Panda/DDM Integration

Torre Wenaus (BNL), Tadashi Maeno (BNL)

BNL DDM Workshop
BNL
September 28, 2006

BROOKHAVEN
NATIONAL LABORATORY

Panda

- PanDA - **Production** and **Distributed Analysis** system
- Started August 2005 in US ATLAS as a full redesign to achieve performance, scalability, ease of operation needed for ATLAS datataking (up to 100-200k jobs/day)
  - Leverages past production experience
  - Designed to inherently support analysis
  - 'One stop shopping' for distributed processing
- In production since Dec 2005
  - Ambitious development milestones met
  - Thanks to productive development team
  - Still in rapid development, esp. analysis

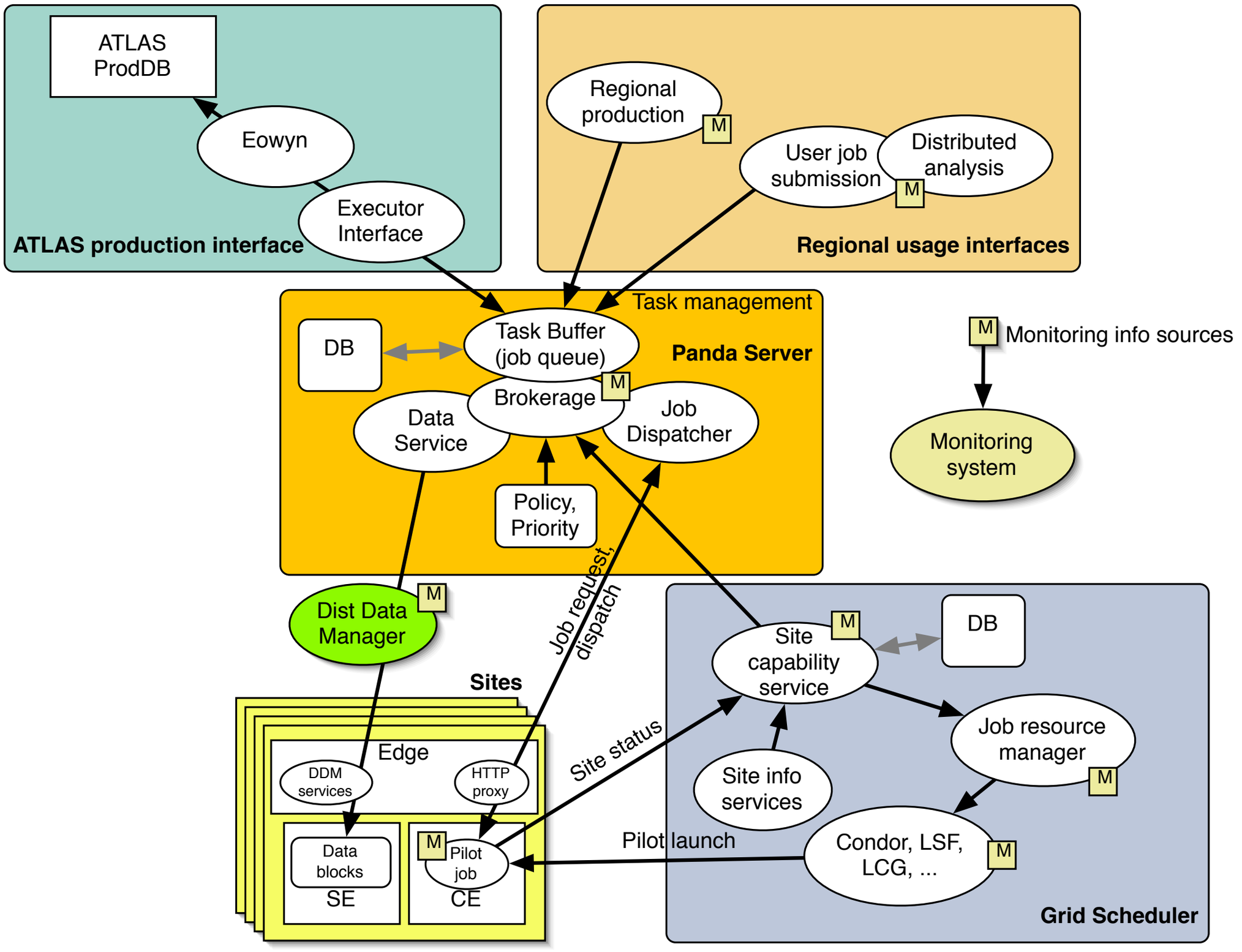
Key Panda Features, and Status

- Designed from beginning to support both managed production and individual users (analysis) via **flexible job spec/injection**
  - interactive analysis, user-level job submission, regional group production
    - first two implemented; regional group production not yet
  - grid-based or farm-based resources
    - currently supported: grid: CondorG, batch farm: PBS
- **Dataset** based organization of Panda matches the DDM system and the **analysis work model** (implemented, based on DQ2)
- Use of DDM to **pre-stage input data** and **immediately return outputs**, all asynchronously, minimizes data transport latencies and delivers earliest possible first results (implemented)
- Management/optimization of workload via **job queue** with **late binding** of jobs to worker nodes gives **dynamic and flexible** system response to **highly variable DA work** (implemented)
- Use of grid and/or farm batch queues to **pre-stage job wrappers** to worker nodes (pilot jobs) and **directly deliver workloads** from Panda allows **fast injection of DA work** (implemented)

