NDGF Site report: Experiences with ARC/dCache

Josva Kleist

Technical Director, NDGF

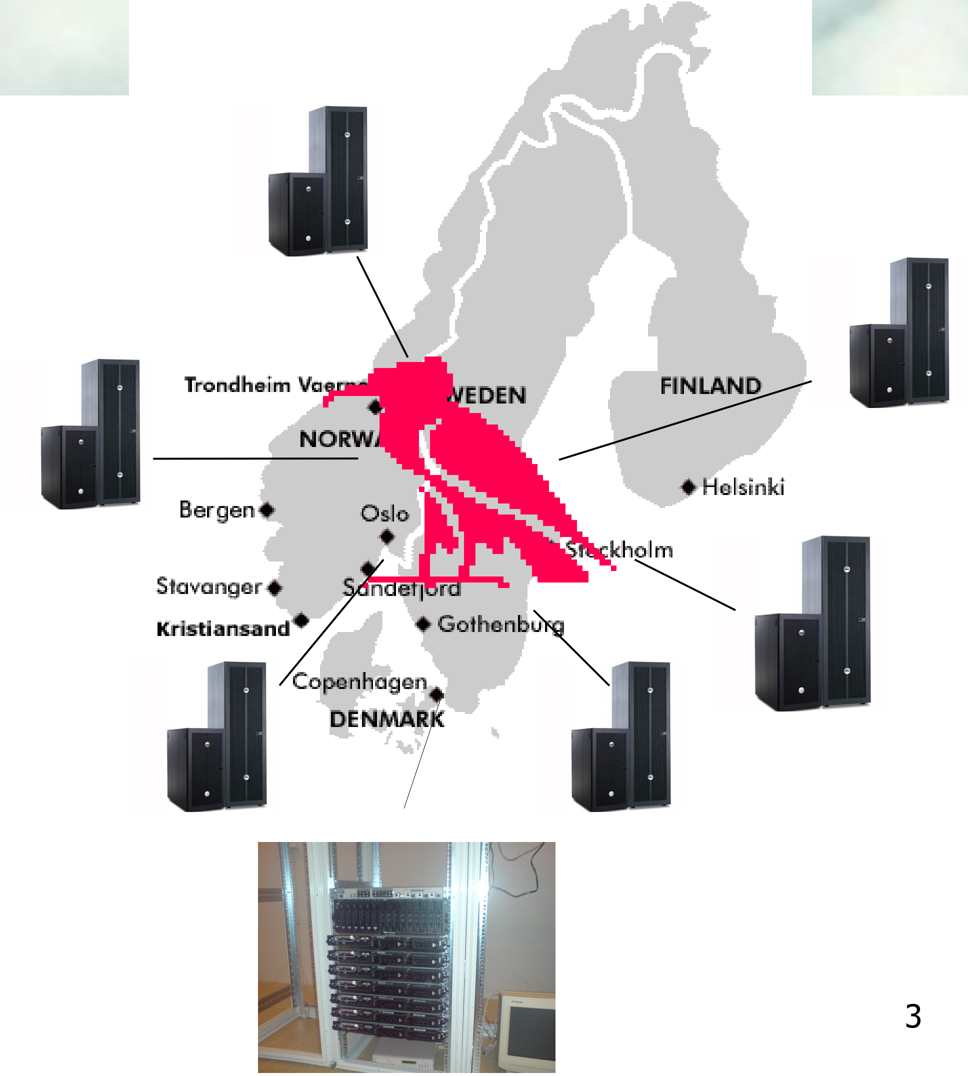
Amsterdam, June 17, 2010

NDGF is different

- A distributed virtual infrastructure (NDGF does not own any resources).
- "The Tier-1 is only about storage".
- Storage and compute resources are placed at national scientific compute centers.
- Computational resources are shared with other research communities.
- All computations carried out at "Tier-2" sites.
- Relatively fat network pipes to most sites.

