 (mechanism now in place, already tested by L1Calo)

From T. Wengler's Talk at Barcelona

# Reprocessing

- The model for full reprocessing of the acquired data set in the first year foresees:
  - Full reprocessing starting from RAW exclusively (saving validation step for other variants such as ESD→ESD)
  - 2-3 reprocessings during 2009/10

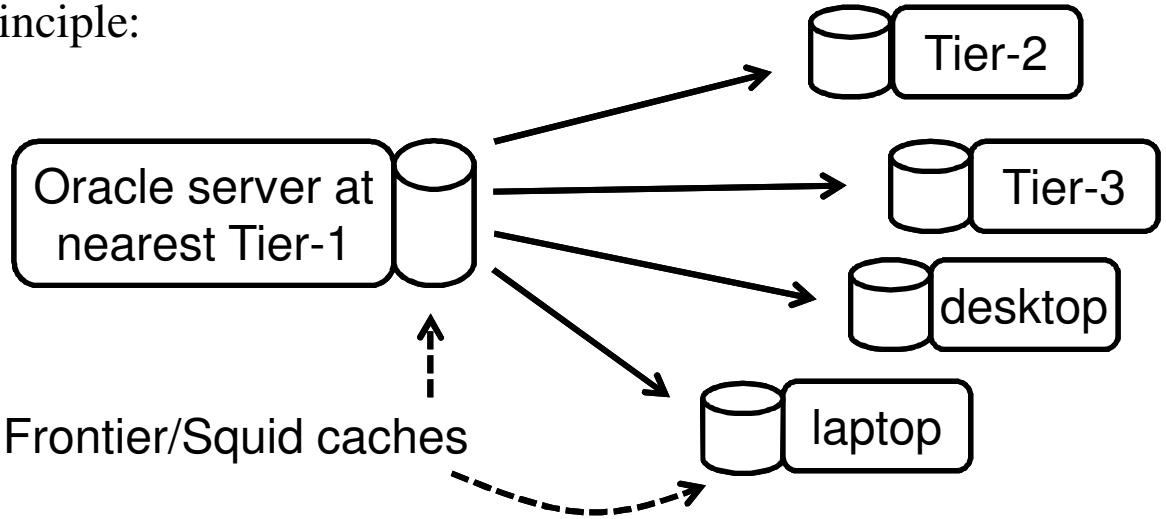
- Data used for public results are expected to have undergone at least one reprocessing
  - i.e. using the output of the Tier-0 reconstruction is normally not sufficient

- In between reprocessings, an AODFix mechanism may be used to apply well-defined corrections to the AOD data:
  - Corrections may be derived from RAW/ESD/AOD studies, but can only be applied based on AOD quantities, which defines possible applications. AODFix versions bound to a specific release/cache.

From T. Wengler's Talk at Barcelona

# Conditions/Meta-data access

- Every (ESD/dESD/AOD/DAOD) analysis presumed to need DB access → crucial to get this right
  - Main principle:




- Every analysis must be able to identify the version of whatever meta-data was used in producing a particular result
  - Centrally supported tools for luminosity calculation and ntuple dumping support saving such information in standard format

From T. Wengler's Talk at Barcelona

# Analysis software

- **Storage and versioning**
  - Code used for public/performance results has to be publicly available and versioned → can be clearly identified.
    - Analysis code to be run in the production system must become part of a release cache, for example for dAOD production
    - Analysis code not part of the release cache (root plot macros etc.) has to be stored and versioned in an equivalent way
  - The group areas used so far should be replicated by the above

- **Analysis tools**
  - If tool is felt to be missing → develop in the context of the relevant group (e.g. PAT, performance), and make it available
    - Private versions of centrally provided tools will lead to requests of replacing them with (or validating against) the centrally provided tools.
      (physics object base selection (medium electron…), luminosity/DQ/trigger tools, ntuple dumper, InSituPerformance …)

