CMS Tier-2 Program for user Analysis Computing on the Open Science Grid

Frank Würthwein
UCSD

Goals & Status

High Level Requirements for user analysis computing

- Code Development Environment
  - Compile, run, debug with fast turn around
  - Very agile & reasonably interactive
  - Complete data access at modest IO
- Large scale processing environment
  - Large scale parallelization
  - Large CPU & IO
  - Perfect bookkeeping that's trivial to use
  - Latencies commensurate with resource consumption

CMS Global Data Grid

- 2500 physicists, 40 countries
- 10s of Petabytes/yr by 2008
- 1000 Petabytes in < 10 yrs?

Boundary Conditions for CMS user analysis computing

- CMS Computing Model:
  - Organized skims at Tier-1.
  - **Physics group activities at Tier-2.**
  - **Free-wheeling user analysis at Tier-2.**
- LHC and the Open Science Grid (OSG)
  - US CMS resources available via OSG

The Role of Tier-2

- Home of **physics group activities.**
- Home of **free-wheeling user analysis**.
- Enable **distributed control & ownership**
  - 4 types of user @ T2
    - Local CMS owners = member of T2 institute/region
    - Local CMS users = member of physics group@T2
    - Opportunistic CMS users
    - Random OSG user not in CMS

Resource utilization policy shall not be limited by technology!

OSG Philosophy

- Provide a core production infrastructure
  - Set of operational procedures
    - Incidence response, AUP, governance, operations, etc.
  - Limited set of core services
- ***Enable Communities to "roll their own".***
  - All high level services CMS depends on are fully controlled by CMS.
- Establish coherence across communities
  - Long term technical groups
  - Short term technical activities
- ***Consortium instead of a project!***

OSG core Infrastructure 2005

CE = Globus GK with **Prima** callout.

SE = SRM/dCache with **gPLAZMA** callout.

Discovery Service based on Clarens.

Diversity of MIS:

MonALISA, MIS-CI, grid cat, BDII & Glue 1.1

Accounting (MonALISA, ACDC)

VO administrative services (**Privilege Project** & VOMS)

Committed to interoperability with LCG/EGEE.

Authorization Infrastructure

(Privilege Project, Prima,GUMS, gPLAZMA)

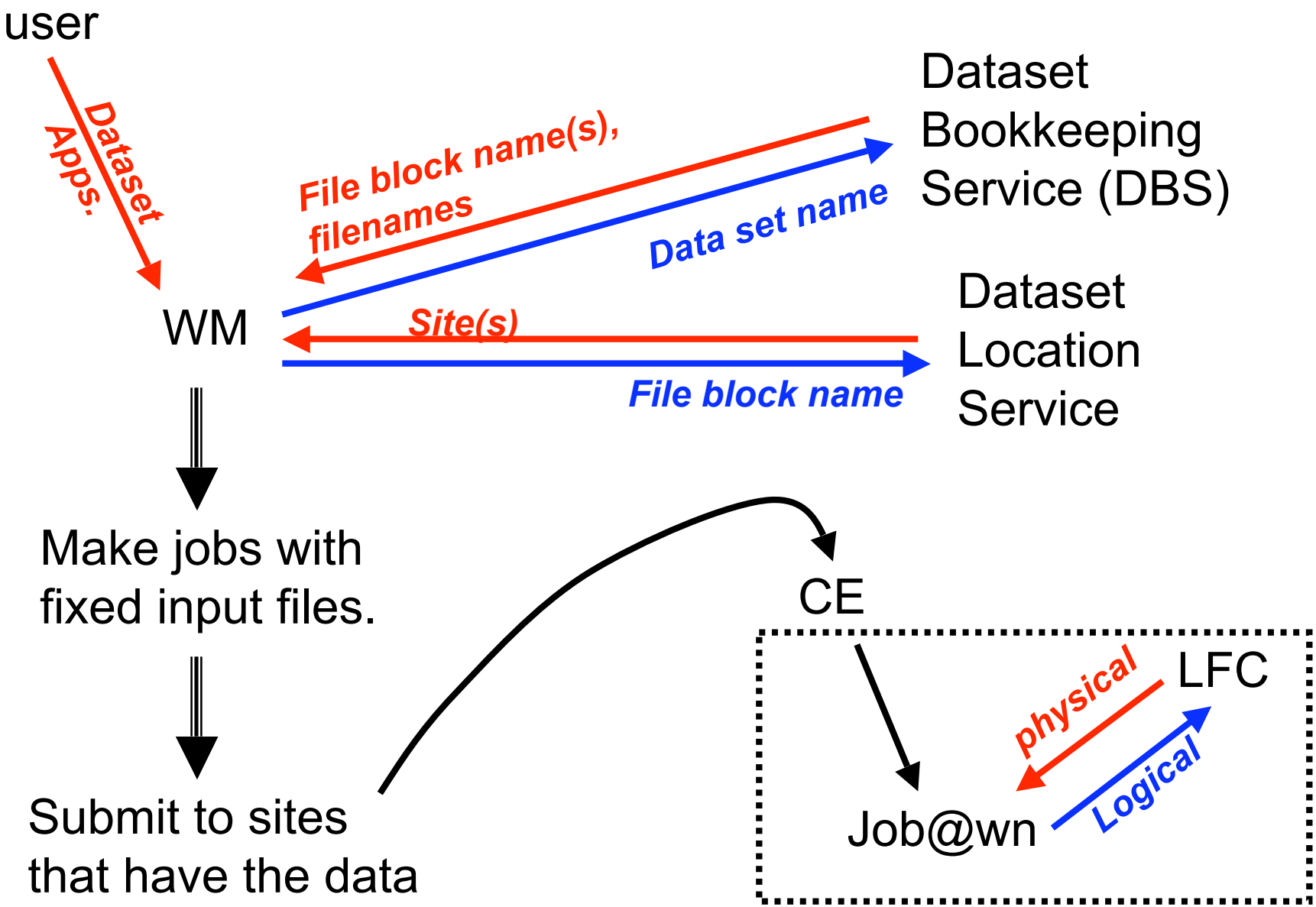
Goals:

