

GridPP
UK Computing for Particle Physics



ATLAS: Resilience and Disaster Plans

Graeme Stewart

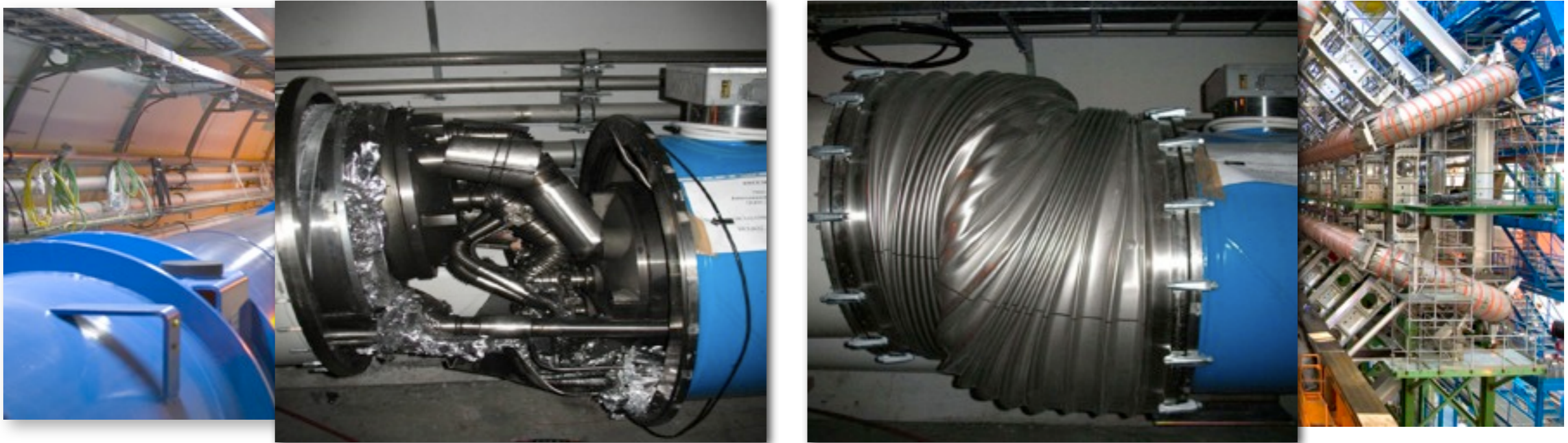
(with thanks to Roger and Peter in particular)

Overview

- From LHC to Online systems
- Offline systems and Grid
 - Central Services
 - Tier-1
 - Tier-2
- Questions
- Conclusions



The Coal Pile in the Ballroom...