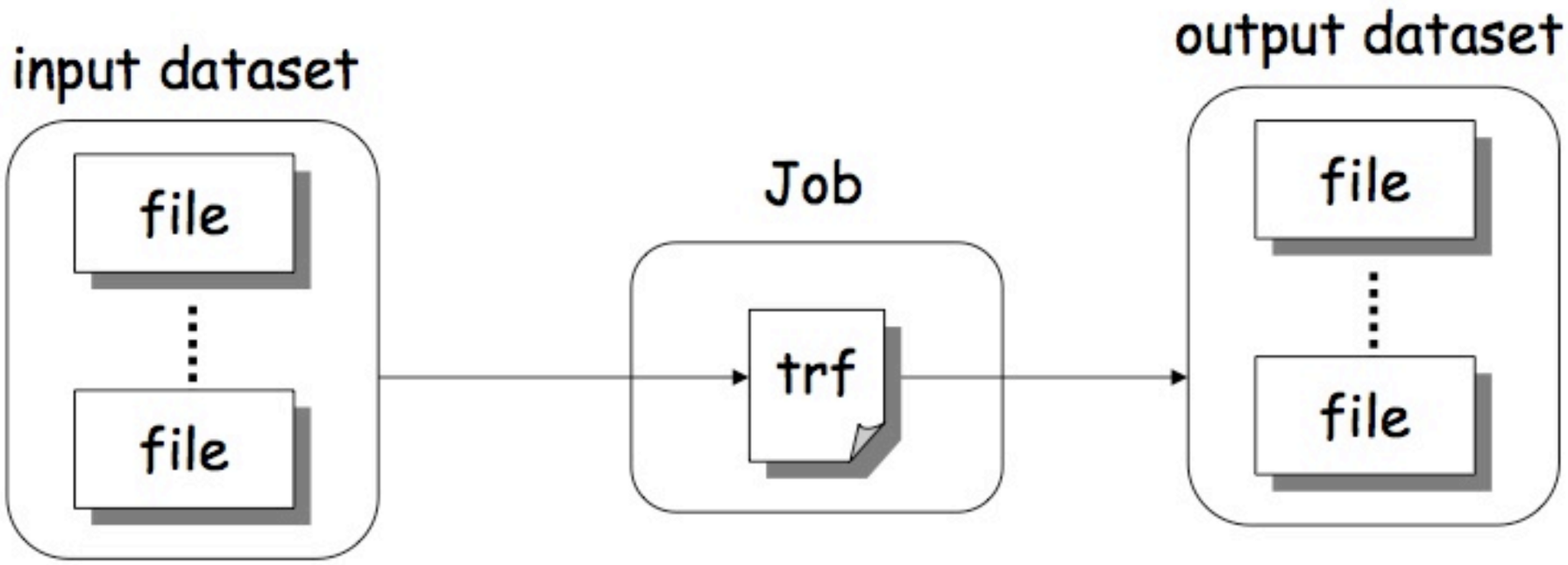
Key Panda Features (2)

- Support for packaging, deploying, running **arbitrary user code/jobs**
  - Implemented - arbitrary scripts can be specified by job and loaded (via http retrieval) for execution
- **Comprehensible system view** offered to users: **heterogeneous** distributed resources appear as one **uniform** resource accessed through standard interface (implemented)
- Easy to **integrate your own local resources**: site requirements are pilot delivery (via local batch queue or grid), outbound http, and access to a DQ2-enabled SE (locally or 'nearby') (only Tier 2s so far)
- Easy-to-use **client interface** makes integration with diverse analysis/ interactive front ends easy (implemented)
- **User ID** built into Panda DB; **monitoring and metadata** extensible to **user level (**User ID (DN) is recorded, user-level extensions not in yet)
- **User-level controls, quotas** directly implementable in Panda's brokerage rules (not implemented yet)
- Extensive **monitoring & browsing** (some specialization for DA, more to come

Panda Architecture



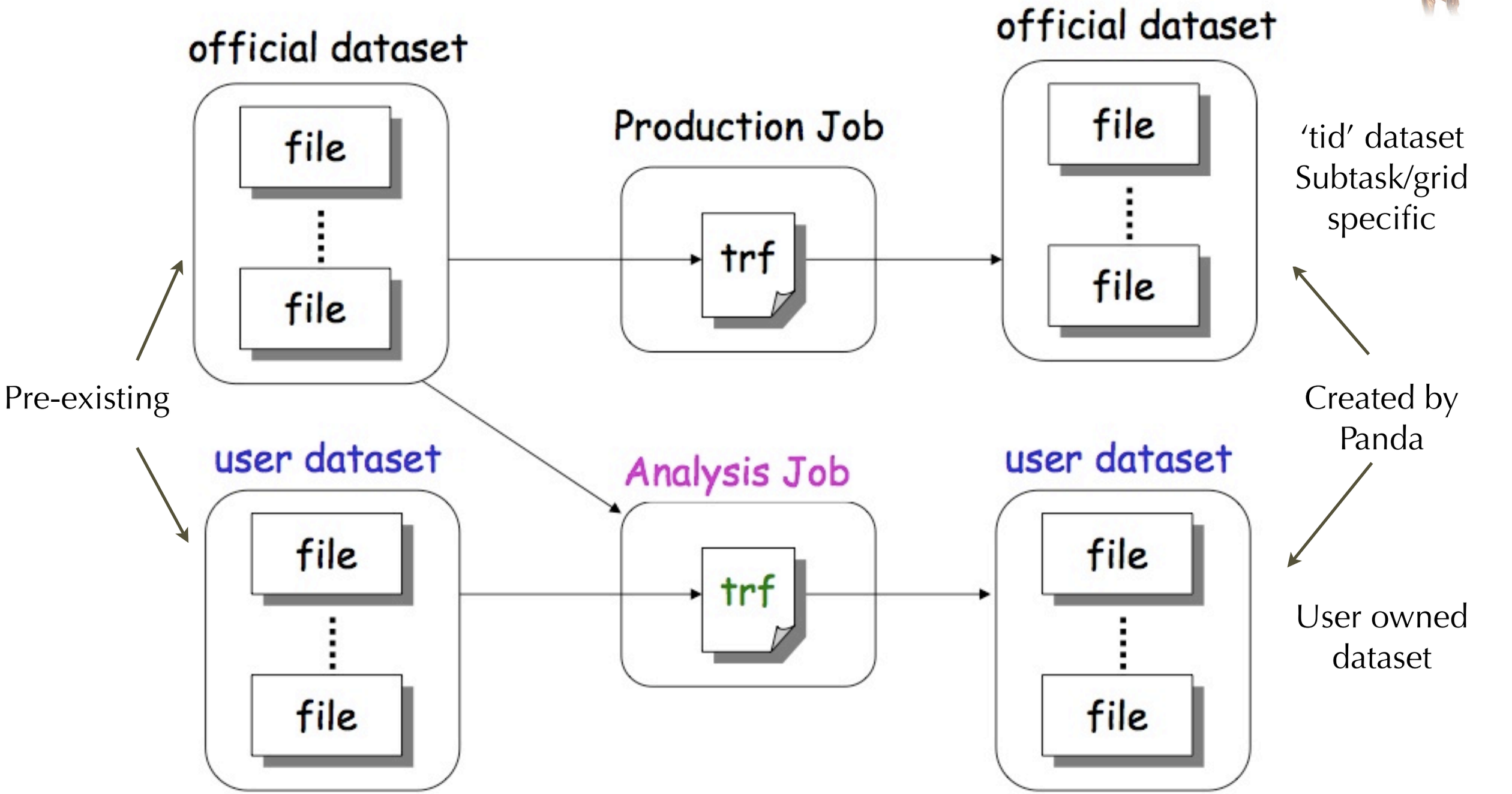
Datasets in Panda



- From a point of view of DDM/PANDA there is no essential difference between production jobs and analysis jobs
- User dataset can be accessed via DDM (e.g., using DQ2 end-user tools)