- Move from host based to site based authz.
  - authz = VO-allowed & !site-vetoed
- Move from host based to service based granularity for authz.
- allow humans to have multiple roles at once.
  - Same DN can have multiple roles with privilege depending on role.
  - Multiple roles per DN means multiple privileges.
- create multi-user environment in which traditional UID based auditing is possible if desired.

Authz Implementation Status

- Prima & GUMS implement authz for CE
  - Being deployed presently as part of OSG
  - Optional component, sites may choose to stick with traditional gridmap file.
- gPLAZMA & GUMS implement authz for SE.
  - Coding & initial testing complete.
  - Integration testing starting.
  - Available only for SRM/dCache, and as such for US CMS T2 & FNAL.

CMS Baseline Services



Some features of CMS Baseline Services

- Blocks of files are statically placed at sites.
- Location service is at “block of files” level:
  - 1 block = O(1%) of Tier-2 disk space
  - 1 block = O(1‰) of CMS data volume (AOD)
- Input files for job fixed at submission time.
- Logical to physical translation is strictly local.

Some Thorny Issues

- **Development Environment on the grid**
  - Quick turn-around for small jobs
  - Debugging in grid environment
- **Managing Scarce Resources**
  - Flexible Policy Infrastructure
  - Dynamic cross-site replication
  - Fault tolerance & reliability
  - Ensure reasonable execution latencies

E.g. Flexible Policy Infrastructure

- Multi-experiment
- Multi-group and sub-group within experiment
- Multi-user within group
- Control of execution latency in Opportunistic resource environment
- User controls relative priority of their own workloads

Little of this exists in today's grids!

Three Conceptual Pieces

- Me - My friends - the anonymous grid
- Hierarchical task queues & late binding of resources.
- High Level Services

Me - My friends - the grid

Laptop	Persistent Services @ T2	Global Computing Grids
Me	***My friends***	***Anonymous world***