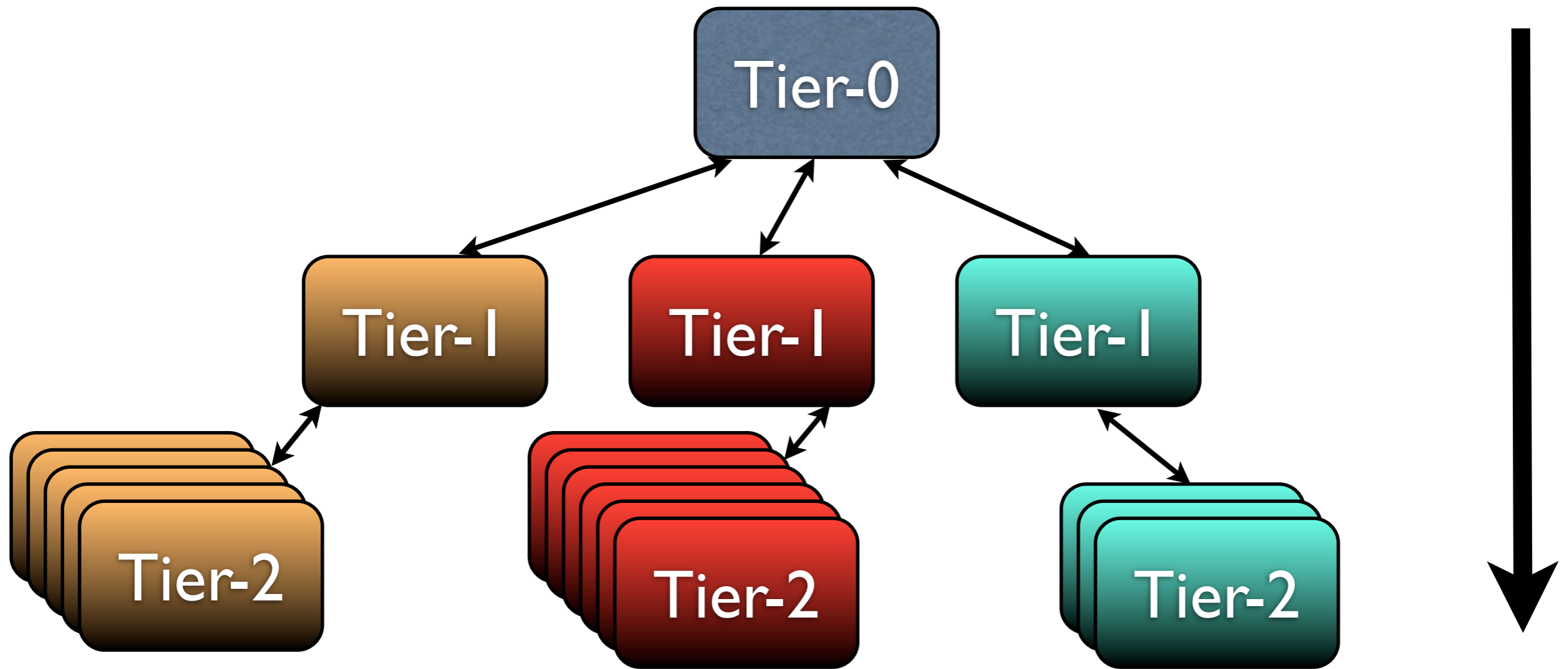


2 most severely damaged interconnects

- There's only one LHC
- There's only one ATLAS detector
- Obviously these systems have their own fail-safes, backups, redundancy
- and we all hope they work



