The Role of the Tiers for CMS Analysis

Thomas Kreß,
RWTH Aachen IIIb
HGF Terascale Alliance - Grid Project Review Meeting
Munich, Feb 10th, 2009

Overview of Tier Levels at CMS

- **Tier-0** - ...

- **CERN CAF** - low latency (w/o Grid overhead) calibration & alignment analysis by priveledged users

- **Tier-1** - only central & scheduled tasks like re-reprocessing; generally no chaotic/individual user analysis (to protect tape systems, ...) allowed

- **Tier-2** Monte Carlo production, resources for detector and physics groups and for national/local community and individual users; interactive access / login not mandatory

- **Tier-3** - as technical term not rigorously defined: [no ... full] Grid functionality Under full national/local control
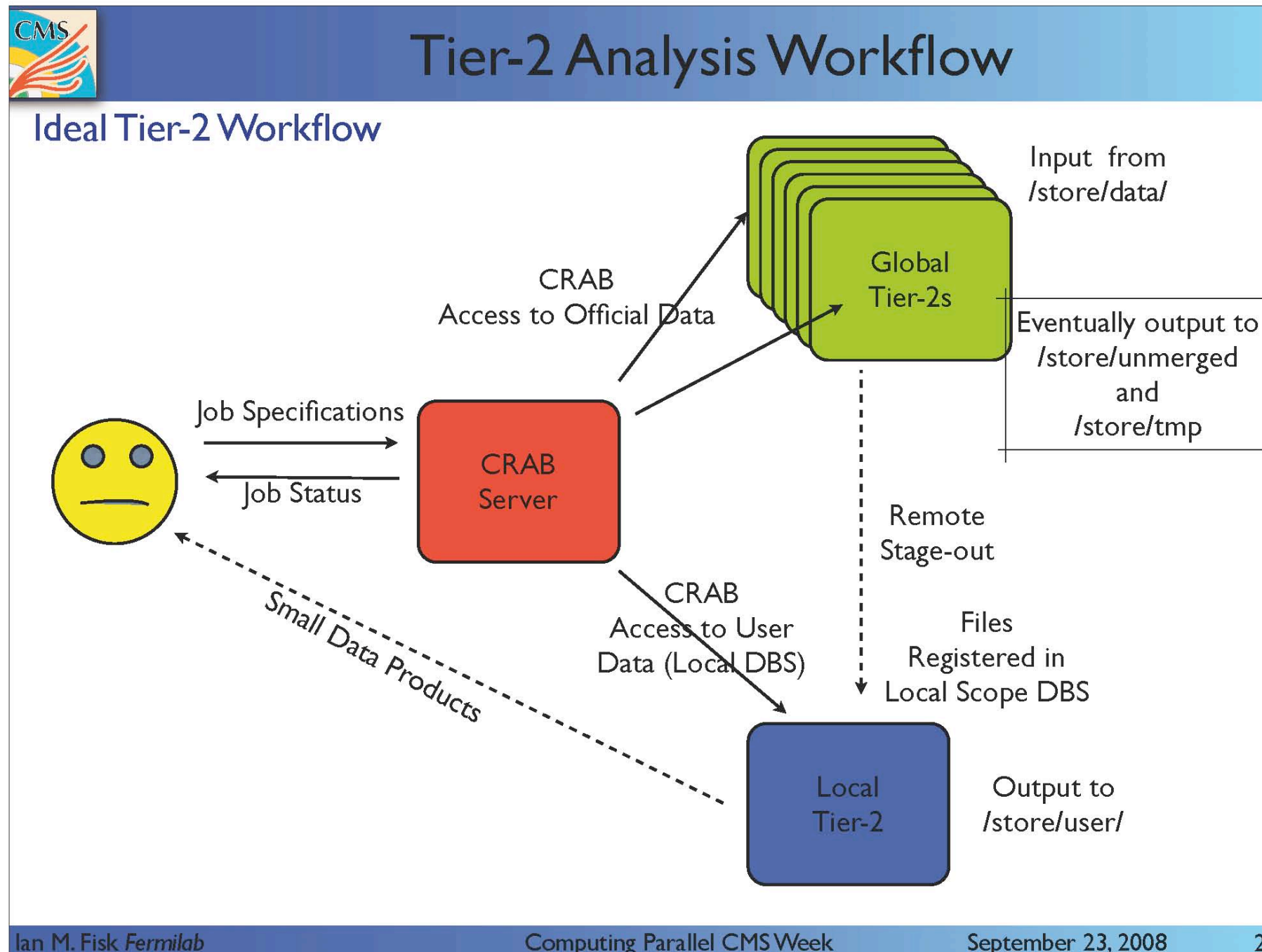
Tier-2 Storage Setup 2009