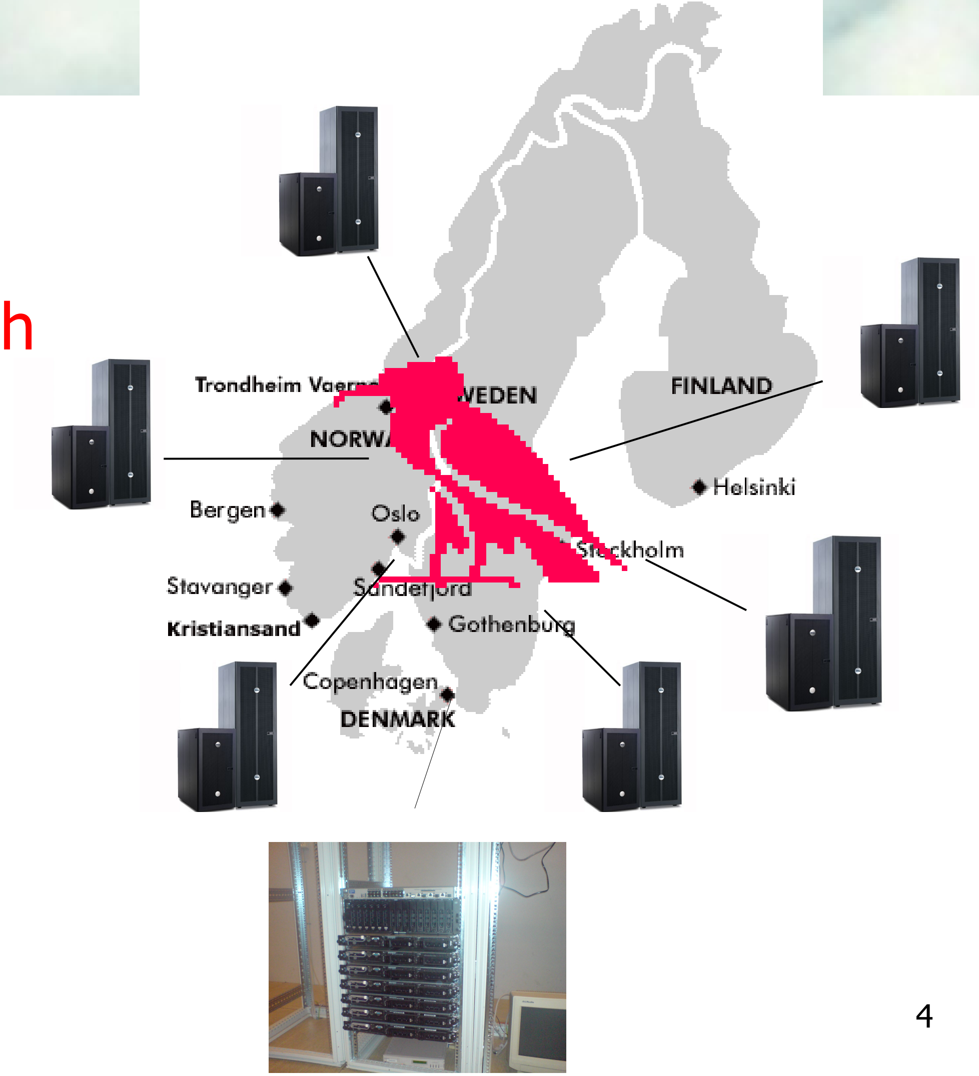
Distributed dCache

- dCache Installation
- dCache head nodes at Nordic GEANT endpoint
- Pools at sites

