Grid-ability

Lessons from deploying WLCG production services

¿Input to Future Grids?

Agenda

3 main “abilities” required for large-scale production Grids

- **Reliability**
- Scalability
- Accountability

- Interoperability is also key...

Background

- 100% of my **Grid** experience relates to the deployment and delivery of **Production Services**
- This started already in the days of EDG with the Replica Location Service and its deployment at CERN and some key (WLCG) Tier1 sites
- In the then-current WLCG Computing Model, the RLS was a critical component which, if unavailable, meant:

➢ **Running jobs could not access existing data**
➢ **Scheduling of jobs at sites where the needed data was located was not possible**

☹ **The Grid – if not down – was at least seriously impaired…**

- **This was taken into account when designing the service deployment strategy & procedures – a taste of things to come!**

WLCG Service Challenges

- **Since January 2005, involved in the WLCG Service Challenge programme**
- ➢ **Get the essential grid services ramped up to target levels of reliability, availability, scalability, end-to-end performance**
- These Challenges, which completed in October 2006, resulted in a **"usable, but not perfect"** service
- With small enhancements & extensions, this is the service that will be used for the Cosmics Runs & FDRs in 2007 and the Engineering & Physics Runs of 2008...
- 💣 **But the service is still very costly to operate in terms of manpower – will this scale to full production?**

Reliability

Some targets for reliability and real life experience in implementing them

Service Availability Targets

- The WLCG Memorandum of Understanding defines:
  - **The services that a given site must provide (Tier0, Tier1, Tier2);**
  - **The availability of these services (measured on an annual basis);**
  - **The maximum time to intervene in case of problems.**

- Taken together, these service availability targets are somewhat aggressive and range from 95% to 99% for **compound** services, e.g.
  - **Acceptance of raw data from Tier0**
  - **Data-intensive analysis services, including networking to Tier0**

- Such 'services' involve many sub-services, e.g. storage services, catalog and metadata services, DB services, experiment-specific services etc.

- Major concerns include both **scheduled** and unscheduled interventions – must design all elements of the service correspondingly
  - **Hardware configuration; procedures & documentation; middleware**

WLCG Tier1 Services[1]

i. acceptance of an agreed share of raw data from the Tier0 Centre, keeping up with data acquisition;

ii. acceptance of an agreed share of first-pass reconstructed data from the Tier0 Centre;

iii. acceptance of processed and simulated data from other centres of the WLCG;

iv. recording and archival storage of the accepted share of raw data (distributed back-up);

v. recording and maintenance of processed and simulated data on permanent mass storage;

vi. provision of managed disk storage providing permanent and temporary data storage for files and databases;

vii. provision of access to the stored data by other centres of the WLCG and by named AF's as defined in paragraph X of this MoU;

viii. operation of a data-intensive analysis facility;

ix. provision of other services according to agreed Experiment requirements;

x. ensure high-capacity network bandwidth and services for data exchange with the Tier0 Centre, as part of an overall plan agreed amongst the Experiments, Tier1 and Tier0 Centres;

xi. ensure network bandwidth and services for data exchange with Tier1 and Tier2 Centres, as part of an overall plan agreed amongst the Experiments, Tier1 and Tier2 Centres;

xii. administration of databases required by Experiments at Tier1 Centres.

- All storage and computational services shall be "grid enabled" according to standards agreed between the LHC Experiments and the regional centres.

[1] WLCG Memorandum of Understanding (signed by each T0/T1/T2)

Problem Response Time and Availability targets Tier-1 Centres

Service	Maximum delay in responding to operational problems (hours)			Availability
	Service interruption	Degradation of the service		
		> 50%	> 20%	
Acceptance of data from the Tier-0 Centre during accelerator operation	12	12	24	99%
Other essential services – prime service hours	2	2	4	98%
Other essential services – outside prime service hours	24	48	48	97%

les.robertson@cern.ch

Problem Response Time and Availability targets Tier-2 Centres			
Service	*Maximum delay in responding to operational problems*		*availability*
	Prime time	*Other periods*	
End-user analysis facility	2 hours	72 hours	95%
Other services	12 hours	72 hours	95%