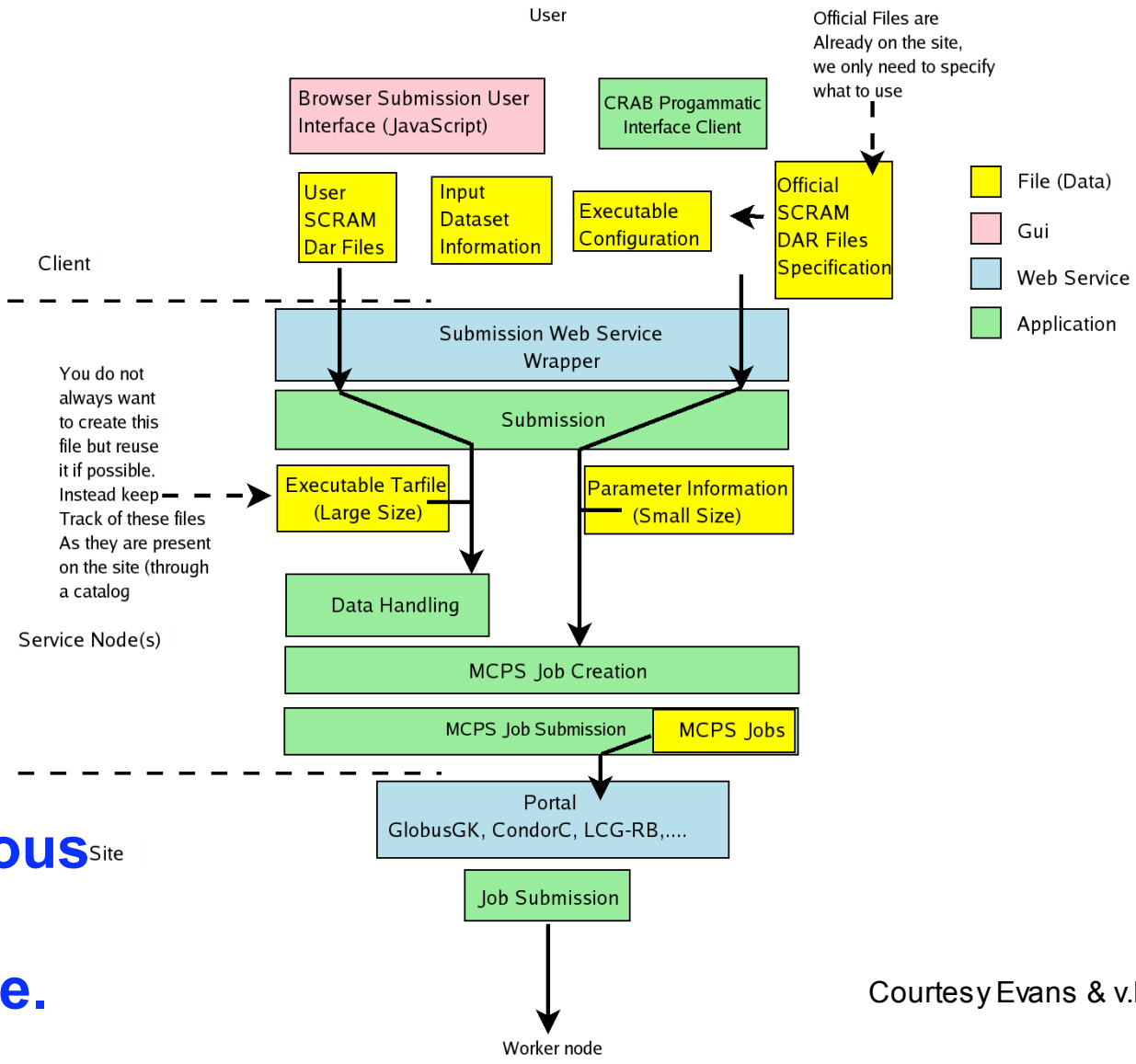
- physicists from UCLA will rely on services at T2 they work with no matter where they are working from.
- All the “heavy lifting” is done by my friends.

Example: Batch Data Analysis

Laptop

Tier-2

Anonymous Site
Grid site
Anywhere.