From T. Wengler's Talk at Barcelona

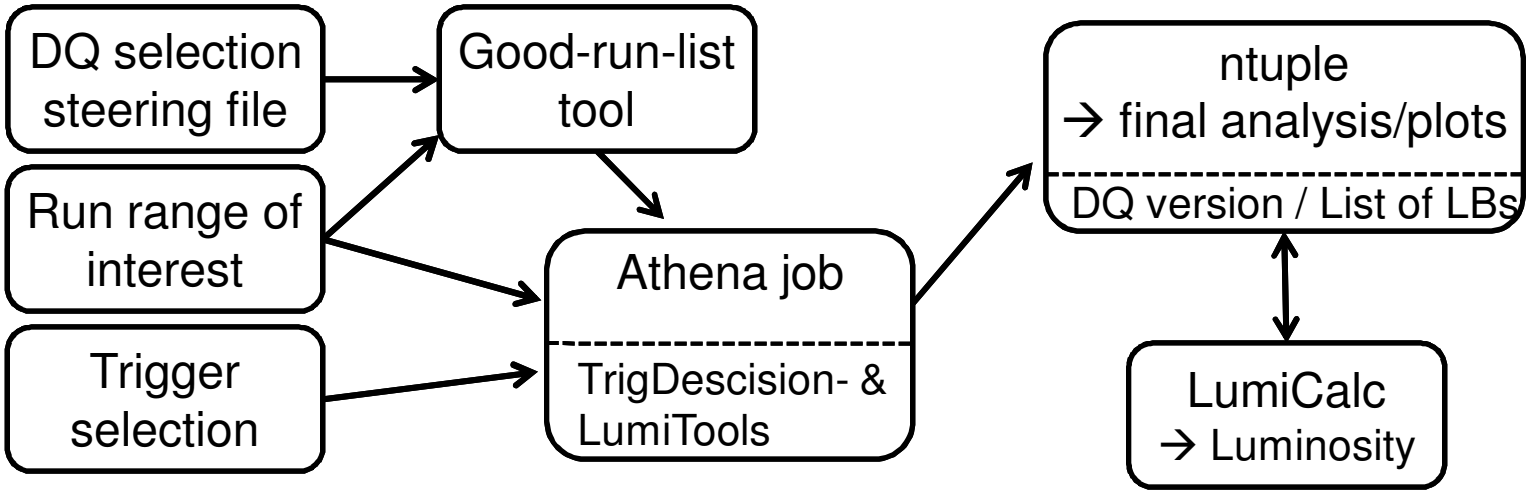
# Detector/Performance analysis

- Main activity during start up phase, carried out within appropriate group
  - Anyone starting such a study (i.e. prompted by something seen in physics analysis) – make the Perf/Det group aware and report
- Main data samples: dESDs
  - Analyse on Tier-2s, output of such analysis to group-space
  - Complemented by using CAF & local CPU on RAW, ESD, commissioning ntuples
  - Centralised tools, e.g. PAT ntuple dumper
  - Code needs to be accessible and versioned
- Performance results
  - To be signed off by CP/Det group and made available through the COOL database and tools like InSituPerformance
  - Should lead to an improvement of the detector monitoring and the MC description of the detector

From T. Wengler's Talk at Barcelona

# Physics analysis

- Main starting point are AOD objects
  - AOD sample, AOD contained in dESD, dAOD samples
- Typical analysis proceeds as:




→ also applicable to performance studies relying on DQ/Trigger/Lumi selection if needed

# Resources

Recall from my talk at UC meeting - US CPU looks to be sufficient for early running
(assuming we cooperate)
- US disk appeared problematic

Will now assume that US will meet our pledge for T1 & T2 disk
- if needed expect US can exceed pledge ?

Likely will not be able to keep all ESDs on T2s (at least not for long)
- may not be necessary to keep all (dESD should help a lot)
- need to be flexible and responsive to users

Michael has already discussed some data distribution plans from DDM group

From S. Campana's talk at Barcelona

# Space Managed by DDM (aka DQ2)

- DATADISK and MCDISK at all T1s and T2s
  - Store on disk (no backup) respectively detector data (including reprocessing) and simulated data
- DATATAPE and MCTAPE at all T1s
  - TAPE archive for detector data (including reprocessing) and simulated data
  - Not accessible by users
- SCRATCHDISK at all T1s and T2s (and possibly at T3s)
  - Volatile disk space for user data as well as centrally managed activities
  - Cleaned after 30 days
- GROUPDISK at some T1s and T2s
  - Disk space for group activities
  - Quota and ACLs defined at the group level
- LOCALGROUPDISK at many T3s
  - Permanent NON PLEDGED disk space for user data
  - Permissions and quota defined by the T3 (institute/lab/center)