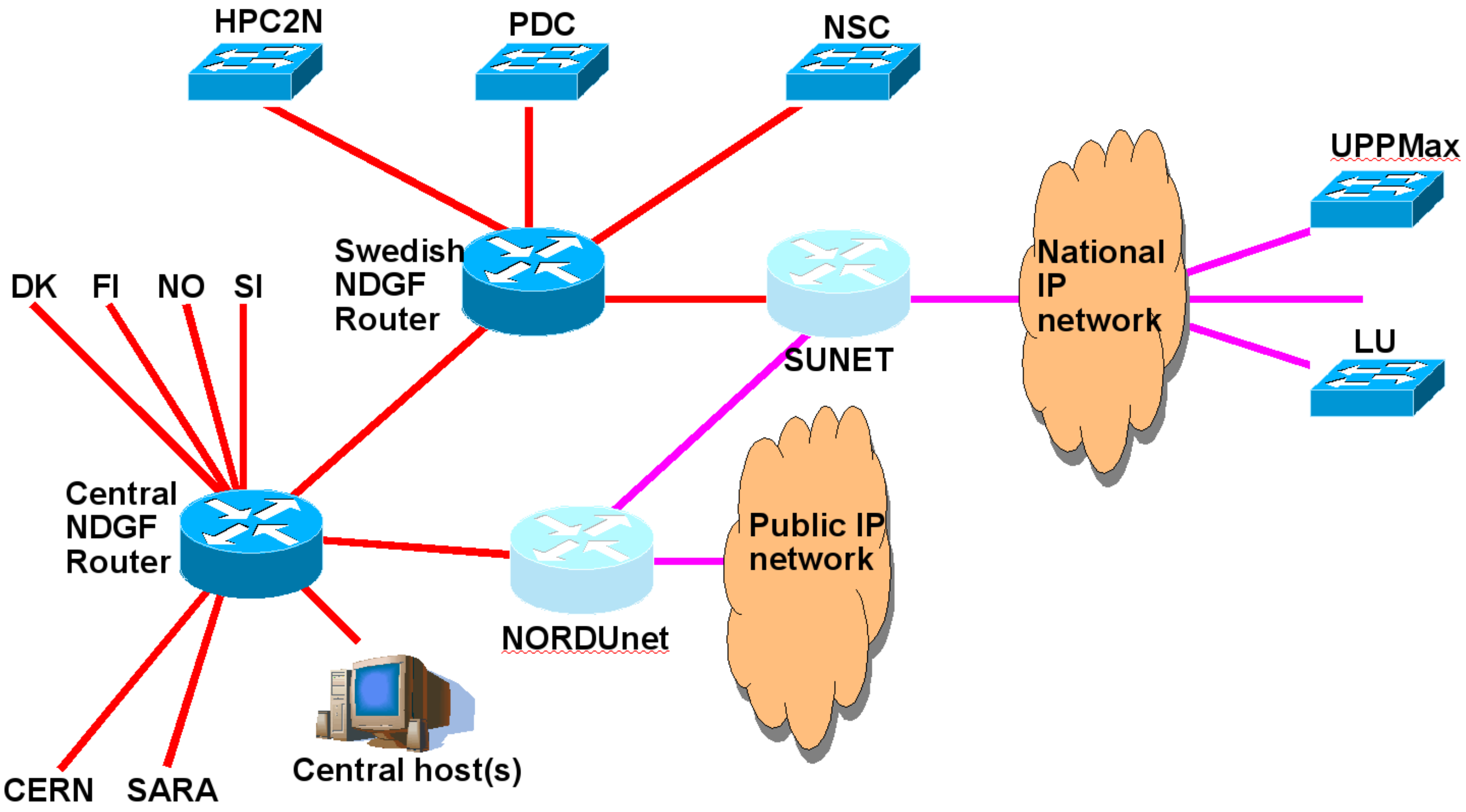


Distributed dCache

- dCache Installation
- dCache head nodes at Nordic GEANT endpoint
- Pools at sites
- In this setting – Not much different from having storage at a single site.

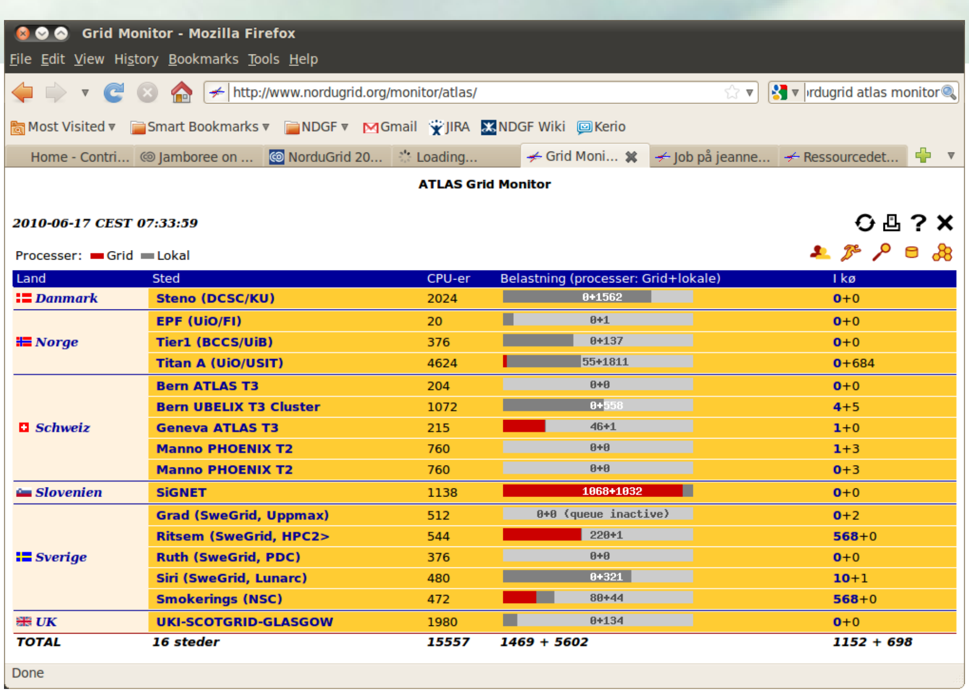


NDGF setup – the network perspective



NDGF Compute model

- Data management is handled by the grid CE.
- Jobs access files on local file system (NFS/GPFS/LUSTRE).
- The CE has the ability to cache files.