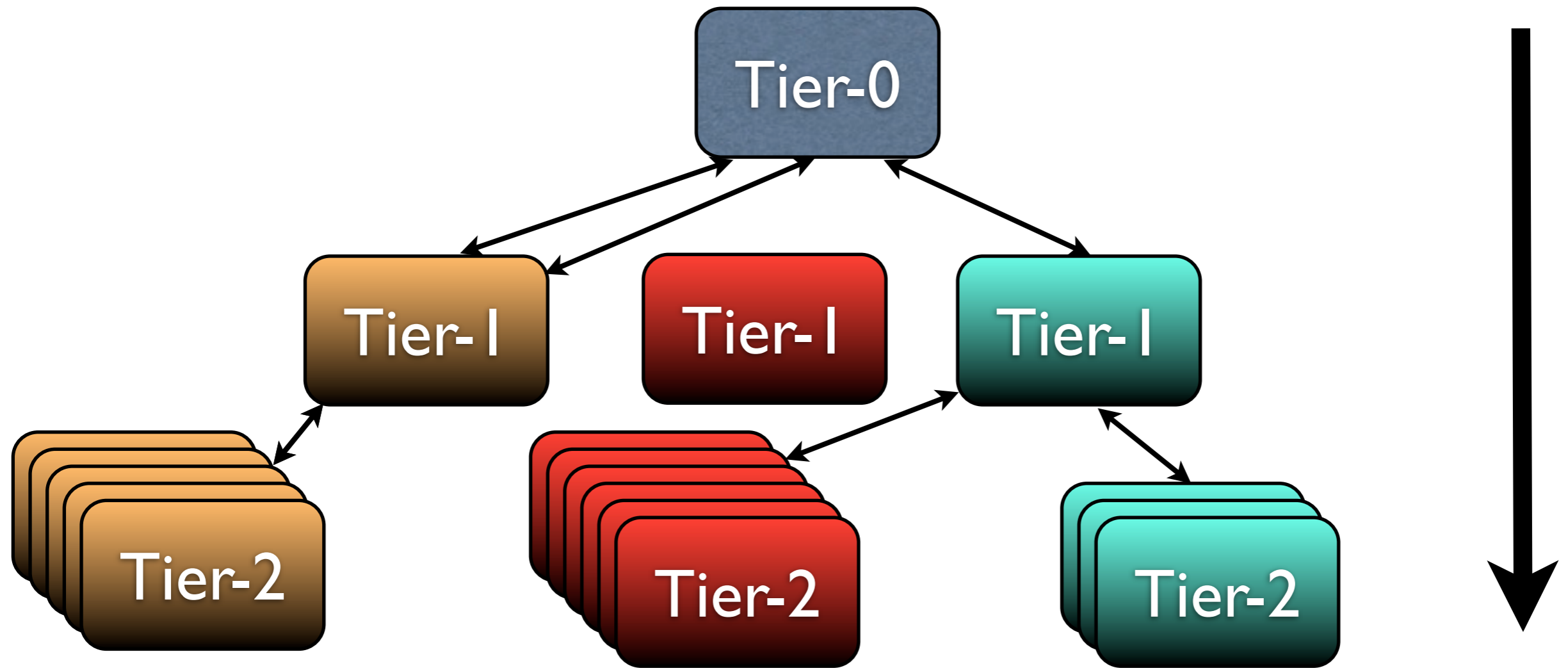
Offline Systems



- Redundancy naturally increases from $T0 \rightarrow T1 \rightarrow T2$



Offline Systems

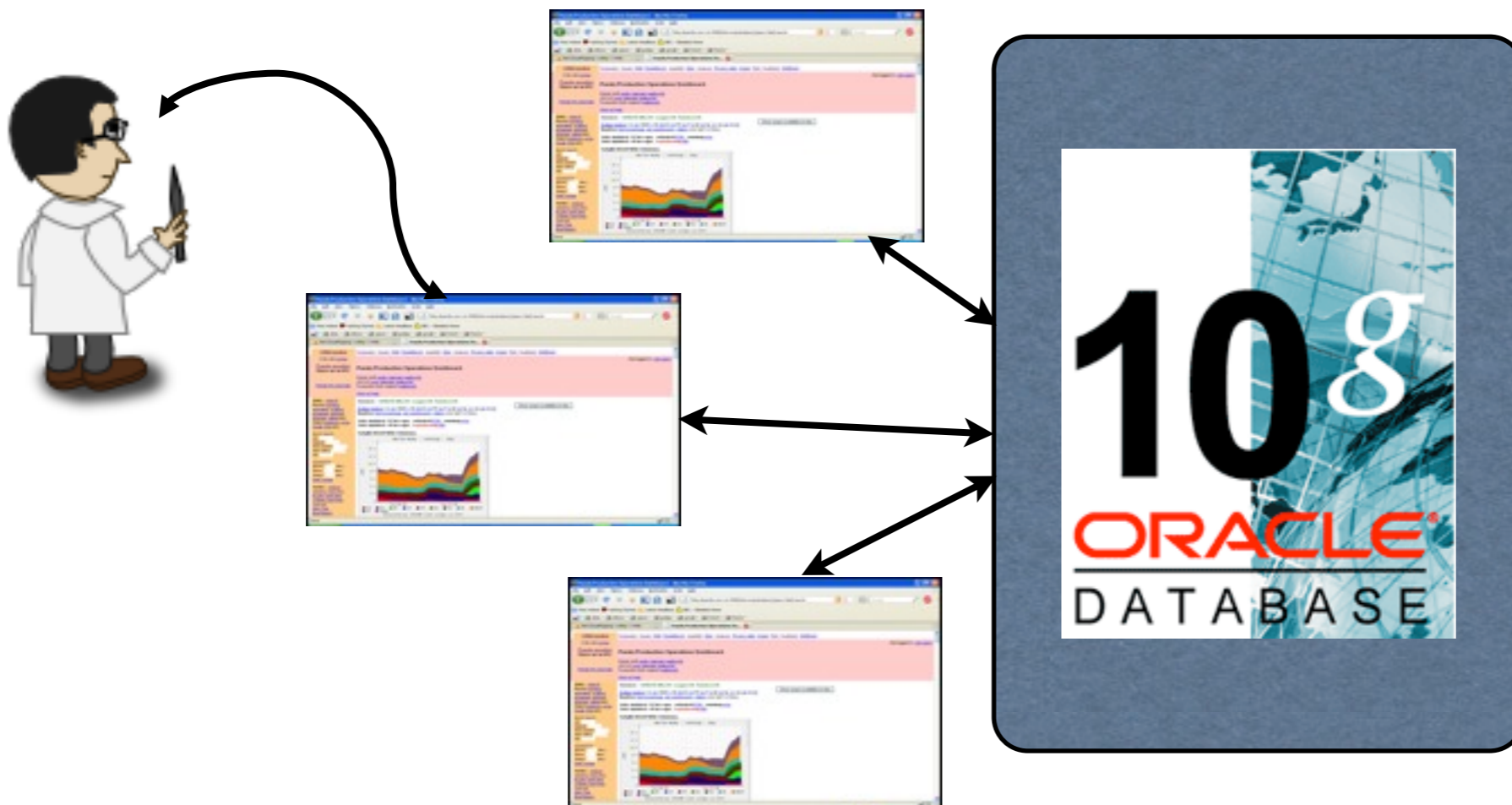


- It's easy to conceptualise reconfiguring the system
 - but can we do it?
 - and what does it involve?



Central Service Design

- Most ATLAS central services are designed around a highly available database back-end with stateless web front ends



- For the ATLAS offline production database (ATLR) Oracle Data Guard is used

Semi-Stateless Services

- Certain central services are ‘semi-stateful’
 - e.g. DDM site services
 - State of subscriptions is monitored, but only while the subscription is active