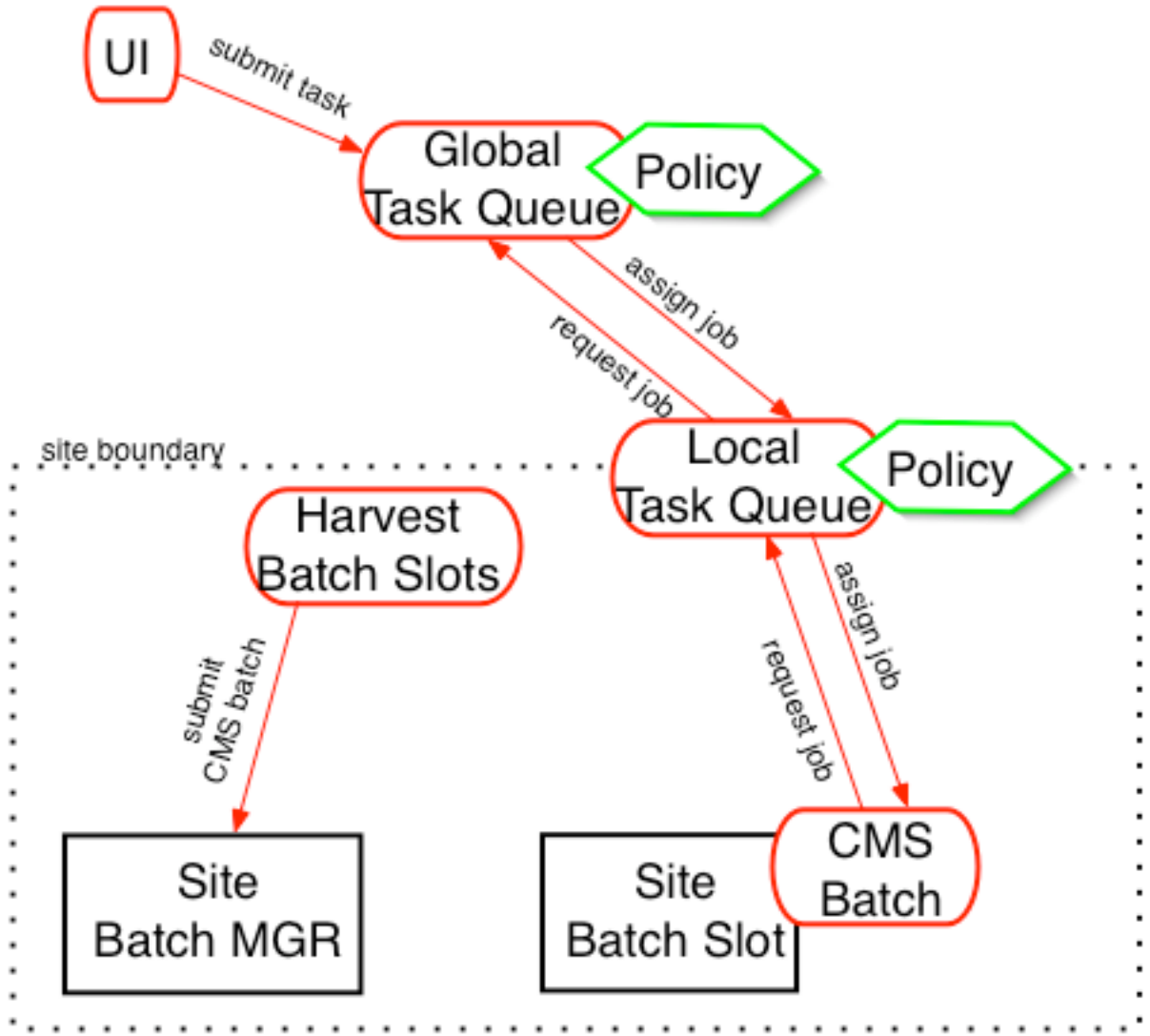


Courtesy Evans & v.Lingen& fkw

“Heavy lifting” done by friends

- Personal & group disk space quota.
- Group CPU quota.
- Grid submission service (incl. ds replication request, fault tolerance through resubmission, …)
- Job bookkeeping & task tracking
- Group & personal cvs (?)
- Group & personal ds & file catalogue (?)
- Group & personal provenance tracker (?)
- …