- Transient & unmanaged to more persistent & centrally managed

    - in 2009 in total O(200 TB) for a nominal Tier-2

    - „home" users' Grid stage-out storage space (O(1/2-1 TB) + private resources / user)

    - O(30-60 TB) for local/national community

    - n * 30 TB detector & physics group space (n=1 ... 4 / T2), n -> MoA credits

    - 30 TB central (RECO/AOD/PAT) data sets (primary skims, bg MCs, ...)

    - O(20 TB) service space (MC prod. buffer, ...)
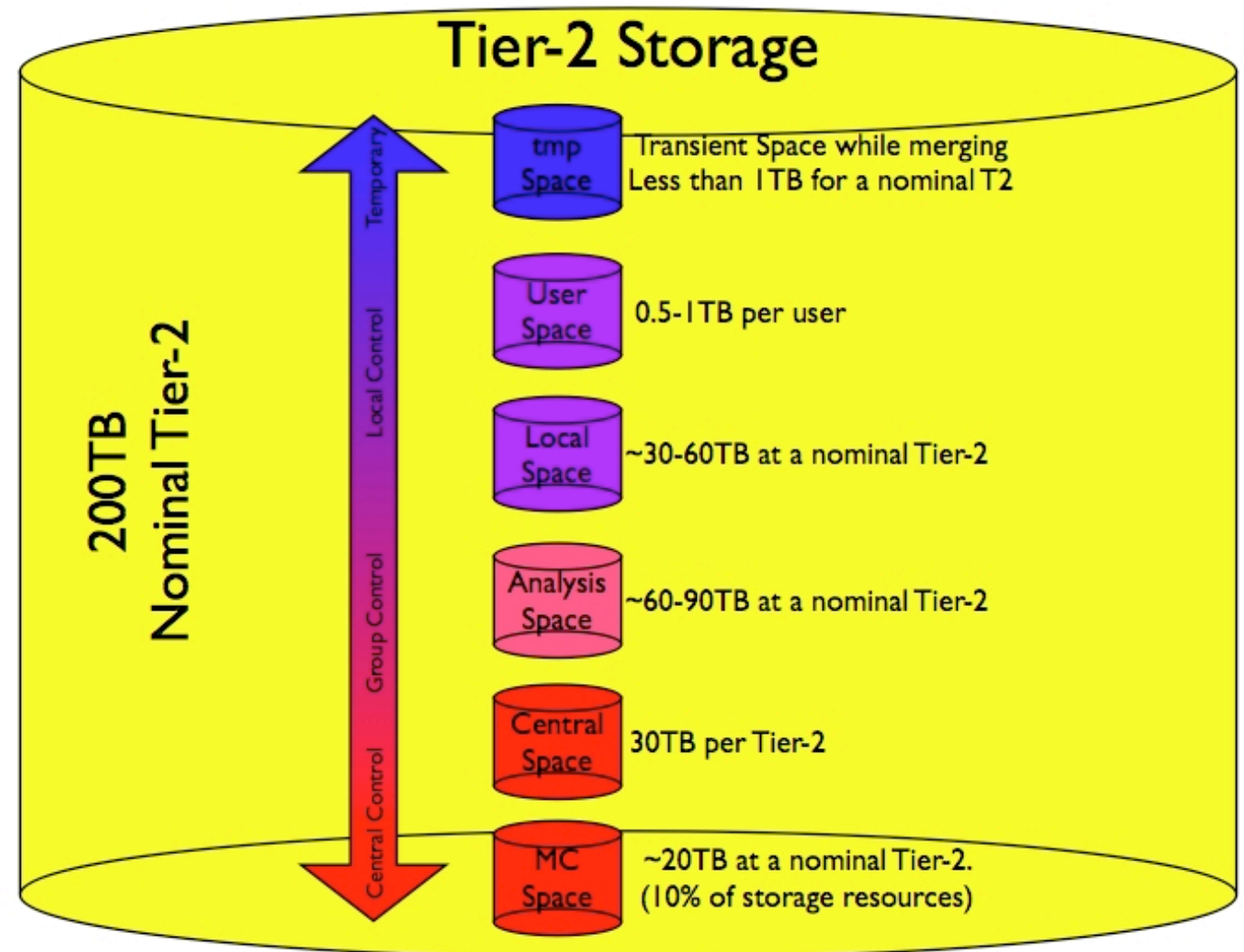
Tier-2 CPU / Batch Setup

- So far only modest usage of VOMS groups / roles

  - high priority for monitoring (SAM, ...) and central CMS tasks (software installation, MC prod., ...)