Grid Monitor - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.nordugrid.org/monitor/atlas/
ırdugrid atlas monitor
Most Visited ▾ Smart Bookmarks ▾ NDGF ▾ Gmail JIRA NDGF Wiki Kerio
Home - Contri... Jamboree on ... NorduGrid 20... Loading... Grid Moni... Job på jeanne... Ressourcedet...
Done

ATLAS Grid Monitor

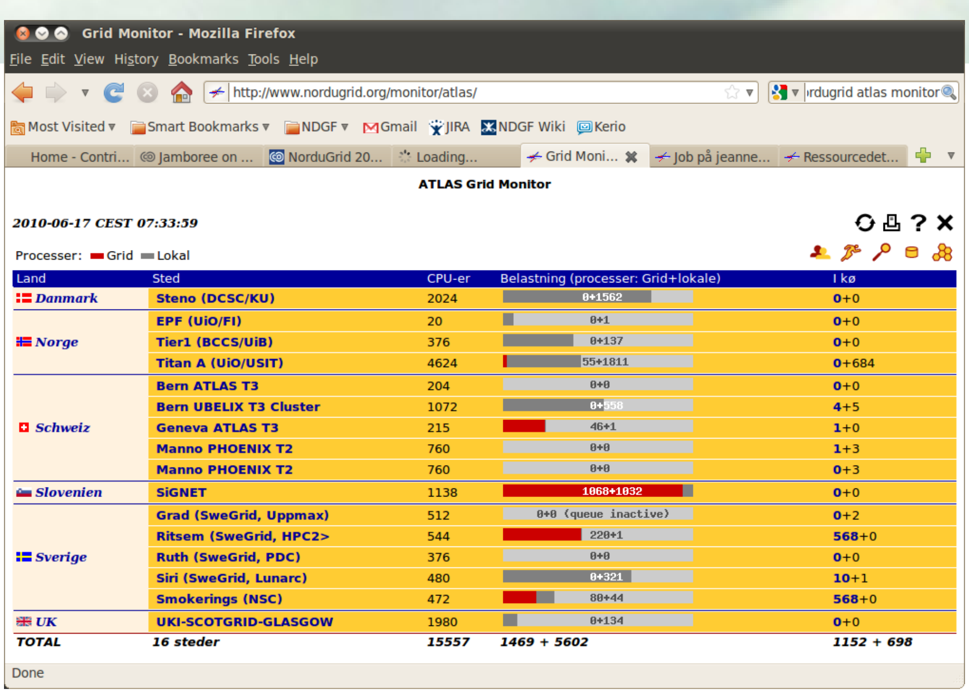
2010-06-17 CEST 07:33:59

Processer: 🟥 Grid ▬ Lokal

Land	Sted	CPU-er	Belastning (processer: Grid+lokale)	I kø
Danmark	Steno (DCSC/KU)	2024	0+1562	0+0
Norge	EPF (UiO/FI)	20	0+1	0+0
	Tier1 (BCCS/UiB)	376	0+137	0+0
	Titan A (UiO/USIT)	4624	55+1811	0+684
Schweiz	Bern ATLAS T3	204	0+0	0+0
	Bern UBELIX T3 Cluster	1072	0+558	4+5
	Geneva ATLAS T3	215	46+1	1+0
	Manno PHOENIX T2	760	0+0	1+3
	Manno PHOENIX T2	760	0+0	0+3
Slovenien	SiGNET	1138	1068+1032	0+0
Sverige	Grad (SweGrid, Uppmax)	512	0+0 (queue inactive)	0+2
	Ritsem (SweGrid, HPC2>	544	220+1	568+0
	Ruth (SweGrid, PDC)	376	0+0	0+0
	Siri (SweGrid, Lunarc)	480	0+321	10+1
	Smokerings (NSC)	472	80+44	568+0
UK	UKI-SCOTGRID-GLASGOW	1980	0+134	0+0
TOTAL	***16 steder***	***15557***	***1469 + 5602***	***1152 + 698***

NDGF Compute model

- Data management is handled by the grid CE.
- Jobs access files on local file system (NFS/GPFS/LUSTRE).
- The CE has the ability to cache files.
- (Almost) all Tier-2 storage is used for cache (approx 1 PB).
- 100+ TB per site.