- Loss of these services provokes a higher level of errors, but no catastrophes
- Resilience strategy here is to have hot spares with a known installation procedure
- This installation procedure is exercised, e.g., during upgrades



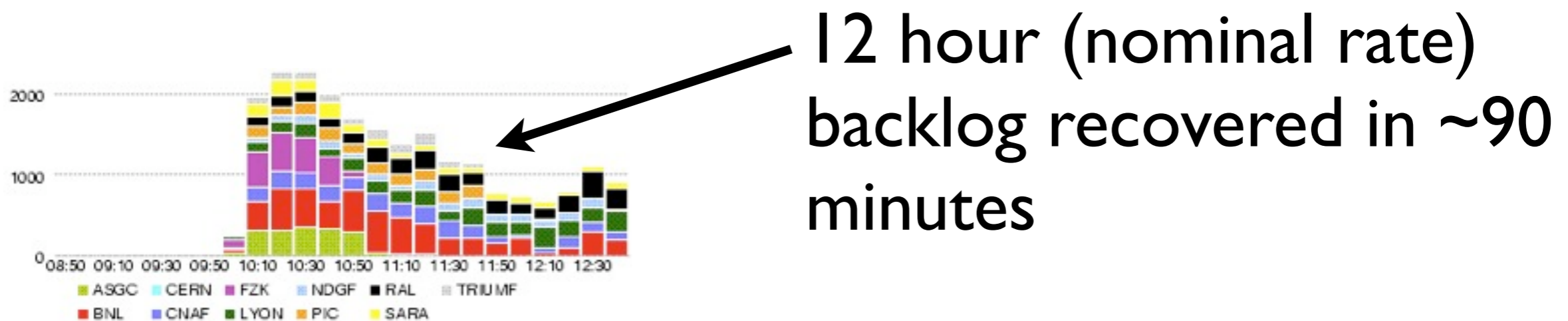
Tier-1 Services

- A reminder of critical T1 functionality
 - Accept RAW data from T0 - custodial
 - Reprocess RAW data
 - Provide services to T2 cloud (LFC, FTS)
 - Data for T2s moves via T1
 - Collect simulation output from T2s



TI Storage Down

- Workflow: data to and through the TI
- If there is a short outage of TI storage
 - DDM will automatically retry and recover
 - Scale tests prove that network, DDM and TI infrastructure can handle such backlogs