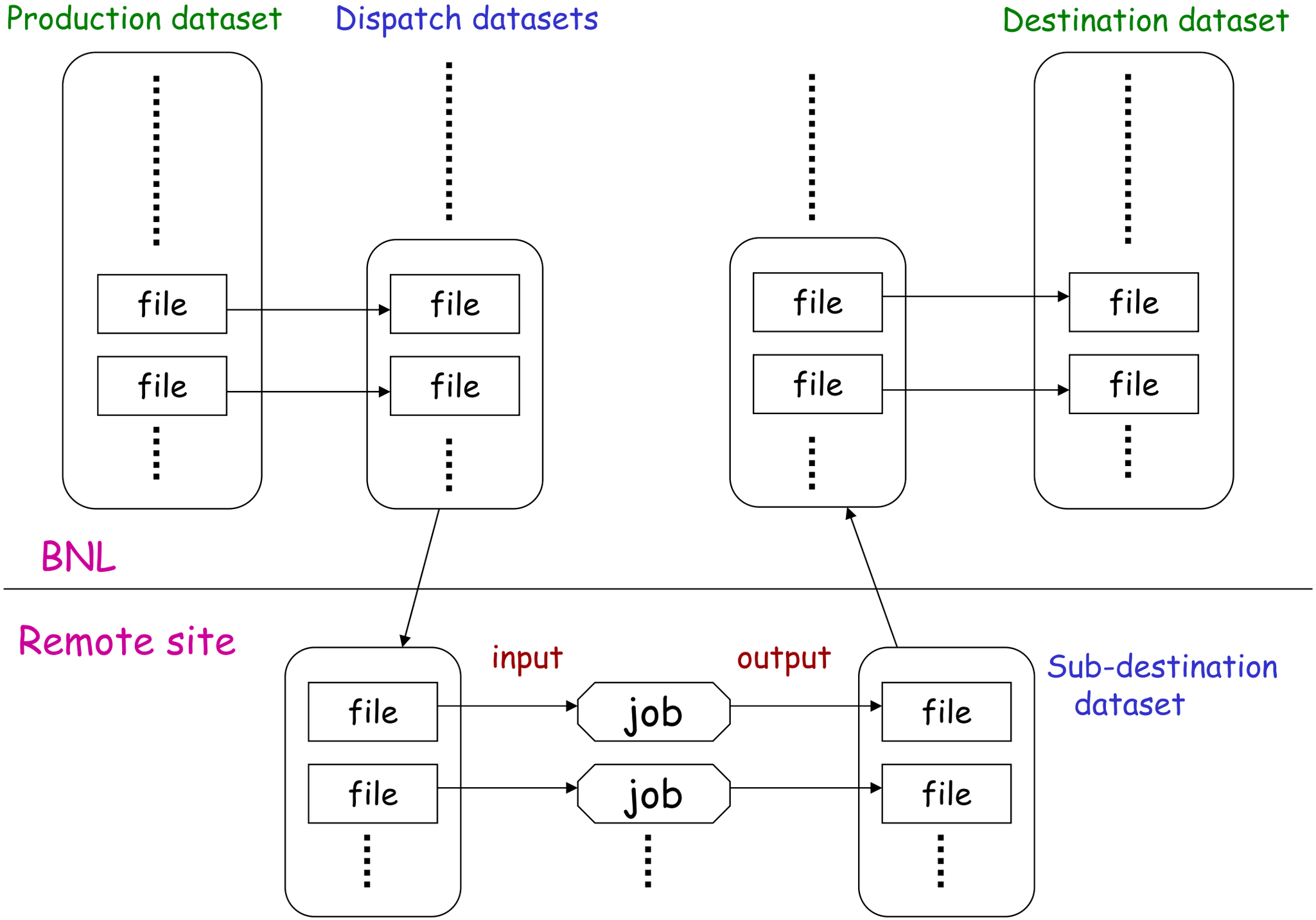
T. Maeno

Datasets in Panda (2)