ATLAS Grid Monitor

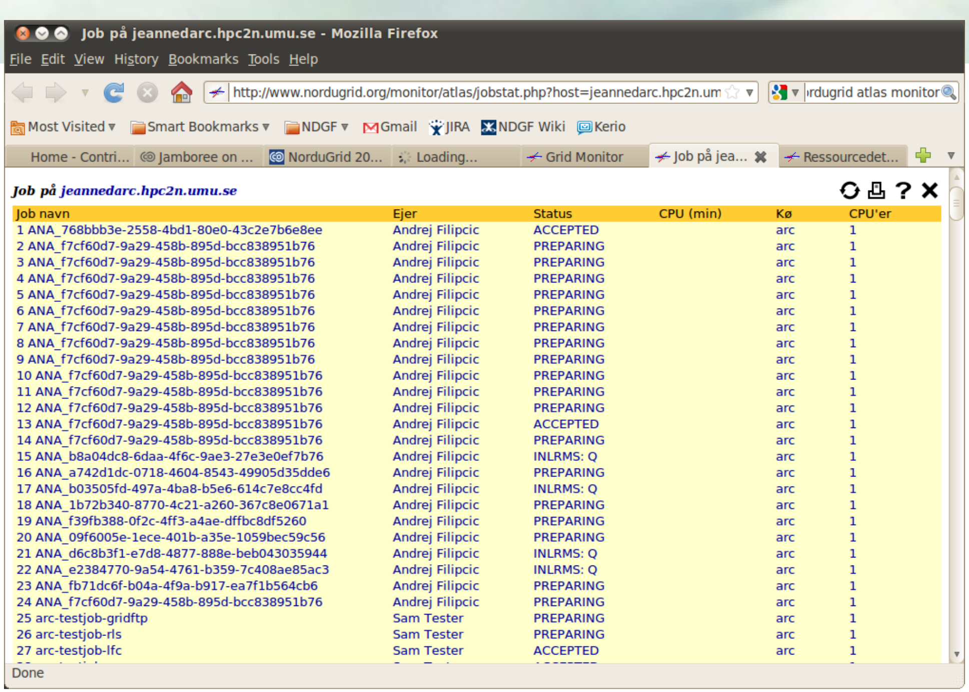
2010-06-17 CEST 07:33:59

Processer: 🟥 Grid ▬ Lokal

Land	Sted	CPU-er	Belastning (processer: Grid+lokale)	I kø
Danmark	**Steno (DCSC/KU)**	2024	0+1562	0+0
Norge	**EPF (UiO/FI)**	20	0+1	0+0
	Tier1 (BCCS/UiB)	376	0+137	0+0
	Titan A (UiO/USIT)	4624	55+1811	0+684
Schweiz	**Bern ATLAS T3**	204	0+0	0+0
	Bern UBELIX T3 Cluster	1072	0+558	4+5
	Geneva ATLAS T3	215	46+1	1+0
	Manno PHOENIX T2	760	0+0	1+3
	Manno PHOENIX T2	760	0+0	0+3
Slovenien	**SiGNET**	1138	1068+1032	0+0
Sverige	**Grad (SweGrid, Uppmax)**	512	0+0 (queue inactive)	0+2
	Ritsem (SweGrid, HPC2>	544	220+1	568+0
	Ruth (SweGrid, PDC)	376	0+0	0+0
	Siri (SweGrid, Lunarc)	480	0+321	10+1
	Smokerings (NSC)	472	80+44	568+0
UK	**UKI-SCOTGRID-GLASGOW**	1980	0+134	0+0
TOTAL	***16 steder***	***15557***	***1469 + 5602***	***1152 + 698***

NDGF Compute model

- Data management is handled by the grid CE.
- Jobs are not submitted to lrms before all input files are available.