T I Storage Down

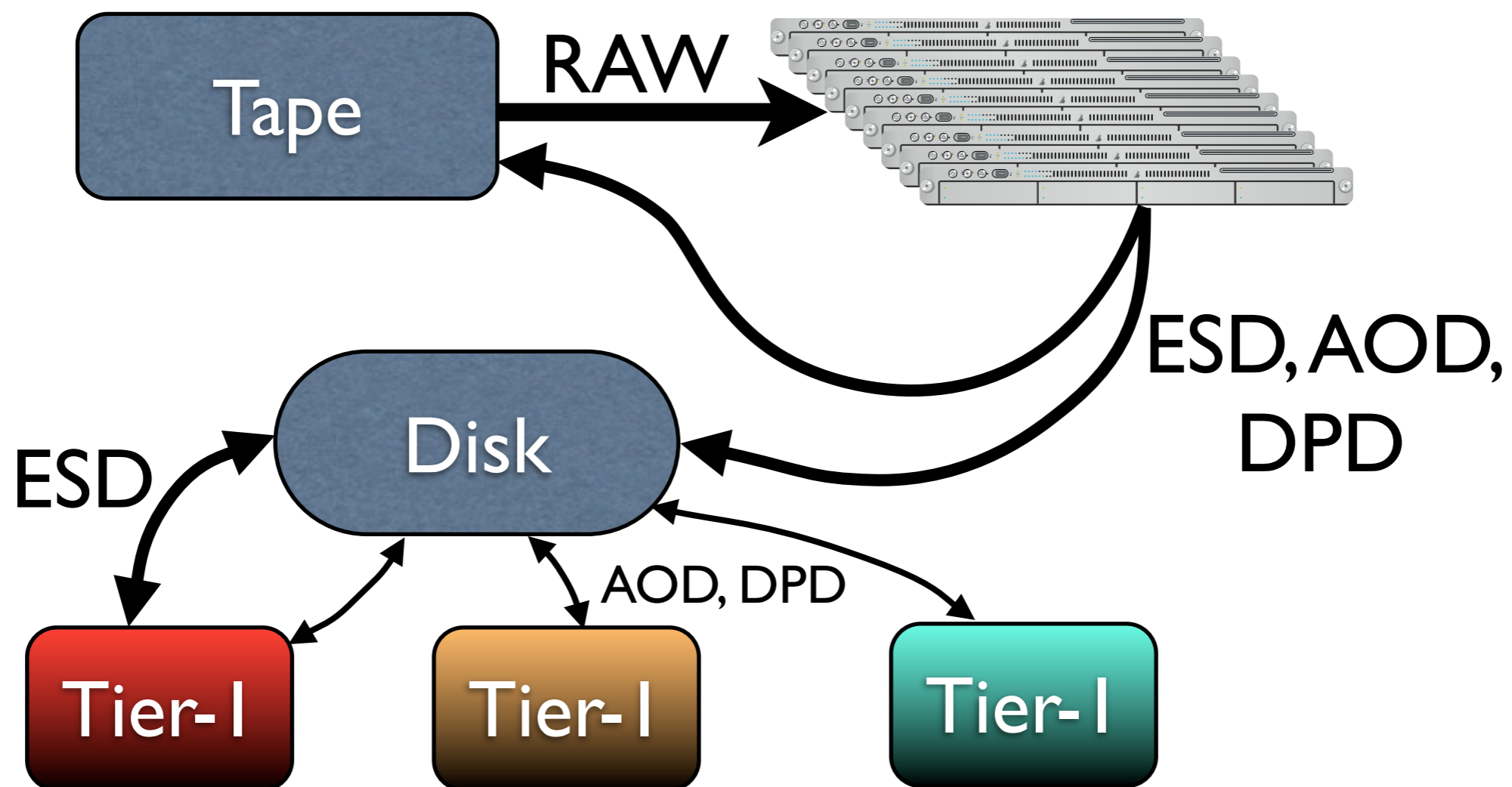


- If the downtime is longer
 - T0 has a buffer size of ~4 days
 - T1 share is redistributed across the other TIs (Santa Claus)
 - Subscribed data-sets are cancelled and re-subscribed to other TIs (Pedro el Negro)
 - AOD and DPD replication to T2s stops
 - At least for a while...

