

The ATLAS Grid Centers and ATLAS Distributed Computing

Andrej Filipčič

Overview

- ATLAS Distributed Computing organization
- ATLAS data and job workflow
- ATLAS Grid Centers, site requirements
- Tier-3 policy

ADC (simplified)

- Central services:

- DDM: data management, dataset transfers, catalogs
- PanDA: central task/job distribution and control system
- Task brokering: job run on sites with input files on local storage
- Autopilot factories submit pilot jobs to CE, payload picked from PanDA when running on node
- Monitoring: EGI/NGI infrastructure + ATLAS system
- ...

- Sites:

- Computing Element (CREAM, OSG, ARC)
- SRM Storage Element
- cvmfs for ATLAS software (ATHENA, database files)
- Squid, frontier: database access + caching
- Additional services in Tier-1 sites (LFC, FTS, DB, ...)

Sites and Tier Levels in ATLAS

- Tier-0: CERN
- Tier-1: 10 large computing centers (disk, cpu, tape), permanent storage
- Tier-2: ~100 medium centers associated to one of Tier-1s (disk, cpu), secondary storage
- Tier-3: ~80 non-pledged resources
 - Full grid sites => ATLAS Grid Centers
 - Analysis sites with no grid infrastructure
- 10 ATLAS clouds: Tier-1 + associated Tier-2 and Tier-3
- Tier-0,1,2: pledged resources, Memorandum of Understanding
- The hierarchical model is not strict any more and is/will change in the future

ATLAS Clouds

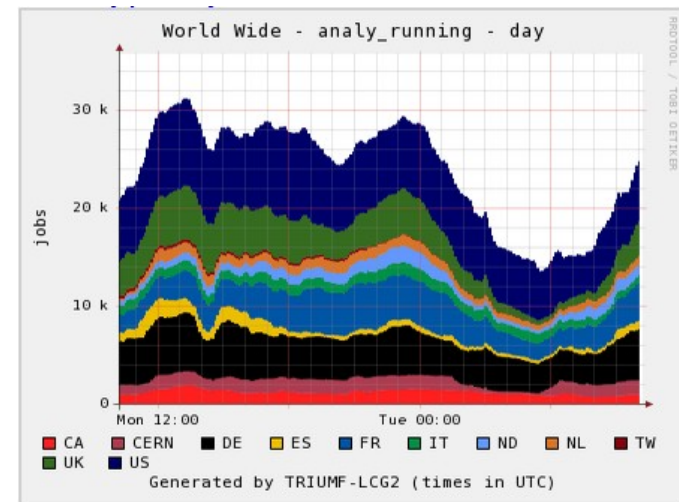
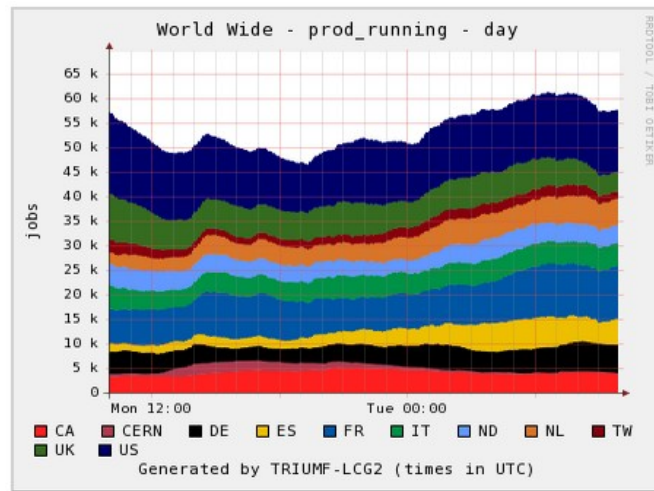
Cloud status: TLo, THi are transfer timeouts for low and high priority jobs