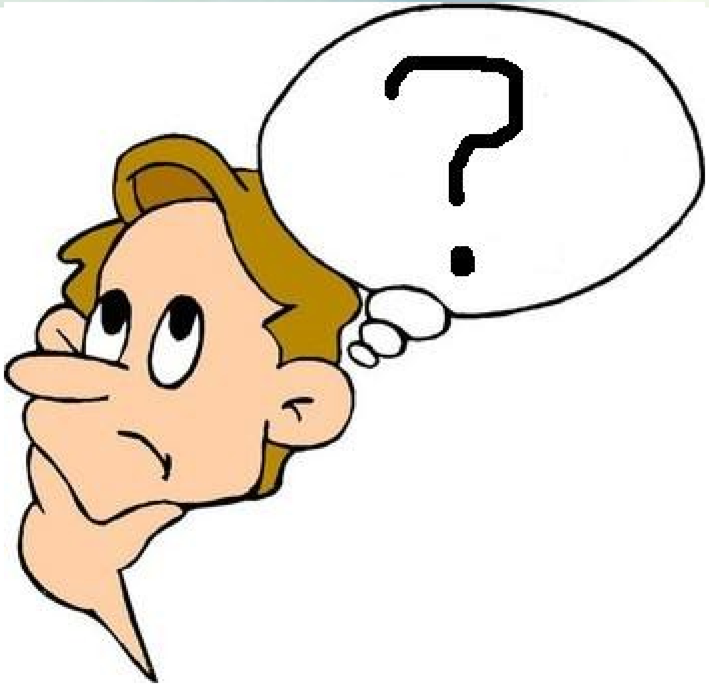


Job på jeannedarc.hpc2n.umu.se

Job navn	Ejer	Status	CPU (min)	Kø	CPU'er
1 ANA_768bbb3e-2558-4bd1-80e0-43c2e7b6e8ee	Andrej Filipcic	ACCEPTED		arc	1
2 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
3 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
4 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
5 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
6 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
7 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
8 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
9 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
10 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
11 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
12 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
13 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	ACCEPTED		arc	1
14 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
15 ANA_b8a04dc8-6daa-4f6c-9ae3-27e3e0ef7b76	Andrej Filipcic	INLRMS: Q		arc	1
16 ANA_a742d1dc-0718-4604-8543-49905d35dde6	Andrej Filipcic	PREPARING		arc	1
17 ANA_b03505fd-497a-4ba8-b5e6-614c7e8cc4fd	Andrej Filipcic	INLRMS: Q		arc	1
18 ANA_1b72b340-8770-4c21-a260-367c8e0671a1	Andrej Filipcic	PREPARING		arc	1
19 ANA_f39fb388-0f2c-4ff3-a4ae-dffbc8df5260	Andrej Filipcic	PREPARING		arc	1
20 ANA_09f6005e-1ece-401b-a35e-1059bec59c56	Andrej Filipcic	PREPARING		arc	1
21 ANA_d6c8b3f1-e7d8-4877-888e-beb043035944	Andrej Filipcic	INLRMS: Q		arc	1
22 ANA_e2384770-9a54-4761-b359-7c408ae85ac3	Andrej Filipcic	INLRMS: Q		arc	1
23 ANA_fb71dc6f-b04a-4f9a-b917-ea7f1b564cb6	Andrej Filipcic	PREPARING		arc	1
24 ANA_f7cf60d7-9a29-458b-895d-bcc838951b76	Andrej Filipcic	PREPARING		arc	1
25 arc-testjob-gridftp	Sam Tester	PREPARING		arc	1
26 arc-testjob-rls	Sam Tester	PREPARING		arc	1
27 arc-testjob-lfc	Sam Tester	ACCEPTED		arc	1

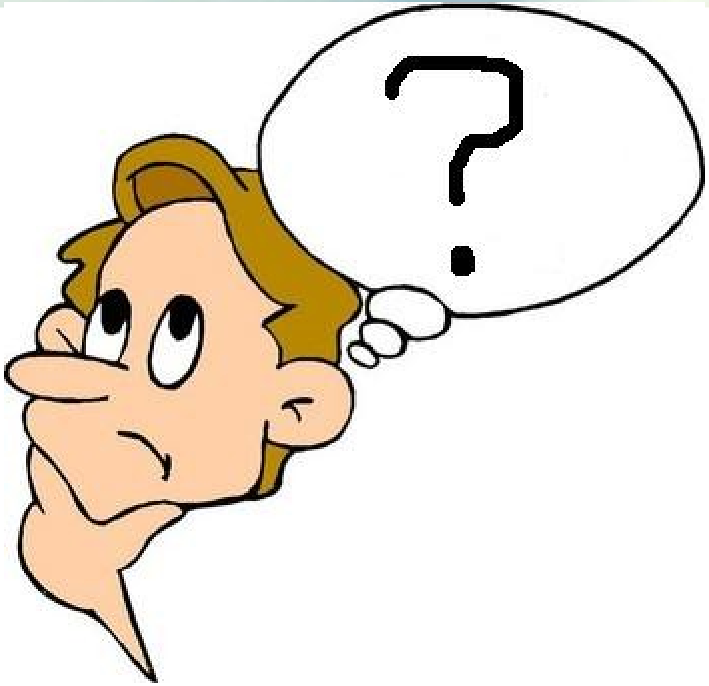
Where is a file cached?