Hierarchical task queues

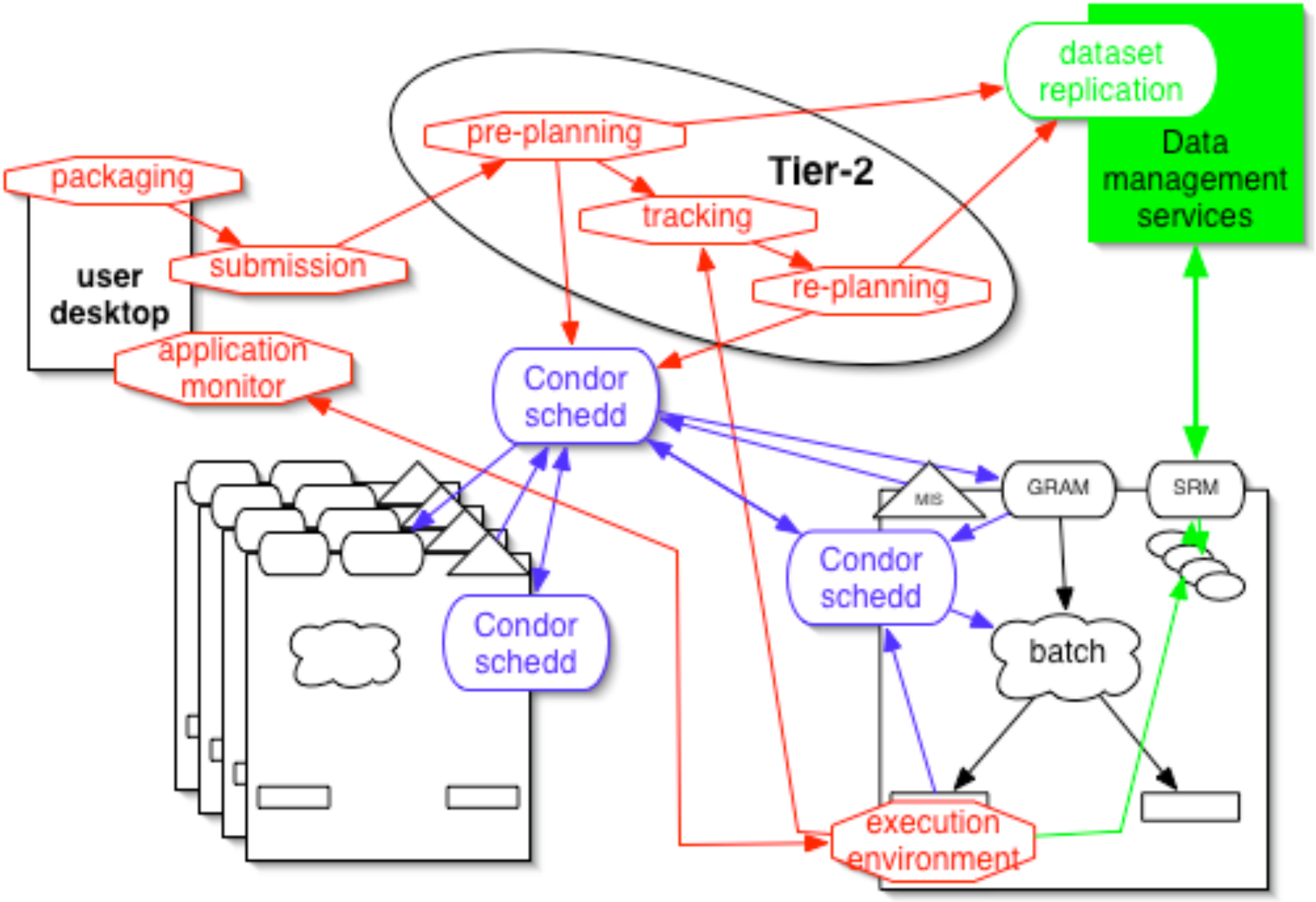


Dynamic policy.
&
“pull” from wn.

Responsibilities

- **CMS User global**
  - Turn user request into jobs.
  - Select sites & schedule jobs “in bulk”.
  - Track workload execution & intervene if needed.
- **CMS site local**
  - Harvest batch slots
  - Schedule jobs based on CMS CPU utilization policies (and file availability?).
  - Request more tasks from CMS User global as appropriate.