Cloud	Tier1	Status	Comment	Tasks	TLo days	THi days	Storage used by Panda	Free space (GB)	as of	Workload sz2kdays	Weight	Nprestige
CA	TRIUMF	online		75	4	3	TRIUMF-LCG2_DATADISK,TRIUMF-LCG2_DATAPE,TRIUMF-LCG2_MCTAPE,TRIUMF-LCG2_MCDISK	196165	09-16 21:44	79786	1000000	3000
CERN	CERN-PROD	online	ended.eos.migration.13sept	34	2	1	CERN-PROD_DATADISK,CERN-PROD_DATAPE,CERN-PROD_MCTAPE	248847	09-16 21:44	25189	1000000	4000
DE	FZK-LCG2	online		80	2	1	FZK-LCG2_DATADISK,FZK-LCG2_DATAPE,FZK-LCG2_MCTAPE,FZK-LCG2_PRODDISK	395191	09-16 21:44	203349	5000	4000
ES	pic	online (XavierEspinal at 09-01 10:28)		57	4	1	PIC_DATADISK,PIC_DATAPE,PIC_MCTAPE,PIC_MCDISK	148220	09-16 21:44	150286	5000	2000
EB	LYON	online		185	2	1	IN2P3-CC_DATADISK,IN2P3-CC_DATAPE,IN2P3-CC_MCTAPE,IN2P3-CC_MCDISK	644181	09-16 21:44	247948	2	3000
IT	INFN-T1	online	elog.29191	92	2	1	INFN-T1_DATADISK,INFN-T1_DATAPE,INFN-T1_MCTAPE	549798	09-16 21:44	62696	5000	2000
ND	ARC	online		85	2	1	NDGF-T1_DATADISK,NDGF-T1_MCDISK,NDGF-T1_DATAPE,NDGF-T1_MCTAPE	315978	09-16 21:44	64683	-1	2000
NL	SARA-MATRIX	online		264	4	1	SARA-MATRIX_DATADISK,SARA-MATRIX_DATAPE,SARA-MATRIX_MCTAPE,SARA-MATRIX_MCDISK	90843	09-16 21:44	464213	10	2000
OSG	BNL_ATLAS_1	online			0	0						0
TW	Taiwan-LCG2	online		28	4	3	TAIWAN-LCG2_DATADISK,TAIWAN-LCG2_DATAPE,TAIWAN-LCG2_MCTAPE	201641	09-16 21:44	133260	2	2000
UK	RAL-LCG2	online	savannah.123292.solved	108	2	1	RAL-LCG2_DATADISK,RAL-LCG2_DATAPE,RAL-LCG2_MCTAPE,RAL-LCG2_MCDISK	519315	09-16 21:44	297829	5000	2000
US	BNL_ATLAS_1	online (fbarreir at 08-29 22:30)	ELOG.123173	188	4	1	BNL-OSG2_DATADISK,BNL-OSG2_MCTAPE,BNL-OSG2_DATAPE	471712	09-16 21:44	448879	2	12000

- Cloud: Tier-1 + associated Tier-2s
- Support team
- 24/7 operation on Tier-1

NL_Tasks	Jobs	assigned	5441	activated	14420	running	9481	holding	229	transferring	8616
ANALY_NIKHEF_GLEXEC	ANALY_NIKHEF_GLEXEC										
IL-TAU-HEP	ANALY_IL-TAU-HEP										
ITEP	ITEP-ce3-atlas-icqpbs										
IIIR-LCG2	ANALY_IIR										
NIKHEF-ELPROD	ANALY_NIKHEF-ELPROD										
BRC-KI	ANALY_BRC-KI										
BU-Protvino-IHEP	ANALY_IHEP										
SARA-MATRIX	ANALY_SARA										
TECHNION-HEP	ANALY_TECHNION-HEP										
TR-10-ULAKBIM	ANALY_TR-10-ULAKBIM										
WEIZMANN-LCG2	ANALY_WEIZMANN										
csTCDie	ANALY_CSTCDIE										
ru-Moscow-FIAN-LCG2	ANALY_FIAN										
ru-Moscow-SINP-LCG2	ANALY_SINP										
ru-PNPI	ANALY_PNPI										

ATLAS Resources



- 100k cores
- 60PB of disk space
- few 10GB/s network capacity
- >200 sites (clusters)
- gLite, OSG, ARC middleware