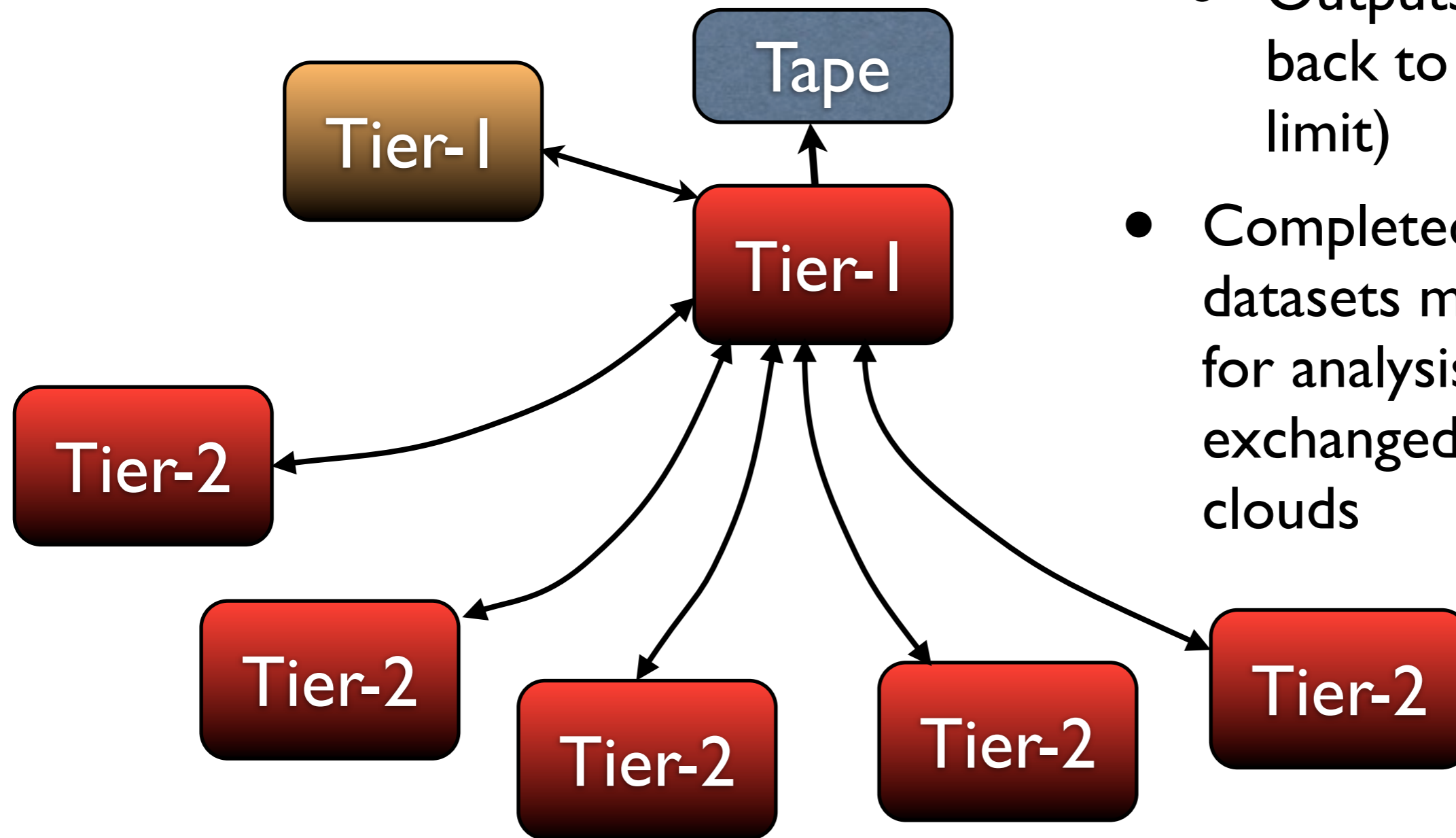


Tier I Reprocessing Workflow

- TI has the only copy of RAW
 - CERN has a backup copy, but limited 'TI' capacity



Tier I Simulation Workflows



- Simulation done at T2s
- Outputs moved back to T1 (4 day limit)
- Completed AOD datasets moved to T2s for analysis and exchanged with other clouds



ATLAS Analysis Workflows

- Group analysis is supported directly at the Tier-1
- Loss of resource will affect ability of physics and performance groups to work
 - Particularly if datasets were unique at the Tier-1 - so this will have an uneven effect
- User and group analysis is supported by Tier-2s
 - Loss of Tier-1 service will affect all T2s
 - Loss of Tier-2 will be more limited in impact
 - But may affect, e.g., completeness of AOD within the cloud



FTS



- FTS is the data movement plumbing between sites
- Loss of FTS at Tier-I means that data cannot be moved to Tier-2s
- Data may generally be pulled from T2s using star channels
- It is possible, *in extremis*, to configure a foreign FTS to serve another cloud's T2s
- The static nature of FTS channels makes this inflexible
- FTS is 'semi-stateful' so a reinstall is not too painful
- *FTS is a critical T1 service for the T2 cloud*
- *Analysis can continue locally to site, but output cannot move*