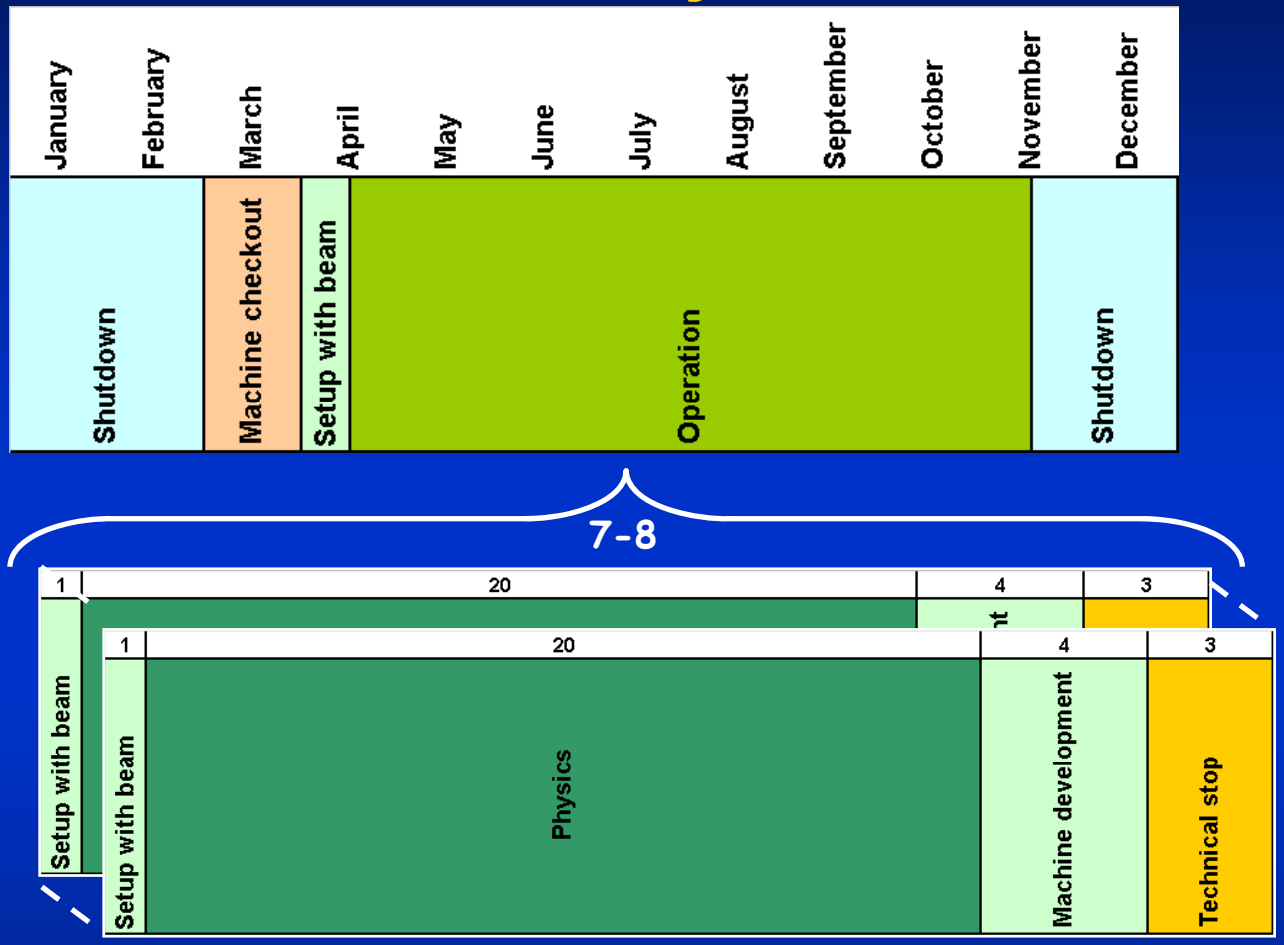
les.robertson@cern.ch

Service Availability - Experience

- Experience to date is that **scheduled** interventions account for far more downtime than unscheduled ones
  - 💣Non-scheduled 'transparent' interventions can be highly pernicious...
- 💣The worst interventions of all so far have been extended downtimes at numerous sites for cooling / power work
- ➢ **The "WLCG Tier0" service is so complex that there are interventions every week – often concurrently**
- Further pressure will be generated from the LHC running schedule (next - hidden) – effectively **reducing** the time slots when such necessary & essential work can take place
- **But** – and it's a big but – apart from 'pathological cases', most interventions **could** be made 'transparently'

Breakdown of a normal year

- From Chamonix XIV -



Service upgrade slots?