ATLAS tasks

Tasks Requests

Task name	Task ID	Req Jobs	Done Jobs	Total events	Prio	Grid	State	PostProd	Timestamp
data11_7TeV.00189781.physics_Egamma.merge.f405_m985_p692	527222	10	0	68000	750	panda@nl	submitting	group	Sep 27 09:25
data11_7TeV.00189781.physics_Egamma.merge.f405_m985_p678	527221	10	0	68000	750	panda@nl	submitting	group	Sep 27 09:25
data11_7TeV.00189781.physics_Egamma.merge.f405_m985_p716	527220	10	0	68000	750	panda@nl	submitting	group	Sep 27 09:25
data11_7TeV.00189781.physics_Egamma.merge.f405_m716_f405_p677	527219	1	0	7000	750	panda@nl	submitted	group	Sep 27 09:25
data11_7TeV.00189781.physics_Egamma.merge.f405_m985_p677	527215	10	0	68000	750	panda@nl	submitting	group	Sep 27 08:23
data11_7TeV.00189813.physics_CosmicCalo.merge.f405_p714	527204	8	3	183000	750	panda@nl	running	group	Sep 27 06:23
data11_7TeV.00189639.physics_Muons.merge.f405_m986_p676	527181	10	0	74000	750	panda@nl	submitting	group	Sep 27 03:24
data11_7TeV.00189693.physics_Muons.merge.f405_m985_p703	527175	126	10	126000	750	panda@nl	running	group	Sep 27 02:21
data11_7TeV.00189693.physics_Muons.merge.f405_m985_p682	527173	10	0	126000	750	panda@nl	submitting	group	Sep 27 02:21
data11_7TeV.00189774.physics_Egamma.merge.f405_m985_p716	527151	8	0	8000	750	panda@nl	submitted	group	Sep 26 22:25
mc10_7TeV.105805.filtered_minbias6.simul.e574_s1339	527039	10000	10	10000000	200	panda@nl	running	atlas	Sep 26 14:25
valid1.128380.Pythia_ggF_H115bb.recon.e830_s933_s946_r2759	526893	10	6	10000	800	panda@nl	running	atlas	Sep 26 10:22
valid1.107054.PythiaWtaunu_incl.recon.e574_s933_s946_r2759	526848	20	7	20000	800	panda@nl	running	atlas	Sep 26 10:22
valid1.105338.HerwigVBFH120tautauhh.recon.e598_s933_s946_r2759	526830	10	9	10000	800	panda@nl	running	atlas	Sep 26 10:22
mc10_2TeV.105014.J5_pythia_jetjet.digit.e913_s1186_s1161_d585	526782	10	0	100000	600	panda@nl	submitting	atlas	Sep 26 09:24
mc10_2TeV.105013.J4_pythia_jetjet.digit.e913_s1186_s1161_d585	526781	10	0	100000	600	panda@nl	submitting	atlas	Sep 26 09:24
mc10_2TeV.105012.J3_pythia_jetjet.digit.e913_s1186_s1161_d585	526780	10	0	100000	600	panda@nl	submitting	atlas	Sep 26 09:24
mc10_2TeV.105011.J2_pythia_jetjet.digit.e913_s1186_s1161_d585	526779	10	0	100000	600	panda@nl	submitting	atlas	Sep 26 09:24
mc10_2TeV.105010.J1_pythia_jetjet.digit.e913_s1186_s1161_d585	526778	10	0	100000	600	panda@nl	submitting	atlas	Sep 26 09:24
mc11_7TeV.119113.Hijing_PbPb_2p75TeV_MinBias_Flow_JV2.recon.e844_s1336_s1300_d581_r2758	526757	2500	1295	25000	600	panda@nl	running	atlas	Sep 26 07:25
mc11_7TeV.117410.AlpgeJimmyWgammaNp0_pt20.merge.e873_s1310_s1300_r2730_r2700_p716	525585	10	9	210000	600	panda@nl	submitting	group	Sep 23 12:22
mc11_7TeV.116392.AlpgeJimmyGamNp3_jetFilter_Nj2Et20.merge.e825_s1310_s1300_r2730_r2700_p7	525554	698	549	10000000	600	panda@nl	submitting	group	Sep 23 12:22
mc11_7TeV.116121.HerwigWj_pt100_met.merge.e825_s1310_s1300_r2730_r2700_p716	525546	15	7	150000	600	panda@nl	running	group	Sep 23 12:22
mc11_7TeV.116101.AcerMC_gg_ttbbQCD.merge.e835_s1310_s1300_r2730_r2700_p716	525537	9	8	90000	600	panda@nl	running	group	Sep 23 12:22
mc11_7TeV.113210.HerwigppjetsJ6.merge.e835_s1309_s1300_r2730_r2700_p716	525533	40	30	400000	600	panda@nl	running	group	Sep 23 12:22
mc11_7TeV.113141.AlpgeJimmyNjetsNp4_J1x.merge.e825_s1310_s1300_r2730_r2700_p716	525480	10	6	100000	600	panda@nl	running	group	Sep 23 12:22
mc11_7TeV.107941.Ttbar_FullHad_PowHeg_Pythia.merge.e887_s1310_s1300_r2730_r2700_p716	525417	99	97	1000000	600	panda@nl	running	group	Sep 23 12:22
mc11_7TeV.107682.AlpgeJimmyWenuNp2_pt20.merge.e825_s1299_s1300_r2730_r2700_p716	525395	375	359	3770000	600	panda@nl	running	group	Sep 23 12:22
mc11_7TeV.107661.AlpgeJimmyZmumuNp1_pt20.merge.e835_s1299_s1300_r2730_r2700_p716	525382	133	129	1335000	600	panda@nl	running	group	Sep 23 12:22
mc10_7TeV.105805.filtered_minbias6.merge.e574_s1339_s1344	525240	2	1	10000000	680	panda@nl	submitting	atlas	Sep 23 10:23
mc10_7TeV.140005.simplifiedModel_wA_slep_26.merge.e838_a126_r2300	525176	4	3	20000	670	panda@nl	running	atlas	Sep 23 08:24
mc10_7TeV.139980.simplifiedModel_wA_slep_1.merge.e838_a126_r2300	525167	25	25	140000	670	panda@nl	submitting	atlas	Sep 23 08:24
mc10_7TeV.139980.simplifiedModel_wA_slep_1.recon.e838_a126	525157	280	275	140000	620	panda@nl	running	atlas	Sep 23 08:24
mc11_7TeV.105922.McAtNlo_JIMMY_WpWm_enumunu.merge.e872_s1310_s1300_r2730_r2700_p716	524968	10	10	200000	600	panda@nl	submitting	group	Sep 22 20:25
mc10_7TeV.105011.J2_pythia_jetjet.merge.e913_s1186_s1161	521545	65	65	100000	590	panda@nl	submitting	atlas	Sep 19 13:22
mc10_7TeV.105011.J2_pythia_jetjet.simul.e913_s1186	521520	2000	1979	100000	400	panda@nl	running	atlas	Sep 19 13:22