From S. Campana's talk at Barcelona

# Space for Groups

- Technical limitations (from DDM devel and ops):
  - No more than 4 groups per site
  - No more than 25 sites with group space
- Any other constraint/limitation/discussion is NON TECHNICAL
  - Organizational
  - Political
  - …
- We started drafting a first implementation of groups per site
  - Based on first (surely not conclusive) discussion with cloud representatives
  - Based on several meetings with group managers and representatives

# Strawman Group Space Allocation

(not including detector space)

very preliminary

| **PERF** | muon | egamma | jets | flavtag | idtracking | tau | TRIG |
|---|---|---|---|---|---|---|---|
| DE | | FZK | DESY | | DESY | | |
| FR | Lyon, muon | Lyon | Tokyo | Lyon | | | Tokyo |
| IT | CNAF, Roma, Napoli | CNAF, Milano | | | | | |
| US | NET2, AGLT2 | SWT2 | MWT2, WT2 | WT2 | WT2 | MWT2 | AGLT2 |
| NL | NIKHEF | | | | NIKHEF | | |
| NDGF | | | | | | | GE |
| TW | | | | | | | |
| ES | | | PIC | | | | |
| UK | | | | | | | |
| CA | EAST | SFU | SFU | | | | |

| **PHYS** | SM | Top | SUSY | Exotic | Higgs | HI | Beauty |
|---|---|---|---|---|---|---|---|
| DE | | DESY, Prague | DESY, LMU | | | CIFRONET | UIBK |
| FR | GRIF, LAPP | GRIF, LPC | Tokyo | Tokyo | Lyon, Tokyo | | |
| IT | Milano | Milano | Napoli | | Roma | | |
| US | AGLT2, WT2 | NET2, SWT2 | SWT2 | NET2 | MWT2 | | WT2 |
| NL | NIKHEF | Nikkhef | Nikkhef | IHEP | NIKHEF | | |
| NDGF | Ljubljana | Ljubljana | T1 | | T1 | | T1 |
| TW | AUS | | | | | | |
| ES | | IFAE | IFIC | UAM | | | |
| UK | GLA | QMUL | RALPP | Oxford | Liv | | LANCS |
| CA | | | | | | | |

# The US T2s (including detector req)

very preliminary

| | |
|---|---|
| NET2 | muon, Top, Exotic, MuonDet |
| SWT2 | egamma, Top, Susy, TileDet |
| AGLT2 | muon, trig, SM, MuonDet |
| MWT2 | Jets, Tau, Higgs, TilDet, LArgDet |
| WT2 | FlavTag, IDTracking, SM, Beauty, InDet |

# Data Distribution and Availability

From S. Campana's talk at Barcelona

- Detector Data from the T0
  - RAW data replicated to one T1 (DATATAPE)
  - ESD replicated to 2 T1s (DATADISK) + BNL (DATADISK)
    - One copy follows the RAW parent
  - AODs replicated to all T1s and many T2s
    - Approx 20 copies world-wide at the moment
    - At least 1 copy per cloud, shared across T2s
- Reprocessed Detector Data at T1s
  - ESD at custodial T1+ replicated to BNL (DATADISK)
  - AOD replicated to all T1s and many T2s (also for dESDs)
    - Same policy as Detector Data
- Simulated Data
  - ESD replicated to BNL (DATADISK)

# Decreasing the number of copies

From S. Campana's talk at Barcelona

- **GUARANTEED** number of copies will decrease after new reprocessings
  - In the new computing model:
    - 12 copies for version N (N=current)
    - 11 copies for version N-1
    - 0 copies for version N-2

- Copies in excess (beyond guaranteed) will be removed when space is needed, based on:
  - Age of the dataset
  - Number of accesses to the dataset
  - Inputs from physics coordination (+ TOB)

- Number of accesses:
  - Determined counting read operations via OFFICIAL TOOLS
    - dq2-get, Ganga, Pathena
    - Anything else (local access via rfcp, dccp, cp …) is not accounted
  - Please use official tools, otherwise your activity will not influence the number of copies