T. Maeno

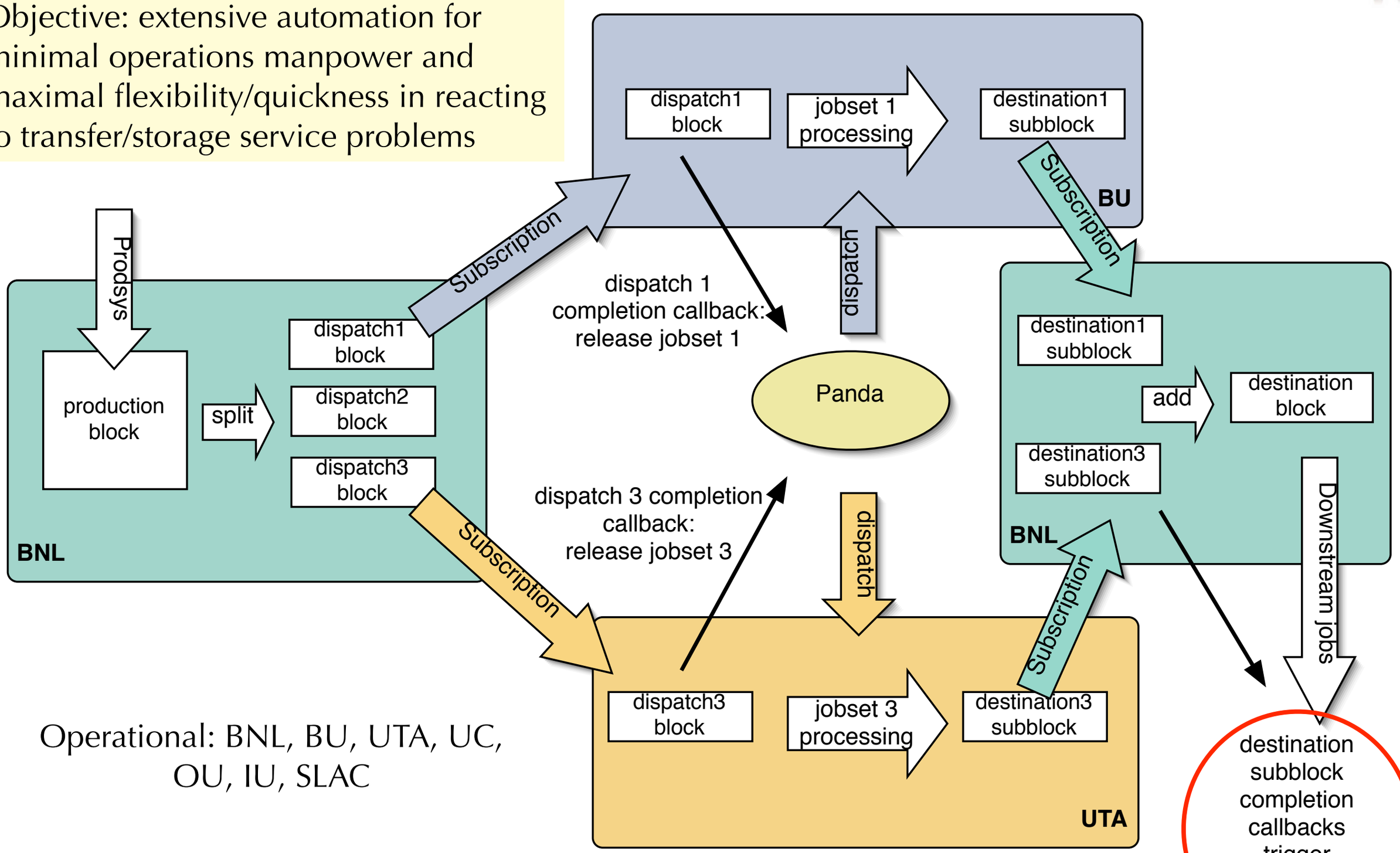
Dataset-based Data Flow in Panda



T. Maeno

DQ2 Data Handling in Panda

Objective: extensive automation for minimal operations manpower and maximal flexibility/quickness in reacting to transfer/storage service problems



Operational: BNL, BU, UTA, UC, OU, IU, SLAC

Possibility, not practice today

Present Panda DDM Operating Scale

TiersOfATLAS

- 7 facilities: BNL, BU, UTA, UC, OU, IU, SLAC
- ~15 DQ2 site installations (site services + LRC)
- ~900 active subscriptions, recent 12 hour period
- ~8k Panda data management datasets (25k total)
- 1.3M LFNs in BNL LRC (POOL FC MySQL)

D. Liko, Sep software week:

CSC11 AOD Data at Tier-1

	Datasets	Complete	Files	Size
ASGC	96	14	5634	520
BNL	226	131	17053	1736
CERN	253	106	16610	1712
CNAF	1	0	217	39
FZK	16	3	1510	172
LYON	72	13	6518	786
RAL	10	1	589	93
SARA	2	0	251	38
PIC	7	3	916	105
TRIUMF	7	3	510	77

CSC11 ESD Data at Tier-1

	Datasets	Complete	Files	Size
ASGC	62	6	3344	2721
BNL	141	96	11494	9507
CERN	99	7	4372	4569
CNAF	2	1	163	213
FZK	0	0	0	0
LYON	1	0	1	1
RAL	8	0	403	428
SARA	1	0	1	1
PIC	1	0	147	193
TRIUMF	5	2	555	18

DQ2 Callbacks

- Implemented in DQ2 at US request for Panda
  - Now extended and heavily used eg. for monitoring
- Useful in implementing data-driven workflow
  - Completion of input dataset dispatch to trigger job release
  - Completion of output dataset archiving to
    - signal data availability
      - data not available (=usable) until replicated to Tier 1 (empirical fact)
    - trigger release of downstream job
  - Automation of reprocessing
    - If the mass store system can tell us when a file has been staged in, daemon can deliver callback to trigger associated job release
- Depends on performant DQ2 subscription system

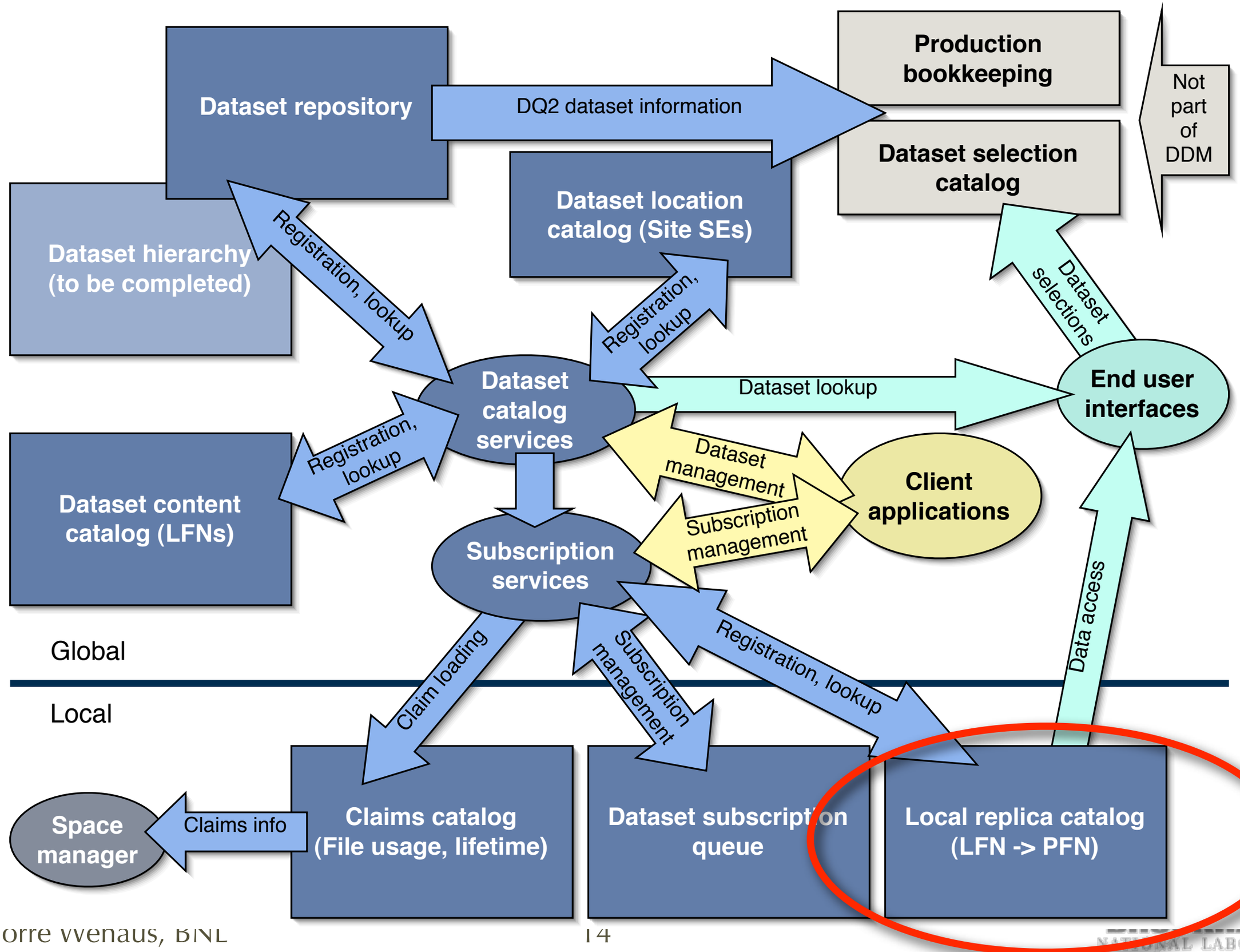
Pilot-Level Data Handling

- Pilot data handling responsibilities:
  - Input data
    - Get input data from local SE (where it's been placed by the dispatch block subscription) to the WN
  - Output data
    - Transfer outputs to local SE
    - Validate the transfer
    - Register the local SE instances with local LRC
    - Inform Panda server of outputs
      - XML FC transmitted to Panda dispatcher, which registers them with output (destination) dataset
      - Destination dataset subscription to BNL then takes care of archival replication to BNL
- Also pilot provides info to Panda (via local DDM http service)
  - Space remaining on local SE