- Without any knowledge the system will not perform...



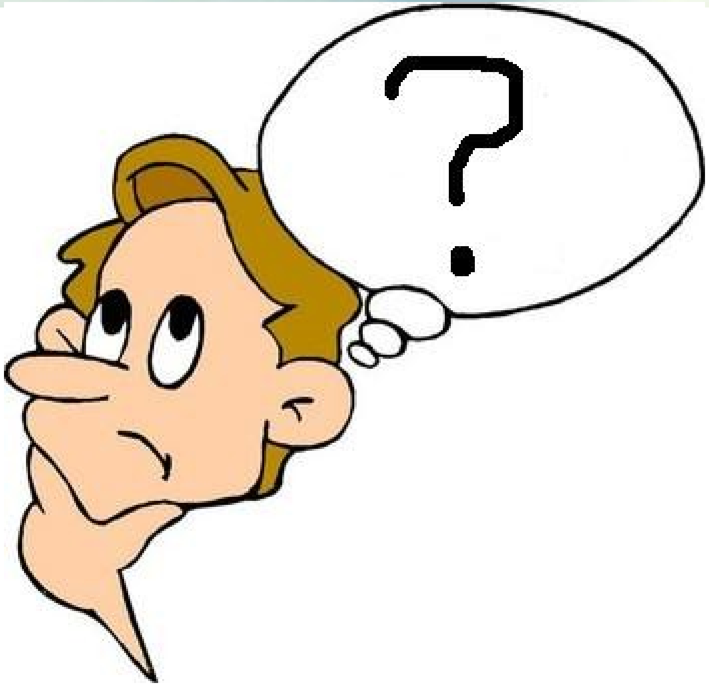
Where is a file cached?

- Without any knowledge the system will not perform...
- 20K analysis jobs consumes approx 400TB of data.
- But only 20-40TB of unique files.



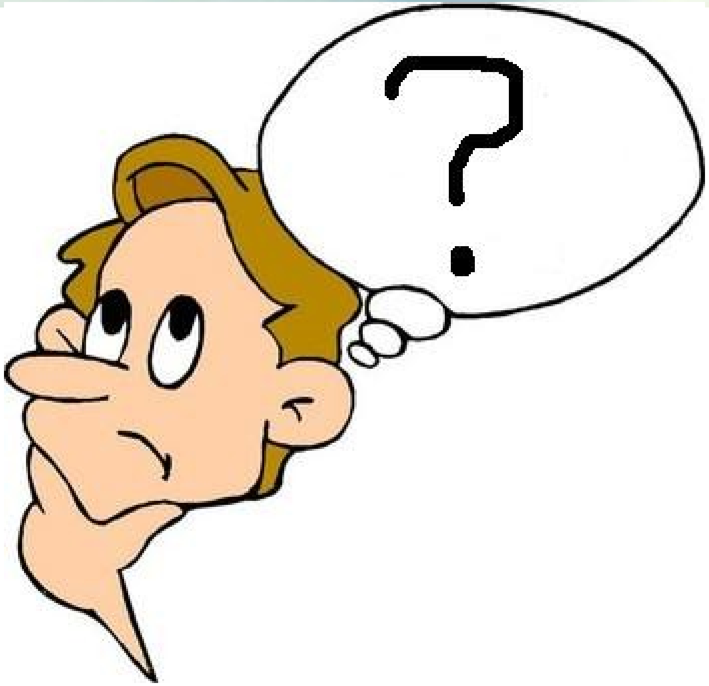
Where is a file cached?

- Without any knowledge the system will not perform...
- 20K analysis jobs consumes approx 400TB of data.
- But only 20-40TB of unique files.
- Without knowledge this took 4 days.