From S. Campana's talk at Barcelona

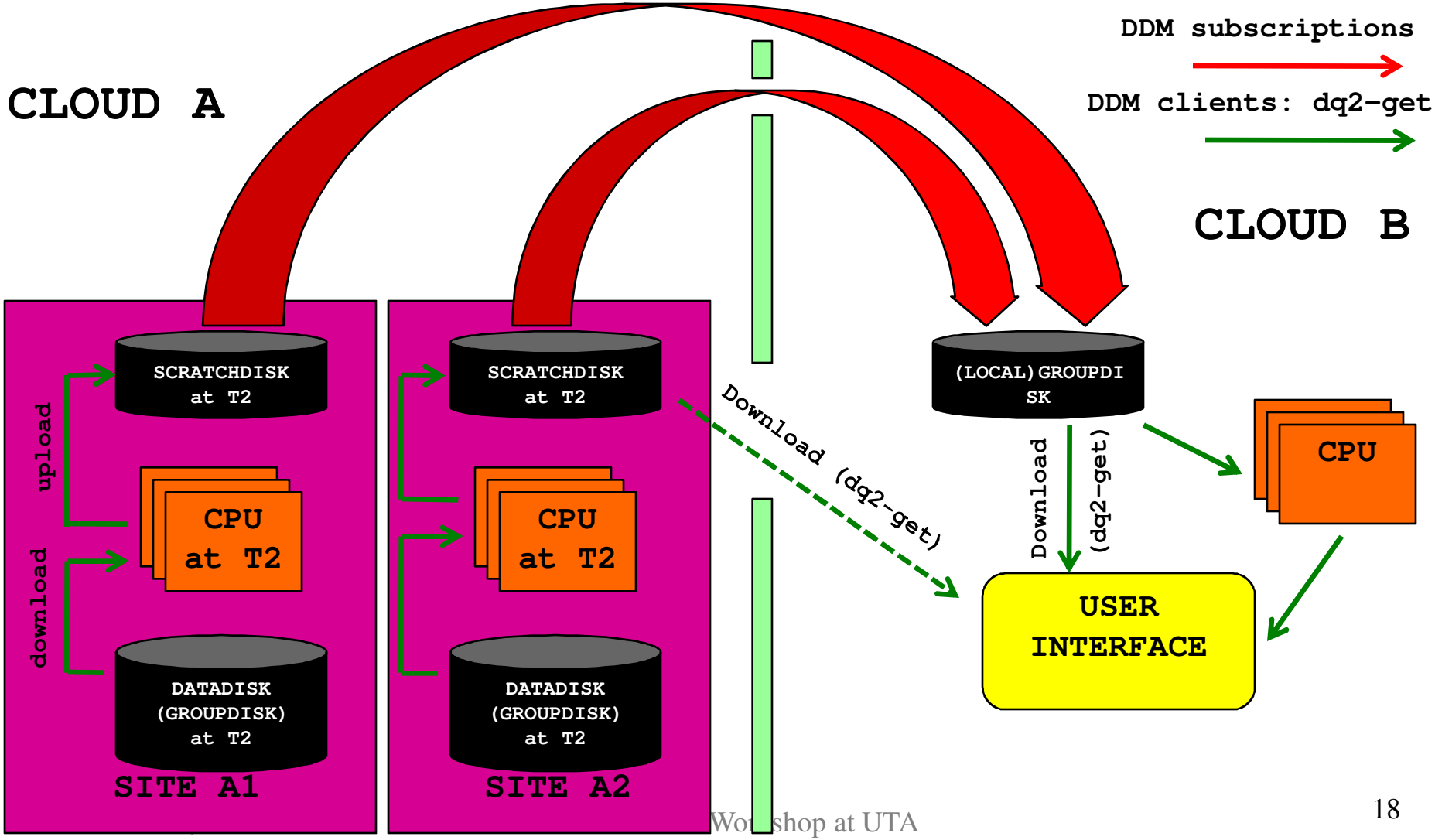
# Analysis on the Grid: main principles

- Jobs are sent to data (and not viceversa)
  - Data are preplaced at Grid sites
  - Analysis tools (Ganga, pAthena) figure out the best place to run a job, given the data availability
- Output are always stored in volatile spaces (SCRATCHDISK)
  - Will be automatically cleaned after N days (N=30 currently)
- Small samples can be downloaded from SCRATCHDISK via dq2-get
  - For debug purposes, do not abuse
- Users/Groups can request output data replication in permanent storage
  - http://panda.cern.ch:25980/server/pandamon/query?mode=ddm_req
  - Based on an approval mechanism