LFC

- *The LFC is the most important critical cloud service*
- Without it T2s are dead, because files cannot be resolved to SURLs
 - Neither can new SURLs be registered
 - Some efforts have been made to provide back-up and resilience advice
 - At the moment these focus on internal availability:
 - Multiple front ends
 - Resilient Oracle back-end
 - CNAF have investigated Data Guard but this is difficult and expensive

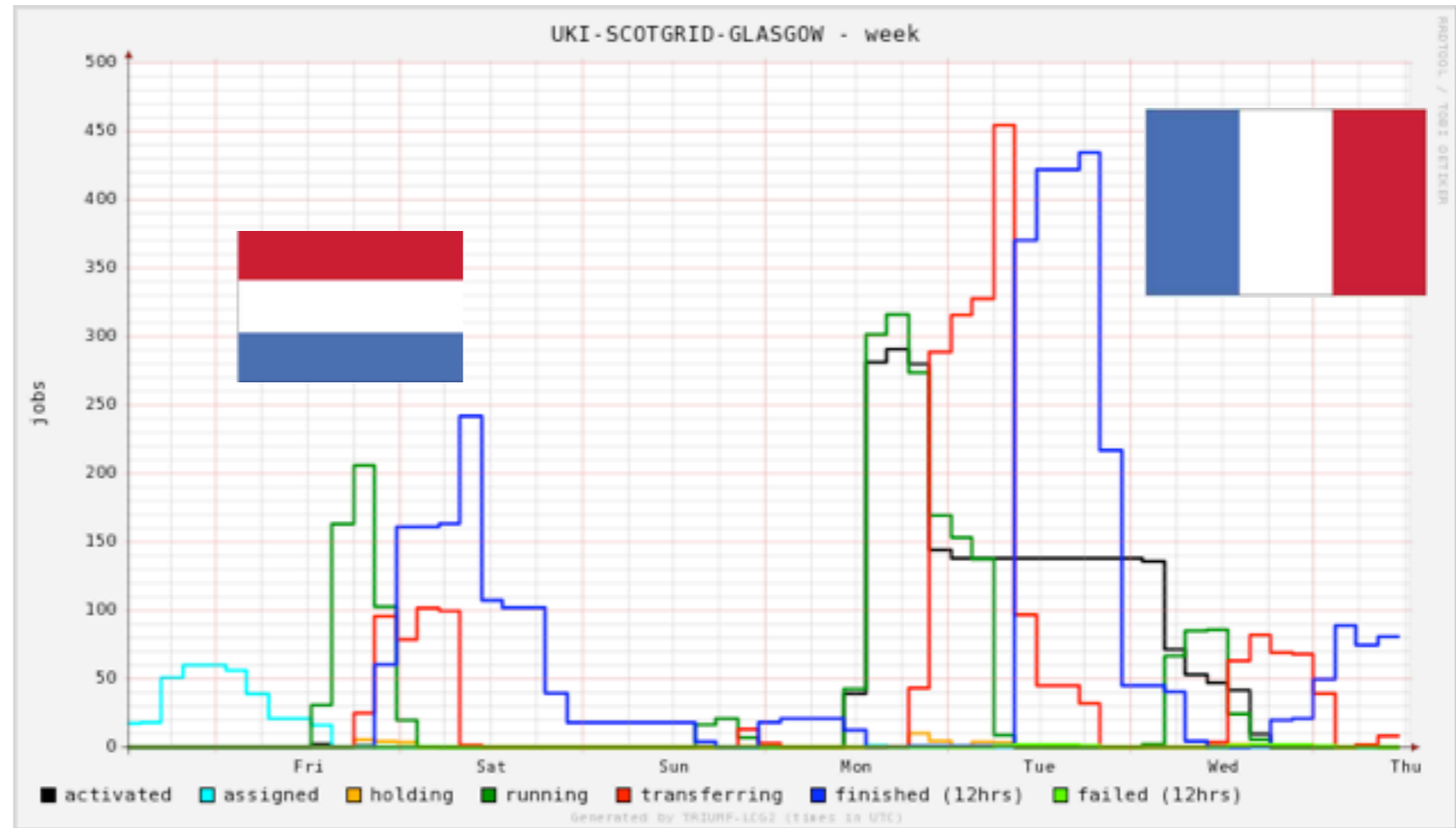


Should I file it or
do you want to
find it again?