~ 140-160 days for physics per year
Not forgetting ion and TOTEM operation

Leaves ~ 100-120 days for proton luminosity running
? Efficiency for physics 50% ?
~ 50 days ~ 1200 h ~ 4 10^6 s of proton luminosity running / year

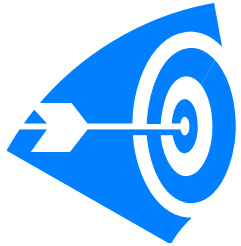
Transparent Interventions - Definition

- Have reached agreement with the **LCG VOs** that the combination of hardware / middleware / experiment-ware **should** be resilient to service "glitches"
- ➢ **A glitch is defined as a short interruption of (one component of) the service that can be hidden – at least to batch – behind some retry mechanism(s)**
- ➢ **How long is a glitch?**
- All central CERN services are covered for power 'glitches' of up to 10 minutes
  - **Some are also covered for longer by diesel UPS but any non-trivial service seen by the users is only covered for 10'**

¿ Can we implement the services so that ~all interventions are 'transparent'?

☺ **YES** – with some provisos

Targetted Interventions



- Common interventions include:
  - Adding additional resources to an existing service;
  - Replacing hardware used by an existing service;
  - Operating system / middleware upgrade / patch;
  - Similar operations on DB backend (where applicable).

- Pathological cases include:
  - Massive machine room reconfigurations, as was performed at CERN (and elsewhere) to prepare for LHC;
  - Wide-spread power or cooling problems;
  - Major network problems, such as DNS / router / switch problems.

➢ **Pathological cases clearly need to be addressed too!**

Major Interventions – C2ALICE

08:45	Ulrich	Stop new Alice Lxbatch jobs from starting
09:00	Miguel, Jan	Stop the instance
09:15	Nilo	Backup stager and DLF databases
10:00	Miguel, Jan	Upgrade the databases
10:30	Miguel, Jan	Upgrade the headnodes + diskservers;
	Miguel, Jan	Reboot diskservers
11:30	Miguel, Jan	Test the instance
12:00	Miguel, Jan	Open service;
	Ulrich	Activate LSF queues
12:15	Miguel, Jan	Announce completion; Update GOCDB

ALWAYS USE AN INTERVENTION PROCEDURE

And Record Openly Any Problems…

- The intervention is now complete and tier1 and tier2 services are operational again except for enabling of internal scripts.

- Two problems encountered.

1. A typo crept in somewhere, **dteam** became **deam** in the configuration. Must have happened a while ago and was a reconfiguration problem waiting to happen.
2. fts103 when rebooted for the kernel upgrade (as were the rest) decided it wanted to reinstall itself instead and failed since not a planned install. Again an accident waiting to happen.

- Something to check for next time.

- Consequently the tiertwo service is running in degraded with only one webservice box. If you had to choose a box for this error to occur on it would be this one.

- Should be running non-degraded mode sometime later this afternoon.

Transparent Upgrades: How do we do it?

- The basic trick: load-balanced servers & rolling upgrades
- ➢ **e.g. take 1 box out of service, upgrade & add back**
- During intervention, load is carried by remaining boxes

- Additional redundancy and availability can be provided by deploying services across multiple sites

☹ **Note that this has a 'dark side' – cross-site problem resolution costs can outweigh the benefits**

➢ **But have in any case to be solved for critical WLCG services, such as the File Transfer Service**

Advantages

- The advantages of such an approach are simply huge!

☺ **The users see a greatly improved service**

☺ **Service providers have significantly more flexibility in scheduling interventions**

☺ **The service provider – user relationship is enhanced**

☺ **Everyone's stress levels plummet!**

➢ **But it must be supported by the middleware…**

Transparency - Caveat

Here's what the Encyclopedia Galactica has to say about alcohol. It says that alcohol is a colourless volatile liquid formed by the fermentation of sugars and also notes its intoxicating effect on certain carbon-based life forms.

The Hitchhiker's Guide to the Galaxy also mentions alcohol. *It says that the best drink in existence is the Pan Galactic Gargle Blaster. It says that the effect of a Pan Galactic Gargle Blaster is like having your brains smashed out by a slice of lemon wrapped round a large gold brick*.

The Guide also tells you on which planets the best Pan Galactic Gargle Blasters are mixed, how much you can expect to pay for one and what voluntary organizations exist to help you rehabilitate afterwards.

The Guide even tells you how you can mix one yourself. The Hitchhiker's Guide to the Galaxy sells rather better than the Encyclopedia Galactica.