  - some countries (US, Germany (DCMS) , ...) give increased batch fair share to their national users to account for additional national/local funding

  - in DCMS so far no VOMS groups on location/institute/lab. level

  - not yet higher fair shares for associated D & P group's users compared to „normal" CMS users

- Every CMS user can transparently run his Grid analysis at every CMS Tier-2 site - „where the (central/group/local/personal) data is" - but stages back the output to his home Tier-2 sites

- For users so far (almost) no pilots jobs, instead push mode(s) used

User Analysis Concept

Tier-2 Analysis Workflow

Ideal Tier-2 Workflow

Input from /store/data/

Global Tier-2s

CRAB Access to Official Data

Eventually output to /store/unmerged and /store/tmp

Job Specifications

CRAB Server

Job Status

Remote Stage-out

Small Data Products

CRAB Access to User Data (Local DBS)

Files Registered in Local Scope DBS

Local Tier-2

Output to /store/user/

Data Placement at Tier-2 Sites

- By CMS tool Phedex, based on FTS & SRM

- Operation teams replicate / delete the „central" data to / from the T2s

- D & P groups nominate data manager(s) to request data replication to group's T2 storage space(s), or group's represen-tative(s) run e.g. Grid analysis sub-skims and stores output at group's T2s

- In principle every CMS user can indivi-dually request data transfers to T2s easily

- Always the local Tier-2 data managers accept or reject the requests

New Transfer Request

E-mail: `Thomas.Kress@cern.ch`

DBS: `https://cmsdbsprod.cern.ch:8443/cms_dbs_prod_global_writer/servlet/DBSServlet`

Data Items: `/a_special_sample_of_interest_for_susy/bla.root`

/Primary/Processed/Tier
or
/Primary/Processed
/Tier#Block
(Use * as wildcard)

More Help

Destinations:

☐ T1_IT_CNAF_MSS	☐ T2_CN_Beijing	☐ T3_GR_IASA_GR
☐ T1_TW_ASGC_MSS	☐ T2_DE_DESY	☐ T3_GR_IASA_HG
☐ T1_UK_RAL_MSS	☑ T2_DE_RWTH	☐ T3_GR_Ioannina
☐ T1_US_FNAL_MSS	☐ T2_EE_Estonia	☐ T3_IT_Napoli
	☐ T2_ES_CIEMAT	☐ T3_IT_Perugia
	☐ T2_ES_IFCA	☐ T3_TW_NCU

Transfer Type: `Replica` ▾ What's this?
Subscription Type: `Static` ▾ What's this?
Priority: `Low` ▾ What's this?
Custodial: `Non-custodial` ▾ What's this?
Group: `susy` ▾ What's this?
Comment: `I have the permission of the SUSY group to transfer this data set to RWTH Aachen into the SUSY group space.`
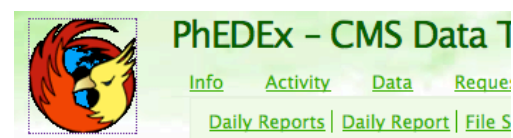
Status of Implementation

- In Summer 2008 detector & physics groups were associated to Tier-2 sites

  - was much more a „political" than a technical challenge

- From fall, data were distributed to Tier-2 central & group space, then blocked at T1s

  - we now have a nice data accounting on DB level for groups and Tier sites

Country M

	T2_AT	T2_BE	T2_BR	T2_DE	T2_CH	T2_CN	T2_EE	T2_ES
FWD phys				1				
QCD				1				
Higgs								1
EWK								1
SUSY	1				1			1
Top			1		1			1
Exotica								
B Physics					1	1		
Heavy Ions								
egamma								
Jets/MissET					1			
Muons								1
B-Tagging	1			1				
Tracker					1			
Tau / Pflow							1	
Trigger DPG								1
Reserve								
Unallocated	?							
Current Resources	0	1	1	3	0	0	1	5
Fall Resources (*)	2	1	1	6	1	1	1	5