From S. Campana's talk at Barcelona

# Data Access: analysis on the Grid

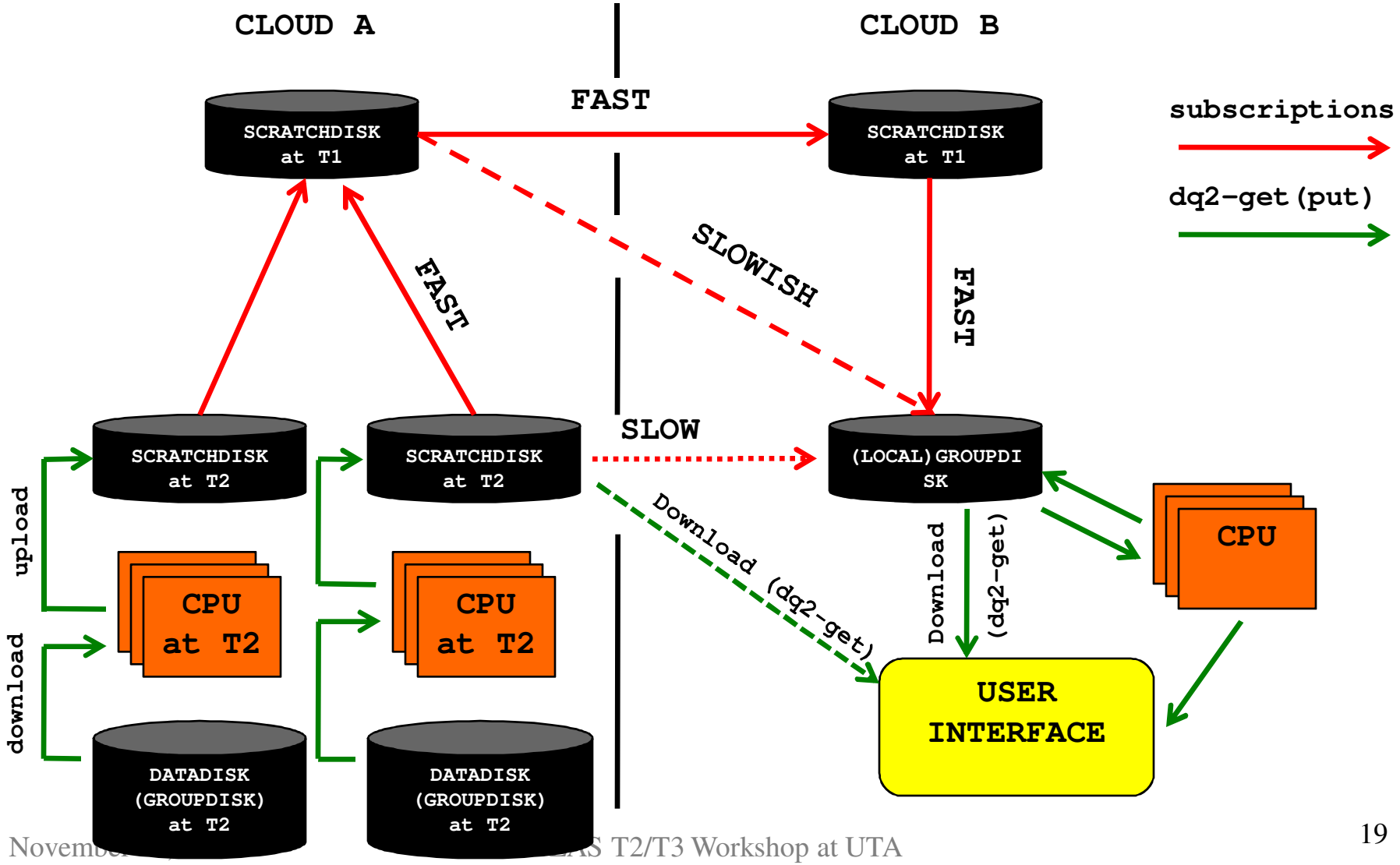
T2s in Cloud A to T3(?) in Cloud B (direct)