Pilot-Level Data Handling Implementation

- panda/pilot2/DQ2ProdClient2.py
  - get_data, put_data functions
  - Site-dependent SiteMover implementations depending on nature and organization of the SE
- Sites currently in Panda production all have site-local SEs
- High current priority: add sites without site-local SEs ('opportunistic sites')
  - Implemented, based on use of uberftp for grid transfer with remote SE
    - uberftp because it supports remote md5sum check, subdirectory placement
  - Not yet in production
- New effort in generic pilot/scheduler 'TestPilot' beginning
  - Intended for non-US ATLAS, and non-ATLAS (OSG) Panda usage -- highly generic and customizable, and low install/deployment threshold
  - Requires DDM plug-ins by which a site, region, application, or VO can insert its own data handling
    - Currently working on LCG ATLAS version based on DQ2
    - To come: generic OSG version (have an OSG customer, CHARMM)
    - Will draw on DQ2ProdClient2.py for US ATLAS site support; explore merging

DQ2 Architecture



Replica Catalog

- MySQL implementation of POOL FC
- Simple implementation of basic functionality, vintage 2002
- Generally stable, scalable, efficient SQL access by bypassing POOL interface (which we do a lot)
- http web service front end provides lightweight client access with no client dependencies beyond http
- Weakness: Authentication. Currently based on standard MySQL user/pass
  - Possible to google user/pass in ToA; even write accounts until recently
  - Solution: grid certificate based authentication, implemented by Sasha Vaniachine et al, Wensheng et al working on production deployment
- Very amenable to wider ATLAS use?! Would (likely) require centralized LRCs at CERN serving Tier 1 clouds (we know how to run MySQL services @CERN)
  - Implement first as LFC backup, used as LFC mirror?