Reliability: Conclusions

- **Service interventions are both inevitable and necessary**

➢ For a reliable and sustainable Grid service, the ability to perform such interventions 'transparently' is essential – but needs to be 'built in'
(m/w + h/w + procedures)

☺ **The technology exists and is in production today!**

- **Some much needed enhancements to the tools for scheduling / announcing / measuring service interventions are also critical**

- I have not talked explicitly about the reliability of the services ***per se*** – but the experience has been (after a **L O N G** period of hardening) that the main Grid services offer sufficient reliability

Scalability

Some targets for scalability and real life experience in implementing them

Scalability – File Transfer Example

- LHC Experiments use a file size ~1GB
- Based on expected data rates & number of sites, the number of files to be transferred Tier0→Tier1 is 10^5 - 10^6 per day
  - Correspondingly higher if Tier2s also included in the game
- 'Manual intervention' to resolve file transfer problems is very time consuming, i.e. expensive and non-scalable
- Target: maximum 1 such problem per site per day

➢ **Service has to be reliable to 1 in $10^{5/6}$**

Scalability – Operations Example

- Current operations model is very ‘eye-ball intensive’
- ***And its not 24 x 7…***
- ***Don’t even mention public holidays…***
- **How will /can this scale to:**
  - Many more users?
  - A production Grid infrastructure?

➢ **It won’t.** Service reliability will of course help, but much more automation is clearly needed...

Scalability – User Support Example

- The story is the same...
- How many Ticket Processing Managers (TPMs) can we afford?
- How many users do we / will we have?
- ***How do we get the service to be so reliable and so well documented that we can survive?***

➢ **Need to think of the cost of each ticket**

- One that takes 1 hour of TPM-time costs €10-30 possibly much more if user / VO costs also included!
  - Whilst TPMs probably rarely spend 1 hour / ticket, 3rd level support often spend considerably longer! Some unscheduled 'transparent' interventions have cost several weeks of expert time and caused extreme user dissatisfaction!
- **This is why call centres charge you per call!**
  - And why they are where they are...

Scalability - Conclusions

- If solutions are to cope with very large numbers of users (or whatever), **great care** must be taken to ensure that the solutions really scale

➢ **The critical issue (in most cases) is the available / required manpower to provide a solution**

- Computers are much better at doing repetitive tasks (rapidly) than humans!

- **If you can write a procedure to be followed, you can also write a script / programme / tool**

Accountability

Some targets for accountability and real life experience in implementing them

Accountability - Definition

- **By ‘accountability’ I mean far more than simply ‘accounting’**

➢ I mean the whole bag of measuring / tracking / logging / monitoring what is going on – and has gone on!

  - Including also middleware versions, intervention dates / plans / logs etc.

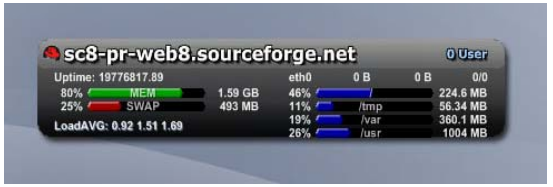
- This is being added rather late to the ‘WLCG’ services

- And again – needs to be built in from the beginning for a large-scale, multi-disciplinary, sustainable e-Infrastructure

- Some examples follow...

The Dashboard – CHEP '06

- Sounds like a conventional problem for a 'dashboard'
- But there is not one single viewpoint…
  - Funding agency – how well are the resources provided being used?
  - VO manager – how well is my production proceeding?
  - Site administrator – are my services up and running? MoU targets?
  - Operations team – are there any alarms?
  - LHCC referee – how is the overall preparation progressing? Areas of concern?
  - …
- Nevertheless, much of the information that would need to be collected is common…
- ➢ **So separate the collection from presentation (views…)**
- As well as the discussion on metrics…



Monitoring - Status

- Since CHEP ’06 the number of monitoring / logging / reporting / dashboard efforts has increased
- There is a tendency for at least **some** of these efforts to attempt to cover the entire space

➢ **But this conflicts with the basic requirements!**

- Sites / VOs necessarily have their own monitoring tools and / or need for specific views

➢ **The ability to select the specific bits of information and correlate views from different sources is fundamental**

- c.f. ‘screen proliferation’ in the early days of LEP...