From S. Campana's talk at Barcelona

# Data Access: analysis on the Grid

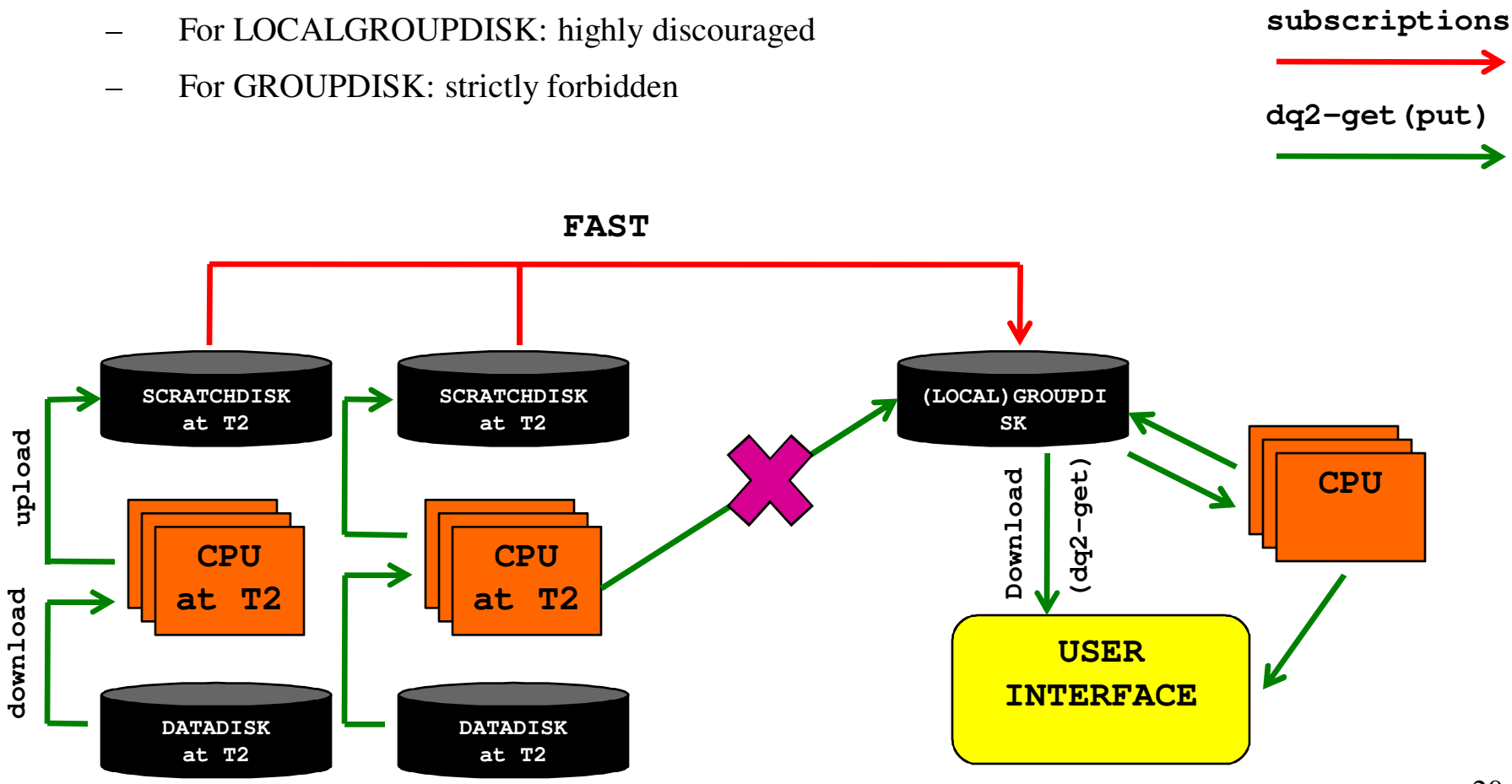
## T2s in Cloud A to T3(?) in Cloud B (via T1s)



# Data Access: analysis on the Grid

From S. Campana's talk at Barcelona

- Data movement is simplified if everything happens in the same cloud
  - No need to multi-hop
- Direct upload from CPUs to final destination
  - For LOCALGROUPDISK: highly discouraged
  - For GROUPDISK: strictly forbidden