Production and Analysis Jobs

Cloud	Pilots	Latest	defined	assigned	waiting	activated	sent	starting	running	holding	transferring	finished	failed	cancelled	%fail
ALL			15099	0	0	74549	7	792	25410	5231	173	132473	20974	27804	14%
CA	1069	09-27 10:05	1883	0	0	2099	2	2	950	205	0	6378	1521	44	19%
CERN	2167	09-27 10:05	455	0	0	417	0	0	1462	72	0	5513	1874	3	25%
DE	3960	09-27 10:05	1786	0	0	9677	1	0	5667	1289	0	26892	3614	2256	12%
ES	1163	09-27 10:05	413	0	0	8334	0	0	916	170	0	5316	770	3496	13%
FR	3811	09-27 10:05	1272	0	0	9081	1	5	4178	581	0	19246	1242	241	6%
IT	1220	09-27 10:05	133	0	0	8910	0	0	915	154	0	8977	174	474	2%
ND	131	09-27 10:05	797	0	0	275	0	782	1128	122	173	4534	608	257	12%
NL	1351	09-27 10:05	617	0	0	8503	0	0	785	147	0	6653	456	136	6%
TW	25	09-27 10:05	132	0	0	4620	0	0	0	0	0	279	13	1	4%
UK	2556	09-27 10:05	6348	0	0	7535	1	0	3313	1752	0	19115	1962	18263	9%
US	5918	09-27 10:05	1263	0	0	15098	2	3	6096	739	0	29570	8740	2633	23%