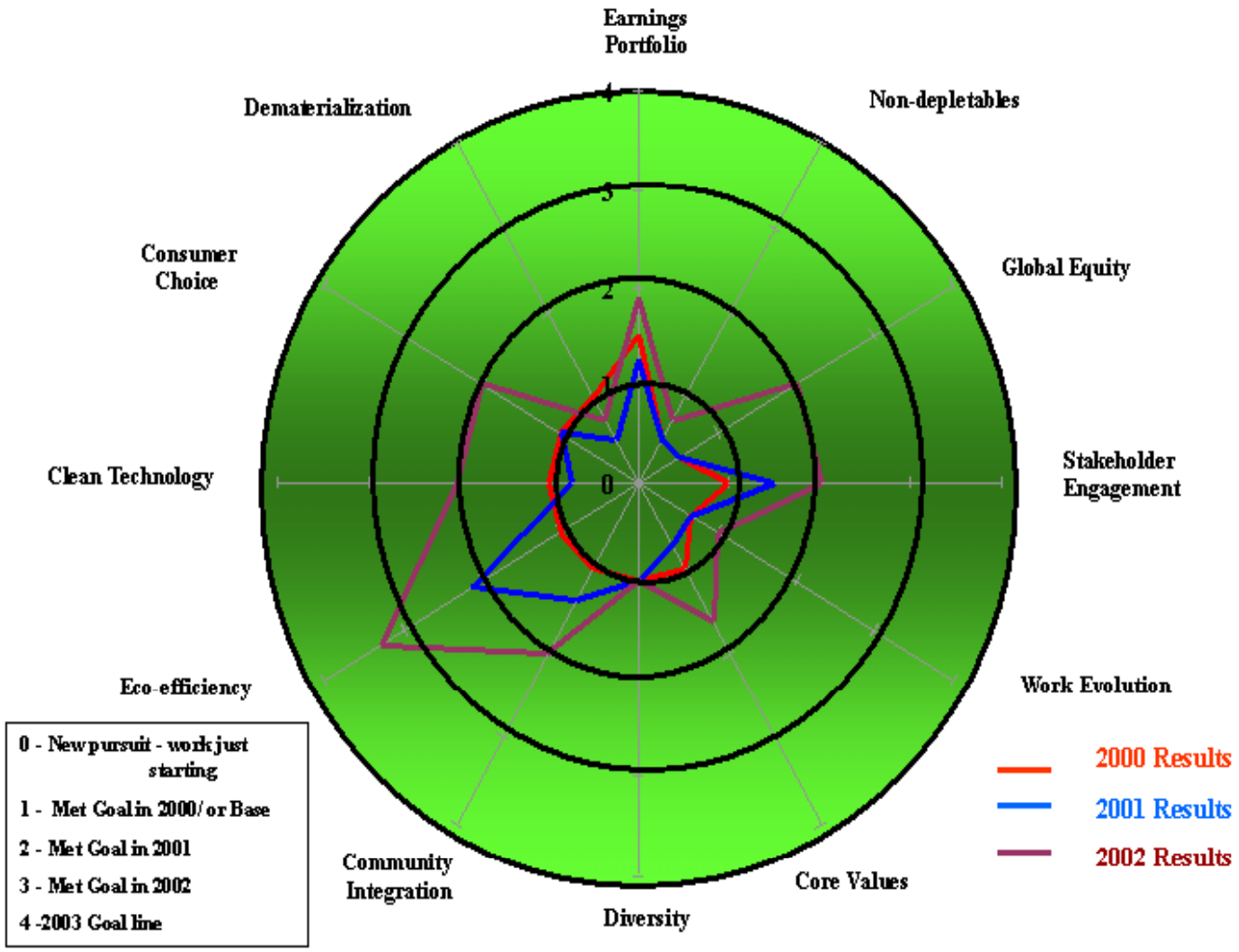
The Requirements on WLCG

- **Resource requirements**, e.g. ramp-up in Tier*N* CPU, disk, tape and network
  - Look at the Computing TDRs;
  - Look at the resources pledged by the sites (MoU etc.);
  - Look at the plans submitted by the sites regarding acquisition, installation and commissioning;
  - ➢ **Measure what is currently (and historically) available; signal anomalies.**

- **Functional requirements**, in terms of services and service levels, including operations, problem resolution and support
  - Implicit / explicit requirements in Computing Models;
  - Agreements from Baseline Services Working Group and Task Forces;
  - Service Level definitions in MoU;
  - ➢ **Measure what is currently (and historically) delivered; signal anomalies.**

- **Data transfer rates** – the Tier*X* ←→ Tier*Y* matrix
  - Understand Use Cases;
  - ➢ **Measure …**

The Dashboard Again…



Accountability - Conclusions

- The key to providing a robust, scalable, manageable and Sustainable **Grid Infrastructure**

- The necessary hooks and infrastructure are fundamental components of the overall Grid service

- And need to be at least as reliable and as available as the hosted services themselves!