# Expected analysis patterns for early data

Same slide shown at UC meeting

Assume bulk of group/user activity will happen on T2s/T3s
(define user accessible area of T1 as a T3 [BAF/WAF])

Assume final stage of analysis (plots) happens on T3s (T2s are not interactive)
[except for WAF]

Two primary modes:

(1) Physics group/user runs jobs on T2s to make tailored dataset (usually $D^3PD$)
(potential inputs: ESD,AOD,$D^1PD$)
resultant dataset is then transferred to user's T3 for further analysis

(2) group/user copies input files to specified T3 (potential inputs: ESD,AOD,$D^1PD$)
On T3 group/user either generates reduced dataset for further analysis or performs final analysis on input data set

Choice depends strongly on capabilities of T3, size of input data sets, etc.

Also, expect some users to run $D^3PD$ analysis jobs directly on T2 analysis queues

# How to manage US GROUP space ?

RAC in conjunction with forum conveners ?

# Backup Slides

# Storage plans (as I understand them)

## Decided

| | |
|---|---|
| Included in LCG pledge: | T1: All AOD, 20% ESD, 25% RAW |
| | each T2: 20% AOD (and/or 20% $D^1$PD ?) |

2 copies of AODs/$D^1$PDs (data+MC) are distributed over US T2s

1 copy of ESD (data only) distributed over US T2s (expect only for 2009-2010)
(may be able to use perfDPDs in some cases)

$D^1$PDs initially produced from AODs as part of T0 production, replicated to T1s, T2s
$D^1$PDs will be remade from AODs as necessary on the T1

# Recommendation on dataset “distributions”
## (my personal suggestion)

# Recall

Included in LCG pledge: T1: All AOD, 20% ESD, 25% RAW
each T2: 20% AOD (and/or 20% D$^1$PD ?)

2 copies of AODs/D$^1$PDs (data+MC) are distributed over US T2s

1 copy of ESD (data only) distributed over US T2s (expect only for 2009-2010)
(may be able to use perfDPDs in some cases)

---

For 1st 2 months:

AOD/D$^1$PD storage should not be a problem
(assume distribution is handled automatically ?)

For ESDs and pDPDs, should have **all** streams available on T2s

who decides these ?
when ?

Should we consider associating specific streams to specific T2s ?

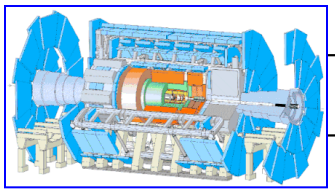
From Adam Gibson's NYU talk

# Attempt to Define Current Model Fails

- Tried to define the model as we saw it then
  - Reach a common foundation for future discussion
- Flowcharts in several views:
  - Format flow
  - Conditions flow
  - Reprocessing cycle
  - Tools
  - As from an analyst for study x,y, z, …
- Attempt to define current model didn't really work
  - Too many pieces currently in flux.
  - A close look at ay piece **requires scrutiny of computing resources, production system, MC strategy**, etc. (at least enough to understand the issues involved)
- Gave up on goal of defining "current" model
  - **Aim for a description, at the end of task force, of envisioned model for first year**
- Probe assumptions, practicalities, and check it with as many people as possible
  - Proposed solutions should not be seen as a definite recommendation [unless we explicitly label it as such], but rather as a basis for discussion

# Format Flow

From Adam Gibson's NYU talk



FE — ROD — ROS — … — SFO

from here everything is stream-wise

RAW data samples

sub-system resources

Tier-0

- Trigger/MinBias (from ESD)
- Muon Comm. (from RAW)
- Tracking Comm. (from RAW)

Comm. NT

RAW

ESD

AOD

Perf DPD x9

AOD

~10%

~10%

CAF

merge Commissioning and Performance DPDs

TAG

on disk

on disk

D[1]PD

on disk as long as possible

on disk

Tier-1

one copy of RAW across all Tier-1s

two copies of ESD across all Tier-1s

Tier-2 x N

One copy of all active AOD/PerfDPD/D1PD sets across all Tier-2s of one cloud

ESD/RAW data sets retrieved on request

user/group data sets (i.e. D2PDs), MC produced on Tier-2, copied to Tier-1s