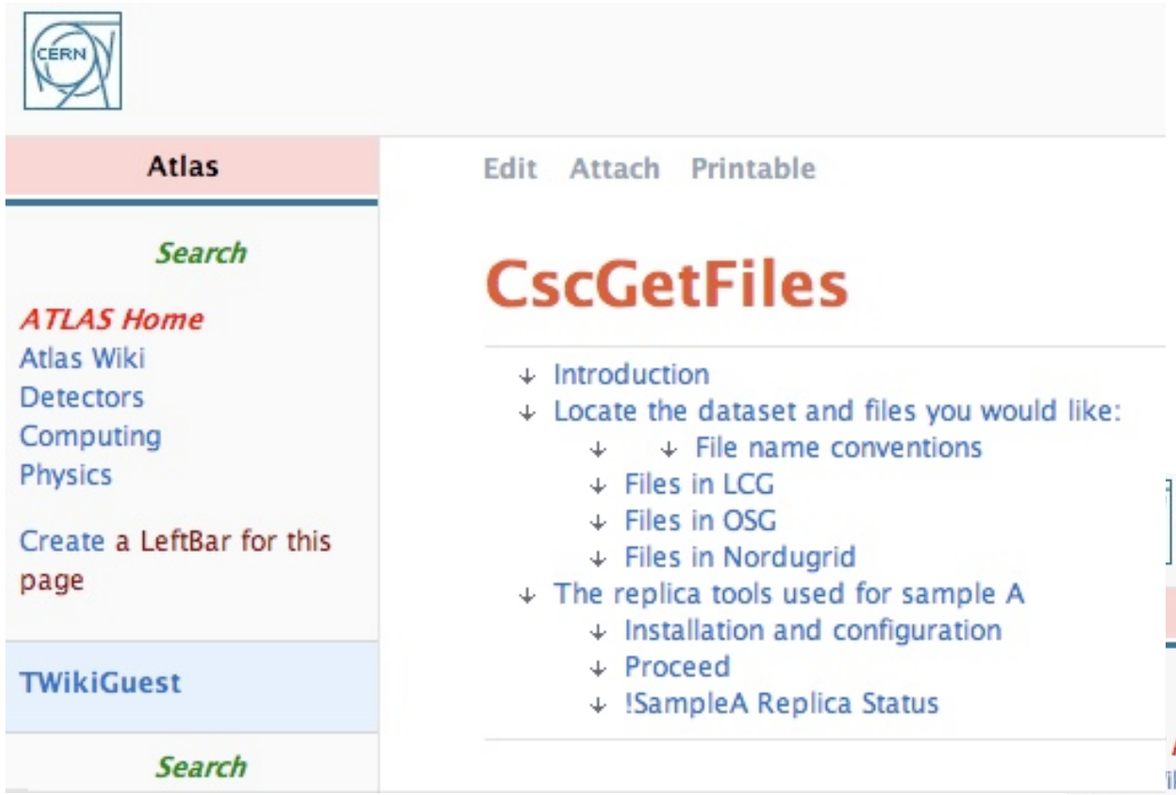
Grid Middleware Fallbacks

Slide untouched from April ATLAS grid workshop at CERN

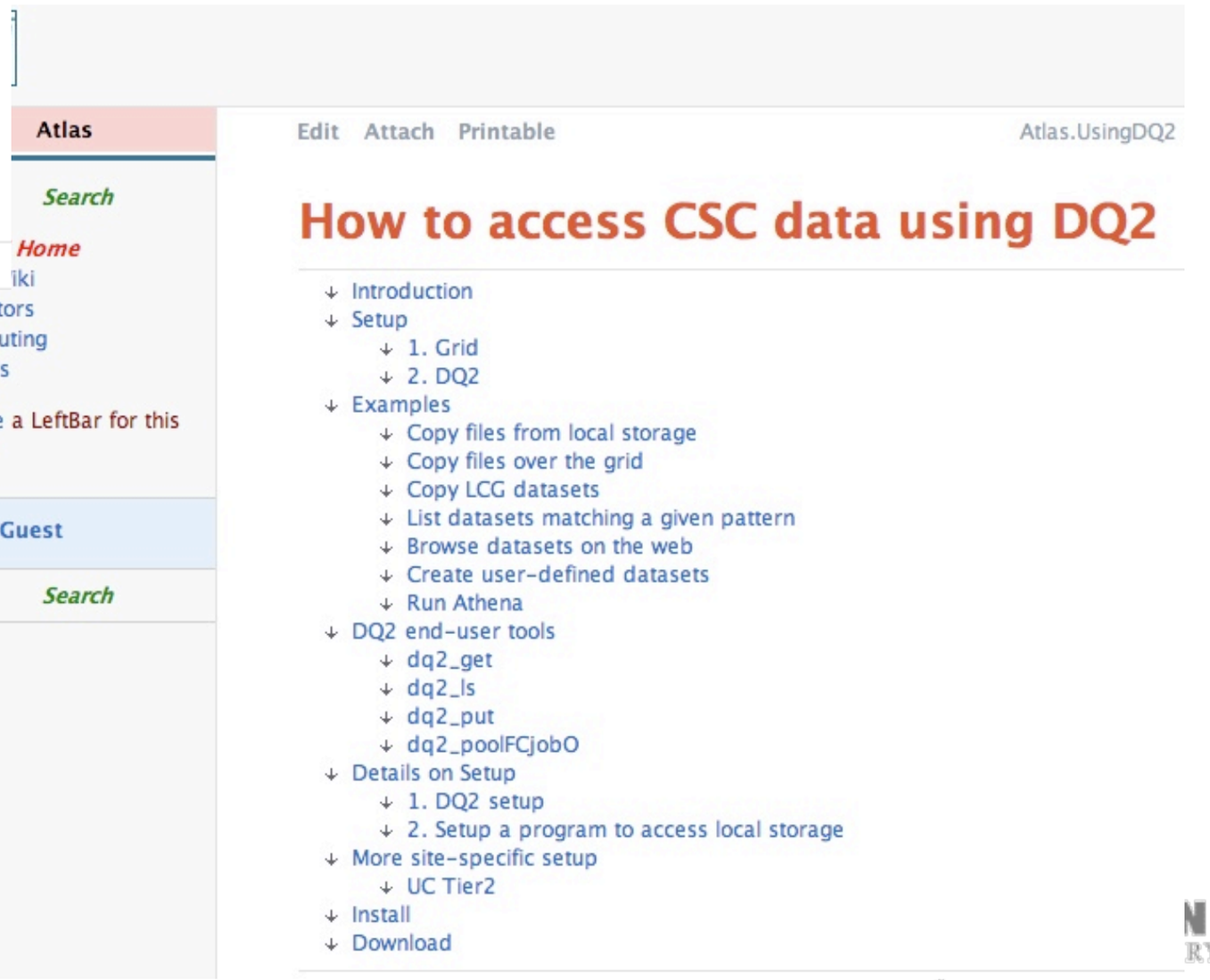
(Today I'd mention LFC! And note Miguel comment: Don't use SRM if you don't have to!)

- An important requirement in Panda and its DDM is support for *fallbacks* for external components, in particular immature unproven middleware
- FTS is in my view still immature unproven middleware
- Panda/DDM uses FTS for all data transfers at the moment (BNL<->Tier 2s) but this will not necessarily always be the case
- Panda/DQ2 has the capability now to swap out FTS for an alternative, and this has been exercised during periods of FTS problems
- If FTS (or any other middleware) fails and as a result DDM fails, it's *our* fault, if there was a potential fallback that we couldn't use
- DQ2 in the new version still supports fallbacks, and we (OSG) want to keep it that way

Data Access Documentation



CscGetFiles
Guidance for locating and accessing CSC data on the different grids



UsingDQ2
Documentation for setting up and using the dq2_* end user data handling tools

Torre Wenaus, BNL

Dataset Browser

Configuration | **Dashboards:** Production DDM Sites & Grids Analysis Physics data Task definition Quota

Update

Panda monitor

Quick guide, twiki

DQ2 dataset browser, category csc, datasets at BNLPANDA

Click for help
Dataset lists last updated 1 min ago

Jobs - search
Recent running, activated, waiting, assigned, defined, finished, failed jobs
Select analysis, production, test jobs

Quick search
Job
Dataset
Task
File

Summaries
Blocks: days
Errors: days
Nodes: days
Daily usage

Tasks - search
Generic Task Req
EvGen Task Req
CTBsim Task Req
Task list
Task browser

Datasets - search
Dataset browser
New datasets
Panda subscriptions
All subscriptions

Sites - see all
BNL BU IU OU SLAC
UC UTA LCG NG
Recent activity

Logging monitor

Select a dataset category *Counts are totals, exclusive of selections*

Category	Count	Description
All	25468	All datasets
T0	862	Tier 0 test
conditions	25	Datasets for conditions data files
csc	3812	Computing system commissioning production
ctb	86	Combined testbeam production
dc2	6	Data Challenge 2 production
destination	6281	Panda destination sub-blocks
dispatch	2105	Panda dispatch blocks
larg	52	LAr commissioning
mc	1129	MC validation production
other	1930	Everything else
rome	208	Rome physics workshop production
testpanda	1484	Panda test datasets
tile	52	Tilecal commissioning
user	1447	User datasets
validation	642	Validation samples (testIdeal* etc)

Restricted to BNLPANDA resident datasets (clear) (Restrict to datasets *not* at BNLPANDA)
Choose another site:

CANADA: ALBERTA, SFU, TORON, TRIUMF, TRIUMFDISK, TRIUMFTAPE, UVIC
CERN: CERNCAF, TIER0DISK, TIER0TAPE
FRANCE: BEIJING, CPPM, LAL, LAPP, LPC, LPNHE, LYON, LYONDISK, LYONTAPE, SACLAY, TOKYO
FZK: CSCS, CYF, DESY-HH, DESY-ZN, FZK, FZKDISK, FZKTAPE, FZU, UNI-FREIBURG, WUP
ITALY: CNAF, CNAFDISK, CNAFTAPE, LNF, MILANO, NAPOLI, ROMA1
NL: ITEP, SARA, SARADISK, SARATAPE, SINP
SPAIN: IFAE, IFIC, IFICDISK, IFICTAPE, PIC, PICDISK, PICTAPE, UAM
TAIWAN: ASGC, ASGCDISK, ASGCTAPE, AST2, AU-UNIMELB, TW-IPAS-T2
UK: RAL, RALDISK, RALTAPE, UKTIER2S, EDINBURGH, GLASGOW, MANC, LANCS, LIV, SHEF, RALPP, OXF, CAM, BHAM, ICL, LESC, RHUL, BRUN, QMUL, UCLCC, UCLHEP, DUR
USA: BNL, BNLDISK, BNLNULL, BNLTAPE, **BNLPANDA**, BNLTEST, BU, IU, IU_BC, OU, OUHEP, SLAC, UC, UC_TP, UC_VOB, UTA, UTA_SWT2, UTA_TEST1

