Tier1 Reliability

John Gordon, STFC-RAL

CCRC09 Planning @CERN

November 13th, 2008

Science & Technology Facilities Council

The Problem

- The MoU commitments do not permit many long breaks before availability drops below the required level.
- The LHC experiment computing models are sensitive to breaks in service at any Tier1.
- Anecdotally the number of T1 breaks in service is felt to be too high and has not dropped recently.
- Recovering breaks in service soak up staff effort both for sites and experiments

Types of Problem

- My questions
- Few answers from T1. Summarise
- How to improve?
- Buy better hardware?
- Redundancy
- Best Practice

My Questions 1

- In 2008, what has been your experience of different types of serious incident (eg >0.5 day down).

  1. **Cat**astrophic failures which affected all your services, eg power failure, air-con failure,
  2. **Har**dware failures (disk crash, cpu died) which resulted in loss of service (ie no failover)
  3. **Mid**dleware failure - where the service failed and needed non-trivial manual intervention to bring it back to service.

My Questions 1

	ASGC	BNL	CNAF	FZK	FNAL	IN2P3	NDGF	NL	PIC	RAL	TRIUMF
1 Cat				1	0	1				1	
2 Har				1	0	0				3	
3 Mid				RS	few	1				8	

FZK – storage slowdown, no middleware breaks
FNAL – FCC has generator, GCC vulnerable but nothing this year.
FNAL – local hardware problems but not for CMS
FNAL – Phedex (now improved) and FTS
IN2P3 – mware services failed due to Oracle patch
RAL – 2 double disk RAID failures
RAL – a variety of different Castor issues

Science & Technology Facilities Council

My Questions 2

1. Which services do you believe that you have hardened.
   - Ie redundancy, failover, UPS, whatever is relevant.
2. Have you identified any services which you plan to harden over the winter?
3. Have you identified any services which you cannot see how to harden sufficiently?

My Questions 2

	ASGC	BNL	CNAF	FZK	FNAL	IN2P3	NDGF	NL	PIC	RAL	TRIUMF
Done				X	X	X					
Winter				X	X	X					
Cannot				X	X	X					

Science & Technology Facilities Council

My Questions 2

	ASGC	BNL	CNAF	FZK	FNAL	IN2P3	NDGF	NL	PIC	RAL	TRIUMF
Done				CE, FTS, WMS	Monitoring	UPS, CE,LFC, WMS, dcache				CE, FTS, WMS, SRM	
Winter				-	Power sep,db UPS pnfs SSD+ new db	FTS				Castor and Oracle	
Cannot				SRM	-	Dcache core nodes, LFC					
Other				BDII, sBDII, LFC							

Best Practice

- Redundancy
  - not all services benefit
  - Independent or round robin
- UPS
- Mirroring system disks,
  - & isolating system disks from service
- Well documented recovery procedures
  - So that anyone called in can restart or replace a service
  - Tested
  - For individual services and full power cuts
- Capacity Planning
  - Plan to cope with the planned load plus a safety margin, not the load you see
  - But what is the planned load?

But…

- Are all sites doing all of these?

Other Issues

- Middleware
- On call not mandatory, cannot work all night.
- Often need many experts
- Reduced capacity
  - Running on half total load is usually simple, reduce batch work
  - But what if one transformer went? Are there instances of critical services on another?

Outcomes?

- Sharing best practice
- Workshops
- Documentation
- Review each other
- Top priority middleware improvements
  - Bug fixes