High Level Services



- **To be integrated out of:**
  - gLite (Condor-C, batch adapter(s))
  - Privilege Project & Globus
  - GAE/Clarens
  - CMS wm (BOSS, runjob, mcps, JobMon,…)
  - CMS dm (DBS, DLS, DP&TS)
  - Core OSG infrastructure
- **Building on experience of:**
  - Grid3/OSG
  - CDF CAF, especially glide-in CAF on INFNgrid.
  - Existing CMS services (PhEDEx, PubDB, RefDB, …)

Status 2 months ago: Tier-2s only for MC production

- No data access
- No multi-user environment
- No code development environment
- No reliable submissions interface

Address this with DISUN, the CMS Tier-2c proposal

DISUN Project Goals

- Provide user analysis infrastructure that is fully integrated with the central UAF at Fermilab, as well as user desktops everywhere.
- Deploy first production quality infrastructure towards the end of 2005.
- Incrementally increase functionality over a 5 year period.
- Expect 10^3-10^4 parallelization by 2007→2009

CMS Data Transfer Status 4/25/05

Transfer status

Age	Node	Files		On Site		Staged		Transferable	
		N	Size	N	Size	N	Size	N	Size
0h09	INFN_Bari	1200	19.6 GB	1200	19.6 GB	1200	19.6 GB	–	–
0h08	INFN_MSS	90228	13.7 TB	–	–	–	–	–	–
0h08	INFN_Transfer	93096	13.9 TB	90393	13.7 TB	–	–	23	4.2 GB
0h07	T1_ASCC_Buffer	237	10.8 GB	–	–	–	–	237	10.8 GB
0h06	T1_CERN_Buffer	17303	1.5 TB	13394	1.0 TB	8742	691.4 GB	–	–
0h05	T1_CERN_MSS	599001	69.4 TB	599001	69.4 TB	–	–	–	–
Current	T1_FNAL_Buffer	74928	5.3 TB	71404	4.9 TB	71402	4.9 TB	2795	285.3 GB
Current	T1_FNAL_MSS	239202	26.4 TB	206644	24.9 TB	–	–	32529	1.5 TB
Current	T1_FZK_Buffer	17207	2.0 TB	17196	2.0 TB	17196	2.0 TB	–	–
Current	T1_FZK_MSS	83180	13.3 TB	83180	13.3 TB	–	–	–	–
Current	T1_IN2P3_Buffer	14634	2.6 TB	14634	2.6 TB	–	–	–	–
Current	T1_IN2P3_MSS	14634	2.6 TB	14604	2.6 TB	–	–	30	5.6 GB
Current	T1_PIC_MSS	66739	7.2 TB	66739	7.2 TB	–	–	–	–
0h16	T1_RAL_Buffer	81000	10.5 TB	80994	10.5 TB	80994	10.5 TB	–	–
0h16	T1_RAL_MSS	80994	10.5 TB	–	–	–	–	–	–
0h16	T2_CIEMAT_Buffer	237	10.8 GB	237	10.8 GB	–	–	–	–
0h15	T2_Caltech_Buffer	963	225.4 GB	950	220.6 GB	–	–	13	4.8 GB
0h14	T2_Florida_Buffer	3107	616.1 GB	3107	616.1 GB	–	–	–	–
0h12	T2_Purdue_Buffer	10981	1.8 TB	2156	467.0 GB	–	–	8825	1.4 TB
0h12	T2_UCSD_Buffer	29957	1.3 TB	5857	286.8 GB	–	–	24108	1.1 TB
	Total	**1518828**	**182.7 TB**	**1271690**	**153.5 TB**	**179534**	**18.1 TB**	**68560**	**4.2 TB**

Status in 2 months

- Data access via PhEDEx & SRM/dCache
- Multi-user environment via Privilege Project.
- CMS software consistently installed & validated at all T2 sites.
- No code development environment
- No reliable submissions interface

Status in 9 months

(Baseline software stack for DC06)

- Data access via PhEDEx & SRM/dCache
- Multi-user environment via Privilege Project.
- Code development environment
  - Dynamic packaging of user code for submission
  - Dynamic installation of core CMS software
  - Interactive read-only access to batch job
- Transition to Condor-C for job submissions

Expect CMS user community to start using Tier-2s by next year.

Effort to make it happen is lead by DISUN

Summary

- CMS is getting organized to make user analysis computing at US Tier-2 centers a reality.
- Effort fully integrated into “distributed computing tools” group in US CMS S&C.
- First useable infrastructure in early 2006.