Selected category: csc (clear category) No field selections active
csc field values:

series (1)	number (207)	physics (202)	stage (7)	format (9)	release (31)
csc11 (1339)	001.Gee_500 (1)	A10_150_Atautau_filter (9)	digit (239)	AOD (276)	root (1)
	005001 (24)	A10_300_Atautau_filter (1)	evgen (215)	CBNT (180)	v11000401 (110)
	005009 (2)	A10_Atautau_filter (1)	merge (28)	ESD (165)	v11000499 (3)
	005010 (10)	A3_Ztautau_filter (4)	recon (624)	EVNT (111)	v11000501 (19)
	005011 (6)	AcerMC_Zbb_4l (2)	recotrig (118)	HITS (111)	v11000502 (11)
	005012 (15)	AcerMCttbb (3)	simul (114)	RDO (140)	v11000504 (9)
	005013 (9)	AlpgenJimmyTopInInNp0 (9)	v11004103 (1)	TAG (8)	v11000505 (103)
	005014 (11)	AlpgenJimmyTopInInNp1 (9)		log (347)	v11000506 (5)
	005015 (19)	AlpgenJimmyTopInInNp2 (9)		oko (1)	v11000508 (16)
	005016 (20)	AlpgenJimmyTopInInNp3 (18)			v11000511 (6)
	005017 (14)	AlpgenJimmyTopInqqNp2 (2)			v11000512 (3)
	005020 (12)	AlpgenJimmyWenuNp3 (8)			v11004101 (8)

Tasks/Datasets In Action

Run by wenaus

Jobs - search
running, activated, waiting, assigned, defined, finished, failed
Analysis jobs
Old archive

Quick search
PandaID
Dataset
Task

Summaries
Blocks: days
Errors: days
Nodes: days

Tasks - search
Generic Task Req
EvGen Task Req
CTBsim Task Req
Full task list
Task browser

Datasets - search
In,out,dispatch,all
Subscriptions

Sites
Site specs
BNL BU OU UC
UTA LCG NG All

System statistics

Logging monitor

Task browser

Click for help

Current selection: STATUS=running&GRID=osg Clear selection

Click to show and select physics types

Releases: v11000301 (46) v11000302 (77) v11000303 (61) v11000304 (12) v11000305 (37) v11000306 (9) v11000307 (6) v11000308 (33) v11000401 (1)
Stages: digit (99) evgen (74) merge (18) reco (24) recon (67)
Outputs: AOD (106) CBNT (97) ESD (88) EVNT (74) HIST (9) HITS (102) RDO (102) TAG (18)
Grids: lcg (28) anygrid (1) nordic (52) osg (81) lcg-cg (120)
Status: aborted (12) done (115) finished (13) rejected (1) running (108) submitted (33)

Total selected events=2160000 jobs=21095 jobs done=16645

Task name	Task ID	Status	Grid	Total jobs	Done jobs	Events	Input files	Release	Formats
mc11.007204.singlepart_mu4.recon.v11000302	694	running	osg	2200	1498	220000	2200	11.0.3	ESD.AOD.CBNT
mc11.007430.singlepart_singlepi_pt2.digit.v11000308	667	running	osg	500	331	100000	20	11.0.3	RDO.HITS
mc11.007200.singlepart_mu2.recon.v11000303	579	running	osg	500	464	50000	500	11.0.3	ESD.AOD.CBNT
mc11.007200.singlepart_mu2.digit.v11000303									RDO.HITS
mc11.007222.singlepart_mu26.recon.v1100030									ESD.AOD.CBNT
mc11.007216.singlepart_mu18.recon.v1100030									ESD.AOD.CBNT
mc11.007211.singlepart_mu10.recon.v1100030									ESD.AOD.CBNT
mc11.007207.singlepart_mu6.recon.v1100030									ESD.AOD.CBNT
mc11.007216.singlepart_mu18.digit.v1100030									RDO.HITS
mc11.007207.singlepart_mu6.digit.v11000302									RDO.HITS
mc11.005001.pythia_minbias.recon.v1100030									ESD.AOD.CBNT
mc11.005001.pythia_minbia digit.v11000304									RDO.HITS
mc11.005800.JF17_pythia iet_filter.reco									ESD.AOD.CBNT
mc11.005800.JF17_pythia_lo filter.digit.v									RDO.HITS
mc11.005055.PythiaPhotonJet1. v1 03									ESD.AOD.CBNT
mc11.005056.PythiaPhotonJet2.reco 3									ESD.AOD.CBNT
mc11.005056.PythiaPhotonJet2.di									RDO.HITS

Task mc11.007216.singlepart_mu18.recon.v11000303

Datasets for task mc11.007216.singlepart_mu18.recon.v11000303

mc11.007216.singlepart_mu18.recon.ESD.v11000303
mc11.007216.singlepart_mu18.recon.AOD.v11000303
mc11.007216.singlepart_mu18.recon.CBNT.v11000303

Parameters for task mc11.007216.singlepart_mu18.recon.v11000303

Task ID	**562**
Project	mc11
Input dataset	mc11.007216.singlepart_mu18.digit.v11000302
Task name	mc11.007216.singlepart_mu18.recon.v11000303
Formats	ESD.AOD.CBNT
Transformation	csc.reco.trf
Trf Version	11.0.3.3
Release	11.0.3
Owner	i_hinchliffe@lbl.gov
CPU/event	100
Memory usage	600
First inputfile number	1
Input files	2200
Events	220000
Events/file	100
Grid	osg

Task browser

Click for help

Current selection: STATUS=running&GRID=osg

Physics type	Events
A3_Ztautau_tightfilter	20000
AlpgenJimmyW4jet	80000
Bs_Jpsi_mu6mu3_phi_KplusKminus	10000
Electron_Pt_25	20000
Electrons_e100	40000
FJ1_fwjets_e200	20000
FJ2_pythia_jetjet	100000
Gmm_500_pythia_photos	250000
H3_120_gamgam	20000
J1_pythia_jetjet	150000
J2_Pt_35_70	20000
J2_pythia_jetjet	150000
J3_Pt_70_140	20000
J3_pythia_jetjet	150000
J4_pythia_jetjet	150000
J5_Pt_280_560	20000
J5_pythia_jetjet	150000
J6_Pt_560_1120	30000
J6_pythia_jetjet	50000
J7_pythia_jetjet	30000
J8_pythia_jetjet	30000
JF17_pythia_jet_filter	3000000
JF17_pythia_loosejet_filter	1800000
LRSM_WR_1800_300	60000
M1_minbias	20000
McAtNloWenu	45000
McAtNloWmunu	60000
P5P_Single211	40000
P7P_Single211	20000
Photon_Pt_60	20000
Photons_e100	40000
PythiaH120gamgam	240000
PythiaH130zz4l	200000

Tasks define production tasks and record their associated metadata

Task Query Form

- *Text fields don't require the exact matching*
- Queries in *italic* not implemetted yet

MC Task

Project :	mc11
Input Dataset	
Transformation :	csc.reco.trf
Transformation Version :	11.0.4.1
Grid Flavour :	osg
Requested By	
Output Task Name	
Priority	
Status :	running

Datasets define and organize the task inputs and outputs

BROOKHAVEN NATIONAL LABORATORY

DQ2 Datasets In Action (Panda)

DQ2 catalogs datasets...