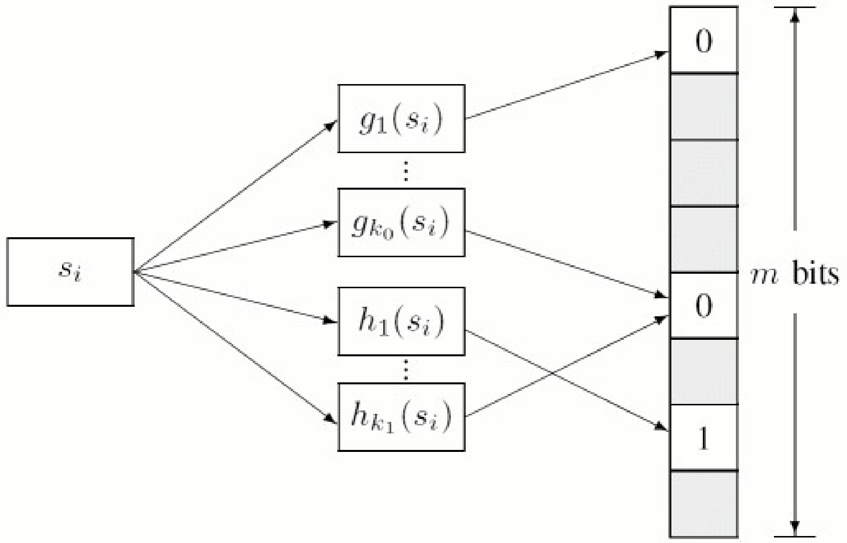
Where is a file cached?

- Without any knowledge the system will not perform...
- 20K analysis jobs consumes approx 400TB of data.
- But only 20-40TB of unique files.
- Without knowledge this took 4 days.
- With the *cache index* only 10h.



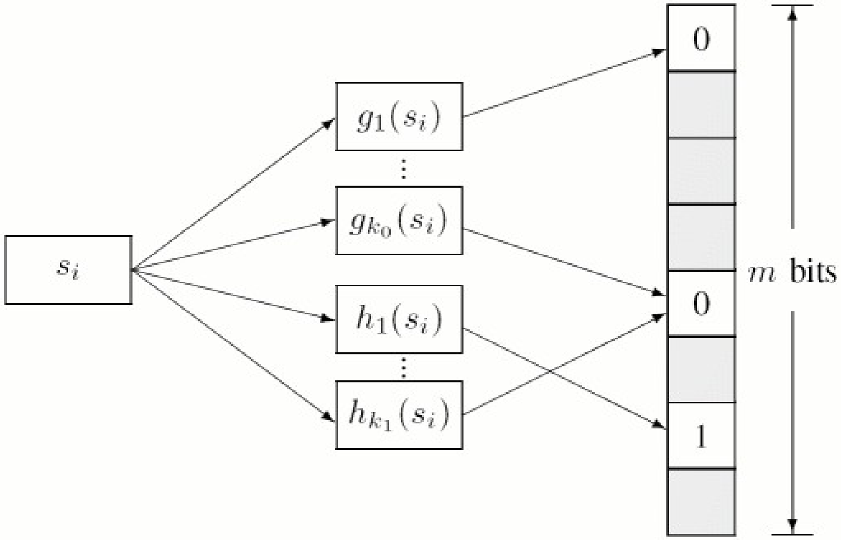
The Cache Index

- We do not need perfect knowledge.
- We do need fast lookup.
- We do need to be able to cope with failure (fast reconstruction of the index).



The Cache Index

- We do not need perfect knowledge.
- We do need fast lookup.
- We do need to be able to cope with failure (fast reconstruction of the index).
- Bloom filters gives us this,
  - incremental update possible,
  - but with some probability of false positives.



Data management at CE - Advantages

- Transfers can be scheduled and prioritized.
- Bandwidth usage can be throttled (e.g. user/VO limits).
- Cache management (LRA discard).
- Somewhere between *d* and *e* in Philippe's list of models.

- Preparation of cache possible.
- Cache2cache transfers can be implemented
  - authorization is the tricky bit.

Data management at CE - Disadvantages

- Pilot jobs does not play well with CE handling data mgmt.
- ARC Ctrl Tower acts as pseudo pilot.
- In most cases only a few minutes delay between job submission before execution starts.

Thank you!

Using the index

- Daemon rebuilds index at site and provides it on request to the index.
- Query is simple http request

  -
    ```
    > curl -k  "https://cacheindex.ndgf.org:6443/data/index?
    url=http://www.nordugrid.org:80/data/echo.sh"
          {"http:\/\/www.nordugrid.org:80\/data\/echo.sh":
          ["benedict.grid.aau.dk"]}
    ```
- Removal of entries difficult → rebuild on a regular basis.
- Basically independent of ARC CE.

Analysis jobs May

titan	10167
pikolit	157862
grad	38508 (17k in one week)
ritsem	35984
pdc	6108
siri	24085
smokerings	3465
swiss	11361
Total	**287504**