Region	Pilots	Latest	defined	assigned	waiting	activated	sent	starting	running	holding	transferring	finished	failed	cancelled	%fail
ALL			15	15176	0	94331	10	5538	57556	1589	71640	147280	17166	303	10%
CA	1067	09-27 09:50	0	358	0	6515	0	3	3971	98	3900	10544	503	4	5%
CERN	577	09-27 09:50	0	0	0	0	0	0	103	17	0	1682	17	0	1%
DE	3173	09-27 09:50	15	4892	0	9108	4	8	6438	82	12515	15481	865	8	5%
ES	1440	09-27 09:50	0	0	0	7335	0	0	1937	172	2622	17708	8975	269	34%
FR	4295	09-27 09:50	0	2877	0	12029	0	0	7152	160	4978	20375	1707	6	8%
IT	2025	09-27 09:50	0	712	0	8356	0	13	4284	117	2343	11312	207	0	2%
ND	761	09-27 09:50	0	0	0	997	0	5456	3648	84	66	7712	241	4	3%
NL	1805	09-27 09:50	0	86	0	7050	0	1	4619	84	3466	11034	970	3	8%
TW	645	09-27 09:50	0	0	0	4717	0	4	1402	116	0	6533	7	0	0%
UK	3535	09-27 09:50	0	5826	0	10185	0	5	7422	205	21896	8453	3085	5	27%
US	5391	09-27 09:50	0	425	0	28039	6	48	16580	454	19854	36446	589	4	2%

Dataset Organization and Transfers

NL	89%	572 MB/s	11609	898	10853	1433	0	0	
Click on the site name to go to the site page, "*" to see statistics for this site per source									
CSTCDIE_DATADISK	0%	0 kB/s	0	0	0	1	0	0	unkno
CSTCDIE_HOTDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
CSTCDIE_PRODDISK	100%	3 MB/s	138	14	132	0	0	0	unkno
CSTCDIE_SCRATCHDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
CSTCDIE_SOFT-TEST	0%	0 kB/s	0	0	0	0	0	0	unkno
IL-TAU-HEP_DATADISK	100%	5 MB/s	25	0	22	0	0	0	unkno
IL-TAU-HEP_HOTDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
IL-TAU-HEP_LOCALGROUPDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
IL-TAU-HEP_PHYS-SM	0%	0 kB/s	0	0	0	0	0	0	unkno
IL-TAU-HEP_PRODDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
IL-TAU-HEP_SCRATCHDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
ITEP_DATADISK	0%	0 kB/s	0	0	0	1	0	0	unkno
ITEP_HOTDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
ITEP_LOCALGROUPDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
ITEP_PRODDISK	100%	214 kB/s	3	2	2	0	0	0	unkno
ITEP_SCRATCHDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
J1NR-LCG2_DATADISK	100%	4 MB/s	284	2	284	0	0	0	unkno
J1NR-LCG2_HOTDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
J1NR-LCG2_LOCALGROUPDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
J1NR-LCG2_PHYS-SM	0%	0 kB/s	0	0	0	0	0	0	unkno
J1NR-LCG2_PRODDISK	100%	4 MB/s	164	18	164	0	0	0	unkno
J1NR-LCG2_SCRATCHDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
NIKHEF-ELPROD_DATADISK	70%	436 MB/s	2438	59	2389	1045	0	0	unkno
NIKHEF-ELPROD_DET-MUON	0%	0 kB/s	0	0	0	0	0	0	unkno
NIKHEF-ELPROD_HOTDISK	0%	0 kB/s	0	0	0	3	0	0	unkno
NIKHEF-ELPROD_LOCALGROUPDISK	0%	0 kB/s	0	0	0	0	0	0	unkno
NIKHEF-ELPROD_PERF-IDTRACKING	0%	0 kB/s	0	0	0	0	0	0	unkno
NIKHEF-ELPROD_PERF-MUONS	0%	0 kB/s	0	0	0	0	0	0	unkno
NIKHEF-ELPROD_PHYS-HIGGS	0%	0 kB/s	0	0	0	0	0	0	unkno
NIKHEF-ELPROD_PHYS-SM	31%	4 MB/s	79	0	79	176	0	0	unkno
NIKHEF-ELPROD_PHYS-SUSY	0%	0 kB/s	0	0	0	0	0	0	unkno
NIKHEF-ELPROD_PHYS-TOP	0%	0 kB/s	0	0	0	0	0	0	unkno
NIKHEF-ELPROD_PRODDISK	100%	7 MB/s	496	70	496	0	0	0	unkno

- Dataset: collection of files of the same type
- Storage Endpoint: disk (tape) space allocated on site for particular activity
 - ➔ Example: NIKHEF-ELPROD_DATADISK
 - ➔ <SITE>-<SPACETOKEN>
- File Transfer Services: copies datasets from one site to another
- Endpoints:
 - ➔ PRODDISK – temporary storage of production job inputs/outputs
 - ➔ SCRATCHDISK – global analysis inputs and outputs
 - ➔ LOCALGROUPDISK – storage for local use
 - ➔ ...