...provides user-level access: dq2_get, dq2_ls, (dq2_put coming)...

...provides automated management/movement tools for use in production...

...organizes validation...

...and ultimately lets you get at the files!

Dataset mc11.007216.singlepart_mu18.recon.AOD.v11000303

Task corresponding to dataset: mc11.007216.singlepart_mu18.recon.v11000303

Produced on grid(s): osg

Availability in DQ2:
Registered in osg DQ2, creation date Mon Jan 23 02:25:50 2006, vuid = 58aed7fc-4057-4b3c-b6e4-73f1f2fa152c
Not registered in lcg DQ2 instance

Data access
Instructions follow on how to access the data on the grids producing or holding it.
The instructions will become simpler once all grids are using DQ2.
In order to set up your environment to use the commands described,
follow the instructions here for running the setup script and initializing your grid certificate.

Access to data on OSG and LCG:
The dq2_get command can be used to retrieve the full dataset

```
dq2_get mc11.007216.singlepart_mu18.recon.AOD.v11000303
```

or a subset of files from it.
Internally dq2_get uses DQ2 where available (OSG) and direct LFC/LCG tools where necessary (LCG).
The dq2_ls command allows wildcarded listings of datasets or files within them, for DQ2 resident datasets.
See the documentation for dq2_get and dq_ls details.

Dataset info from Panda:
Type=output Status=running Created 2006-01-23 02:25:54 Modified 2006-01-26 14:57:57

Look for recent or all Panda jobs using this dataset for input or output
Look for recent or all Panda jobs using this dataset for input
Look for recent or all Panda jobs using this dataset for output

Click to show 512 subdatasets

BNL-validated dataset mc11.007216.singlepart_mu18.recon.AOD.v11000303_bnl available

OSG sites holding the dataset: BNL

338 constituent logical files in dataset mc11.007216.singlepart_mu18.recon.AOD.v11000303:

LFN	Click to view site replicas	
mc11.007216.singlepart_mu18.recon.AOD.v11000303._00001.pool.root.1	BNL	At BNL
mc11.007216.singlepart_mu18.recon.AOD.v11000303._00002.pool.root.1	BNL	At BNL
mc11.007216.singlepart_mu18.recon.AOD.v11000303._00003.pool.root.1	BNL	At BNL
mc11.007216.singlepart_mu18.recon.AOD.v11000303._00004.pool.root.1	BNL	At BNL

DDM Wish List

- Dataset catalog system OK but still some holes to fill
  - More bulk ops -- eg. efficient dataset metadata retrieval
  - Versioning system, closed/frozen datasets now in place but, from usage, not sufficient
    - Very few datasets frozen; need more awareness in system of partial datasets, %complete, empty datasets
- Subscription system much more problematic
  - Extremely long subscription servicing cycle
  - All the issues we're discussing here of replication robustness and performance
- Local replica catalog
  - US MySQL LRC generally working well, LFC is a problem (for US as well)
    - Need production deployment of grid certificate based authentication for MySQL LRC (expected very soon?)
  - We need scalable robust LRCs ATLAS-wide

DDM Wish List (2)

- Clean, simple, documented packaging/installation for OSG
  - As much coherence as possible between LCG and OSG
  - cf Miguel's comments, we should use distutils directly? Cleanly split off LRC installation?
- Eventually: system partitioning. US DQ2 system publishing to ATLAS
- DQ2 monitoring much better; next is monitoring of the monitoring
  - 'Intelligent layer' watching for prevalent errors, raising alarms
- 'Claims system' that became a couple more LRC attributes
  - Need a real plan for dataset- as well as file- level space management
- More end-user support
  - DDM down to the laptop
  - User subscriptions, upgraded dq2_get (copy to local SE and register)
  - Opening up FTS site-to-site access -- Dan's proposal
- More developers! Particularly tool-making: consistency checking, site data management, etc. Have had very strong contributions from site people, more welcome!

DDM Wish List (3-10)

- Plus all the things I've forgotten ;-)

Current/Future Activities

- Top priority: extending Panda to opportunistic US ATLAS sites, which is mainly a DDM issue
  - And extending Panda yet further -- LCG ATLAS in particular
- Top priority: robustness, stability, fault tolerance
  - DDM failures both in production and analysis way too common. Partly a facility issue, partly fault tolerance (retry, timeouts etc.) in the DDM system
- User-level support
  - Today neither dq2_get nor asking users to use subscriptions is adequate
- Many efforts/improvements coupled closely to ATLAS level program and decisions
  - LRC, grid authentication, http interface
  - Subscription implementation, monitoring, quality of service
  - DDM monitoring
  - System partitioning (better to do it *before* we urgently need it); partitioning by region (OSG production) and function (user catalogs)
- Site services
  - Space management, consistency between storage and LRC, uniform SRM/xxxx deployment, SRM 2.2, stable ToA endpoints

From Aug Tier 2 Workshop

http://www.usatlas.bnl.gov/twiki/bin/view/Admins/TierTwoStorageDataServices

- Tier2 DDM responsibilities, tasks
  - Connect to Tier 1 via FTS
  - Connect to all other US Tier 2s via FTS
  - Provide SRM based storage. Current baseline: SRM/dCache
  - Investigate xrootd for possible role (esp. SLAC)
  - Provide DQ2 site service
    - Site should have single, stable endpoint; ToA config should be stable
    - Site (re)configuration should be internal, not exposed at the site service address
  - Provide US standard LRC and associated http service
    - Separate MySQL from site service strongly recommended
  - Support agreed, US supported space usage controls
  - Support agreed, US standard & supported end user data access tools
    - Today, requirement includes LCG UI if direct LCG data access is desired (available from BNL AFS)
    - But data aggregation at BNL should make this mostly unnecessary