- Some re-use of existing technologies could make sense – e.g. database mirroring / replication for Grid infrastructure databases as for m/w ones – *sharing of techniques also for other services & DBS, such as dCache/PostgreSQL etc.*

Interoperability

Interoperability is literally taken for granted in the world of the Internet / Web. Surely it is as fundamental a principle to the Grid too?

Grid Computing – A Definition

- The definitive definition of a Grid is provided by [1] Ian Foster in his article **"What is the Grid? A Three Point Checklist"** [2].

- The three points of this checklist are:

  1) Computing resources are not administered centrally;

  2) Open standards are used;

  3) Non-trivial quality of service is achieved.

➢ **With a fair degree of success, we have demonstrated the importance of Interfaces, rather than Implementation (SRM)**

- **The use of open standards surely has to a corner-stone of any future Grid infrastructure**

WLCG as a Virtual Organisation

- According to Wikipedia:

  *In grid computing, a **Virtual Organization** is a group of individuals or institutions who share the computing resources of a "grid" for a common goal.*

- VO-specific services are clearly and logically a requirement – but must satisfy constraints of hosting Grid

- To the VO in question, these services are as fundamental as any other services that the VO depends upon...

Agenda

By harnessing the following 3 “abilities”:

- ✓ Reliability
- ✓ Scalability
- ✓ Accountability

we gain both “manageability” & (critically) “usability”

Usability: an Analogy with the Web...

- In the early days of the Web, to make content **available** you had to type raw HTML
  - Some of us still do...

- To **view** content, on many systems your only option was a clunky[1] line-mode[2] browser[3]

➢ **It is inconceivable that it would now be so ubiquitous if higher-level, user-friendly tools had not emerged**

- **Are we today in the 'HTML-age' – waiting for the wheel & later steam engines to chug into view?**

Conclusions

- An analysis of service interruptions from the EDG-RLS days on shows that **scheduled interventions** are by far the main cause of downtime

- **This continues to be true today and is valid for all services** (there are many non-scheduled interventions that need to be resolved)

- Significant improvements in service level – including both ease of use and ease of delivery – could be achieved through resilient services

Reliability / Scalability / Accountability / Usability

- **These issues will be key to building a long-term scalable and affordable e-Infrastructure**

- This work has been re-launched following discussions at WLCG GDB & MG – see Operations slot at WLCG Collaboration Workshop

- **Proposed target: solve in production << CHEP 2009**

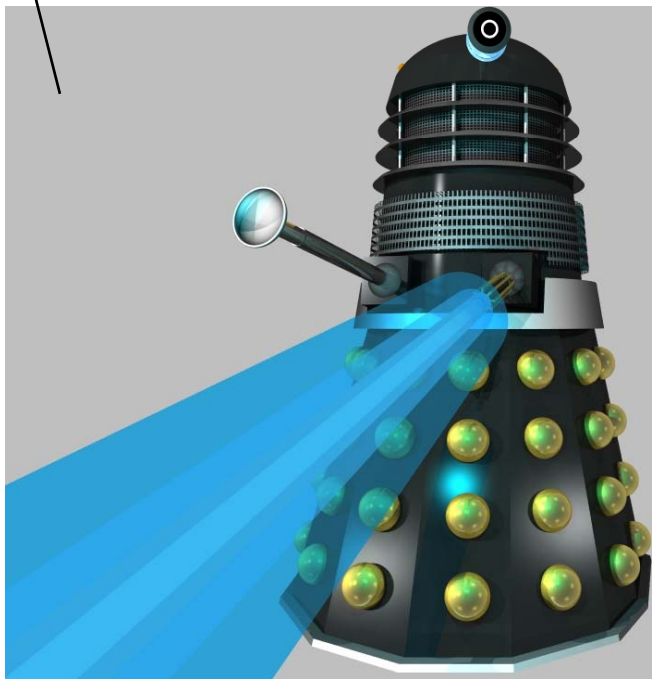
Services - Summary

- Its open season on SPOFs...

Seek!

Locate!

Exterminate!



The End