TI Down (continued)

- During problems with RAL CASTOR 2.1.7 upgrade we tested moving a T2 for simulation production into another cloud
- Expanding this 'test' to data distribution from TRIUMF → Melbourne



Tier-2 Sites

- Tier-2 sites do not have unique copies of any collaboration data
- Unavailability is manageable, across the grid
- However, they may have unique copies of user or group data
 - So the impact is not uniform
 - If the downtime is very long then jobs should be re-run
 - We plan to have some geographic redundancy for groups



2009 Production Summary

site	success	failure	success (walltime)	failure (walltime)	efficiency	efficiency (walltime)
RAL-LCG2	368530	92202	4193093601	1176694107	80%	78.1%
UKI-SCOTGRID-GLASGOW	116058	16855	2756814276	74604805	87.3%	97.4%
UKI-NORTHGRID-MAN-HEP	26314	64951	1286172166	18619931	28.8%	98.6%
UKI-LT2-QMUL	50930	4942	1477293083	58678607	91.2%	96.2%
UKI-LT2-RHUL	33478	1992	815234554	35228206	94.4%	95.9%
UKI-NORTHGRID-LANCS-HEP	26600	6494	728881952	54726802	80.4%	93%
UKI-NORTHGRID-LIV-HEP	23697	2428	920979900	41298009	90.7%	95.7%
UKI-NORTHGRID-SHEF-HEP	18492	777	406419292	7603613	96%	98.2%
UKI-SCOTGRID-ECDF	16919	1919	236323447	12205352	89.8%	95.1%
UKI-LT2-IC-HEP	13740	2076	346725033	30329772	86.9%	92%
UKI-SOUTHGRID-CAM-HEP	14278	1015	287185345	10377608	93.4%	96.5%
UKI-SOUTHGRID-OX-HEP	11255	3481	256478843	46340174	76.4%	84.7%
UKI-SOUTHGRID-RALPP	10733	678	292750434	10756710	94.1%	96.5%
UKI-SCOTGRID-DURHAM	4978	200	39746251	121970	96.1%	99.7%
UKI-LT2-Brunel	3110	383	100145894	3245722	89%	96.9%
UKI-SOUTHGRID-BHAM-HEP	737	23	453517	172312	97%	72.5%
UKI-LT2-UCL-HEP	0	205	0	0	0%	-
<i>total</i>	<i>739849</i>	<i>200621</i>	<i>1.4144697588e+10</i>	<i>1581003700</i>	<i>78.7%</i>	<i>89.9%</i>

- Caveat emptor...



Dave's Scenarios of Doom...

- Specifically answering questions raised by GridPP PMB...



Action Plan for Major T1 Outage

1. Restore LFC and FTS services
 - Then prioritise CASTOR recovery
2. Move T2s to other cloud(s) for data distribution
3. Move T2s to other clouds(s) for simulation production (if necessary)



Data Loss Strategy

- Please clean your SE of dead SURLs
- Copy from elsewhere:
 - Specifically CERN for RAW
 - Priority depends on experiment phase
- Regenerate:
 - For derived data
- Delete:
 - If unrecoverable



T1 Resource Loss Strategy

- We reduce the scale of the T1 to fit the new resources
- Some reprocessing can be moved to CERN, but share may have to go to other clouds
 - You don't need to worry about fairshare - it's under our control
 - Requires over-provision of bandwidth to T1s
- Some T2s may have to get data delivered from other T1s and move their simulation capacity to other clouds
- But it may be someone else's disaster...
 - Increased RAW share capacity
 - Ability to adopt other T2s



Communications



- Email:
 - ATLAS Operations Lists, UK Operations Lists
- Jabber:
 - Most of us have it
- Skype:
 - We're mostly around
- Phone:
 - Yes, it still exists
- People:
 - Need to identify human points of contact and responsible people on ATLAS side



Conclusions

- TI unavailability has been well exercised
 - Unfortunately, this was not usually a drill
- We also have experience of most data loss and service outage events
- Communication is very important
- As well as flexibility to tailor the response to exactly the problem we face