Task/Job workflow

- Task brokered (allocated) to particular cloud
- Input dataset transferred to the site SRM endpoint
- PanDA defines and activates the jobs for the site
- Pilot jobs on the site, when started on node:
 - Pick a job from PanDA site queue
 - Copies inputs from local SE
 - Runs the job
 - Copies output to local SE
- Transfer of output datasets from the site SE to associated Tier-1 storage

Overview of Job Types

- Monte-Carlo simulation:

- Event Generation (100MB input, 10 minutes cputime, 1GB output)
- Simulation, tracking through the detector (1GB shared input, 10h cputime, 100MB output)
 - . This is the only job type that can run on low bandwidth sites

- MC Reconstruction:

- 1GB input, 5h cputime, 5GB output

- Data reprocessing:

- 5GB input, 1h cputime, 5GB output

- Merging:

- 5GB input, 10 minutes cputime, 5GB output

- User Analysis:

- 20GB input, 1h cputime, 100MB output (but varies wildly)

- Global analysis with PanDA: user runs analysis on the container (dataset collection)

- Up to 10000 jobs on 100TB of input data by a single user
- Very heavy for nodes and storage: typical local I/O 10MB/s per job

ATLAS Computing Sites

- Sites, that pledged resources (Tier-0,1,2)
 - ➔ ATLAS “owns” the pledged resources (central disk space management, node job allocation)
- Voluntary sites:
 - ➔ Smaller clusters with cpus for ATLAS
 - ➔ Analysis sites for private use by ATLAS collaborators
 - ➔ National grid sites
 - ➔ Large clusters with extra resources for ATLAS
 - ➔ (commercial) clouds
 - ➔ ...

How to become ATLAS Grid Center

- **First step: International Computing Board (ICB)**

- ➔ Agreement on cloud association support (NL)
- ➔ Agreement of national physics contact and funding agency representative
- ➔ Site requirements: 50-cores, 5TB of SRM storage (2TB proddisk, 3TB scratchdisk), 10Mb/s network connectivity
- ➔ 0.5 FTE of site support
- ➔ EGI/NGI rules, 75% uptime

- **Second step: ATLAS Distributed Computing (ADC) registration**

- ➔ Site testing and validation (test jobs, test transfers)
- ➔ Site certification, registration in central system

- **Functional categories: defines job types which can run on the site**

- ➔ High priority production
- ➔ ATLAS-wide analysis
- ➔ Low priority production
- ➔ Distributed data site
- ➔ Other ATLAS computing activities

- **Site Job Configuration: through PanDA queues.**

- ➔ brokeroff → jobs will be submitted to the site only if the site is explicitly requested
- ➔ Site (or ADC) can choose whether to run global production, global analysis, or allow only custom user jobs

Detailed site requirements

- Full EGI/NGI site certification + production status
- ATLAS requirements:
 - ➔ ATLAS VO support in grid services
 - ➔ Squid service
 - ➔ cvmfs on nodes
 - ➔ 5.5 GB of free disk space per job (core)
 - ➔ Configured SRM storage endpoint with ATLAS ACL
 - 2TB PRODDISK if running production jobs
 - 3TB of SCRATCHDISK if running analysis jobs
 - 5TB of LOCALGROUPDISK if site wants to store local or private datasets
- Resources not strict, site can start with lower resources if it plans to grow in the future
- SRM storage mandatory, no jobs possible without it

ARC alternative to gLite

- Used in Nordugrid cloud
- Single server: ARC frontend
- Input file caching => less transfers
- No permanent storage needed => inputs/outputs directly transferred from/to NDGF-T1. ARC server takes care of controlled/throttled transfers
- Job types that run on a particular site can be limited

Yerevan Tier-3 Status

- Two sites: YERPFI, AM-04-YERPFI
- Both EGI production status
- Configured services:
 - BDII, CREAM-CE, Classic-SE
 - Atlas specific: xrootd, squid, cvmfs
- Need SRM endpoint, DPM the easiest to setup
- Testing the site can start after that
- Need to clarify what the site functionality
- (NL) Cloud support essential → the base cell to solve the problems or implement the new requirements. The cloud support seems to be insufficient for all the new sites. Eastern countries encouraged to form a support unit around Dubna.

Conclusions