PhEDEx – CMS Data T

Info   Activity   Data   Reque

Daily Reports | Daily Report | File S

View node: T2_US_Nebraska ▼ Submit

T2_US_Nebraska Group Usage

Group	Subscribed	Resident
DataOps	3.97 TB	3.69 TB
FacOps	1.04 TB	1.04 TB
b-tagging	4.59 TB	4.58 TB
top	2.71 TB	2.71 TB
tracker	1.40 TB	1.40 TB
undefined	71.22 TB	71.04 TB
	84.94 TB	84.46 TB

T2_US_Nebraska Custodial Data

Custodial	Subscribed	Resident
Non-Custodial	84.94 TB	84.46 TB
	84.94 TB	84.46 TB

T2_US_Nebraska Non-subscribed Data

Source	Non-subscribed
3.82 TB	0.00 B

Group Usage

DataOps

Node	Subscribed	Resident
T2_DE_RWTH	867.50 GB	867.50 GB
T2_US_Nebraska	3.97 TB	3.69 TB
	4.82 TB	4.54 TB

FacOps

Node	Subscribed	Resident
T2_US_Nebraska	1.04 TB	1.04 TB
	1.04 TB	1.04 TB

b-tagging

Node	Subscribed	Resident
T2_IT_Pisa	6.25 TB	6.25 TB
T2_US_Nebraska	4.59 TB	4.58 TB
	10.84 TB	10.82 TB

ewk

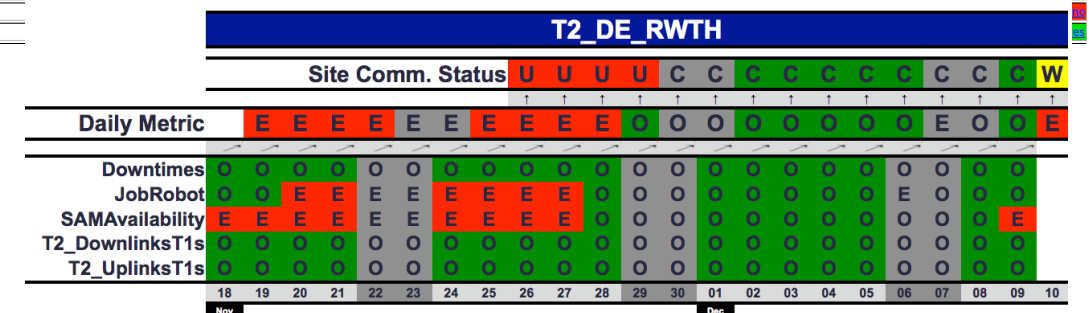| Node	Subscribed	Resident
T2_IT_Pisa	7.85 TB	7.79 TB
	7.85 TB	7.79 TB

tau/pflow

Status of Implementation ... con't

- We see analysis job activity at T2s !

- The deployment of „user home space" is just under way

  - user has at least one home area with guaranteed space from CMS T2 pledges

  - usually at his local/national T2 (special arrange-ments for CERN users and countries w/o Tier-2)

  - beside a T2, a local T3 with sufficient CMS Grid capability is OK, but only best effort support

- Tier-2 sites' reliability is improving

jobs per user

At RWTH Aachen, period 24 h

submitted · app-succeeded · application-failed · site-failed · aborted · cancelled · app-unknown · pending · running

Last Steps in User's Analysis

- Usually user's Grid output files are in .root format

- Can be registered automatically in „local-scope" database(s)

- Output can be read (by every CMS user) by further Grid jobs (if in DB), from CMS software framework, a „light" version, or directly from root package

- Output on Grid Tier-2 space can sometimes be accessed directly from local desktops (e.g. in Aachen read-only by dCap protocol), from the NAF, or can be copied over to local disks with CMS web tool or Grid tools from an UI

Concerns

- Sufficient data protection (ACLs, Quota, ...) is still missing

- WAN-SRM transactions frequency is very low (O(few Hz))

  - alternative: local (protocol) stage-out, harvesting and WAN transfer of merged output ?

- CE, WMS, ... scalability ?

  - alternative: pull mode / pilot jobs ?